MOD 25 CHI-SQUARE TEST FOR INDEPENDENCE

Learning Goal

- Conduct a chi-square test for independence between two categorical variables.

INTRODUCTION

Previously, we learned how to use the ANOVA F-test whether two or more population means differ. You can also think of the ANOVA test as examining the relationship between two variables. The explanatory variable is categorical and has 3 or more values (e.g. the groups or populations). The response variable is quantitative.

In this activity we work with two categorical variables to determine if there is a relationship between them. Another way to say this is: Are the categorical variables independent of each other or dependent on each other?

This new statistical test is called a *Chi-Square Test for Independence.*

We use the Chi-square test of independence when both the explanatory variable and the response variable are categorical. The Chi-square test of independence is a hypothesis test used to determine whether two categorical variables are likely to be related (dependent) or not related (independent). In other words …

> ***If the explanatory and response variables are dependent (related) this means the explanatory variable impacts the distribution of values across the categories of the response variable***.

As we work through this activity, the preceding sentence may become more clear. But before we move on, use the space below to note the key points from this page (excluding that indented sentence in bold type.) You'll discuss the key points in your group.

CHECK YOUR UNDERSTANDING: WHICH HYPOTHESIS TEST APPLIES?

1) For each research question, identify the appropriate hypothesis test: z-test for one proportion, t-test for one mean, t-test for a difference in two means, ANOVA, Chi-square.

   a) Is there a relationship between amount of credit card debt and employment status (unemployed, employed part-time, employed full-time)?

   b) Is there a relationship between whether or not a student owns a credit card and employment status (unemployed, employed part-time, employed full-time)?

   c) Is gender associated with hybrid car ownership?

   d) Do the majority of community college students drive to their college?

   e) Do women drive more miles weekly on average than men?

STATING HYPOTHESES FOR THE CHI-SQUARE TEST FOR INDEPENDENCE

The null hypothesis can be stated in several equivalent ways:

- There is no relationship between the categorical variables.
- There is no association between the categorical variables.
- The categorical variables are independent.

The alternative hypothesis says there is a relationship (association), which means the categorical variables are dependent.

CHECK YOUR UNDERSTANDING: STATE THE HYPOTHESES

2) The Disability IAT measures implicit attitudes toward disabled people. Prior to taking any implicit association test, participants respond to various survey questions. One question asks participants to identify their disability status. Another question asks participants to rank their preference for abled/disabled people.

   - The explanatory variable is *disability status* (abled vs. disabled).
   - The response variable is *prefers* (seven categories of preference from "*strongly prefers disabled people*" to "*strongly prefers abled people*.")

   State the null and alternative hypotheses.

ANALYZING THE DATA FOR A CHI-SQUARE TEST

Here is the two-way contingency table for a random sample of 3000 participants in the 2020 Disability IAT.

	Strongly prefers disabled	Moderately prefers disabled	Slightly prefers disabled	No preference	Slightly prefers abled	Moderately prefers abled	Strongly prefers abled	Total
Disabled	12	13	48	450	85	34	13	655
Abled	14	24	54	1616	409	163	65	2345
Total	26	37	102	2066	494	197	78	3000

Suppose we want to know whether the variables *disability status* and *preference* are related, i.e. whether the variable *preference* depends on the variable *disability status*. To investigate this we would ask the following research question.

Research question: Are *preference* and *disability status* independent?
What does independent mean?

Answer: Two categorical variables are independent if the explanatory variable does not impact the distribution of the response variable (in the contingency table). **In other words, *disability status* and *preference* are independent if *preference* looks the same for abled and disabled people in the table when we ignore *disability status*.**

What? How do we ignore *disability status* (the explanatory variable)? Well … let's find out.

Check your understanding: How do we “ignore” the explanatory variable?

3) First, fill in the missing percentages to complete the distribution of *preference* for abled and disabled people. **These are the interior cells in the table**. Notice: that when we calculate these percentages, we divide by the totals for the appropriate category of the explanatory variable, *disabled* or *abled*. **In other words, we are NOT ignoring the explanatory variable, *disability status*.**

   (No need to show your work as we have; just enter the percentages.)

	Strongly prefers disabled	Moderately prefers disabled	Slightly prefers disabled	No preference	Slightly prefers abled	Moderately prefers abled	Strongly prefers abled	Total
Disabled	12	13 13/655 = 1.98%	48 48/655 = 7.33%	450	85	34 34/655 = 5.19%	13	655
Abled	14	24 24/2345 = 1.02%	54	1616	409	163 163/2345 = 6.95%	65 65/2345 = 2.77%	2345
Total	26	37 1.23%	102	2066	494	197	78	3000

Now fill in the missing percentages for all people together. This is the bottom row of the table. Notice that we divide by the total number of participants (regardless of their disability status). So, we ARE ignoring *disability status* when we fill in the bottom row.

If *disability status* and *preference* are independent, then the distribution of the response variable, *preference*, will be the same for each population of the explanatory variable, disabled and abled people (when we compare percentages, not counts).

So, we can conclude that the two variables are independent, if for each category of the response variable, *preference*, the population percentages for disabled and abled people are the same.

a) But, looking at each category of the response variable (*preference*), we see that percentages are not the same for disabled and abled people. So, should we conclude that the variables are dependent? Why or why not?

b) How can we know if the difference in the percentages for each category of the response variable, *preference*, are insignificant and therefore due to the variances we expect to see in random sampling?

c) So, in this scenario, what does a P-value represent?

The P-value is the probability of getting a sample with these differences in percentages given the sample was randomly selected from a population with no differences in percentages for each category of the response variable, *preference*.

THE CHI-SQUARE TEST STATISTIC & EXPECTED COUNTS

The Chi-Square test statistic (written χ^2) measures how much the observed data in our sample differs from what we expect to happen when the null hypothesis is true.

Here is the formula:

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

We will not calculate the χ^2 test statistic by hand (StatCrunch will do this for us). However, we will spend a little time doing some calculations to help you understand the expected counts and their importance.

The expected counts are the counts we expect to see if the null hypothesis is true, e.g. if the variables *disabled status* and *preference* are independent (not related).

If the variables are independent, then the percentages should be the same for each category of the response variable, *preference*. The explanatory variable should not impact these percentages. In other words …

4) If the variables (*disability status* and *preference*) are independent, we expect the percentages of disabled and abled people to be the same in each category of the response variable, *preference*. (For your convenience, our table with the percentages is included below.)

	Strongly prefers disabled	Moderately prefers disabled	Slightly prefers disabled	No preference	Slightly prefers abled	Moderately prefers abled	Strongly prefers abled	Total
Disabled	12 (1.83%)	13 (1.98%)	48 (7.33%)	450 (68.7%)	85 (12.98%)	34 (5.19%)	13 (1.98%)	655 (100%)
Abled	14 (0.6%)	24 (1.02%)	54 (2.3%)	1616 (68.91%)	409 (17.44%)	163 (6.95%)	65 (2.77%)	2345 (100%)
Total	26 (0.87%)	37 (1.23%)	102 (3.4%)	2066 (68.87%)	494 (16.47%)	197 (6.57%)	78 (2.6%)	3000 (100%)

So, if the *disability status* and *preference* are independent, what percentage of disabled and abled people should we expect to see for each category of the response variable? To answer this question, fill in the table below.

Preference	**Disabled: expected %**	**Abled: expected %**
Strongly prefers disabled		
Moderately prefers disabled		
Slightly prefers disabled		
No preference		
Slightly prefers abled		
Moderately prefers abled		
Strongly prefers abled		

Expected counts

For each category of the response variable, the expected counts are calculated as follows.
(expected percentage) X (total of the explanatory variable category)

5) Fill in the expected counts for each category of the response variable. We filled in the expected counts for *no preference* (our work is shown below the table).

   (No need to show your work; just enter the expected counts.)

	Strongly prefers disabled	Moderately prefers disabled	Slightly prefers disabled	No preference	Slightly prefers abled	Moderately prefers abled	Strongly prefers abled	Total
Disabled	12 (1.83%)	13 (1.98%)	48 (7.33%)	450 (68.7%) **451.10**	85 (12.98%)	34 (5.19%)	13 (1.98%)	655 (100%)
Abled	14 (0.6%)	24 (1.02%)	54 (2.3%)	1616 (68.91%) **1615.00**	409 (17.44%)	163 (6.95%)	65 (2.77%)	2345 (100%)
Total	26 (0.87%)	37 (1.23%)	102 (3.4%)	2066 (68.87%)	494 (16.47%)	197 (6.57%)	78 (2.6%)	3000 (100%)

If the variables *disability status* and *preference* are independent, the expected percentage of disabled and the expected percentage of abled people who have *no* preference is 68%.
Here are the calculations of the expected counts for *no preference*.

Disabled: 68.87% of 655 = $(0.6887)(655) = 451.0985$

We expect 451.10 of the disabled people in our sample to like disabled and abled people equally.

Abled: 68.87% of 2345 = $(0.6887)(2345) = 1615.0015$

We expect 1615.00 of the abled people in our sample to like disabled and abled people equally.

Note that in the table, the expected counts for *no preference* are very close to the actual counts. But, this may not be the case for the remaining categories of the response variable.

Why are expected counts important?

Before we can use StatCrunch to find the χ^2 test statistic and the P-value, we need to make sure conditions are met for use of the χ^2 density curve. Here are the conditions:

- We have a random sample from the population, or we have random assignment of treatments.
- Each expected count is at least 5. (This is the same idea as requiring success and failures to be at least 10 when we used the normal distribution for inference on proportions).

6) Are conditions met for use of the χ^2 model to determine whether *disability status* and *preference* are related? Explain.

Let's Use StatCrunch to Assess the Evidence

To conduct the Chi-square test of independence, we used StatCrunch and the familiar contingency table (Stats → Contingency → With Data). We chose the explanatory variable (*disability status*) as the row variable and the response variable (*preference*) as the column variable. We also chose to display *Row percent* and *Expected count.* Here are our results.

Contingency table results:

	Strongly prefers disabled	Moderately prefers disabled	Slightly prefers disabled	No preference	Slightly prefers abled	Moderately prefers abled	Strongly prefers abled	Total
Disabled	12 (1.83%) (5.68)	13 (1.98%) (8.08)	48 (7.33%) (22.27)	450 (68.7%) (451.08)	85 (12.98%) (107.86)	34 (5.19%) (43.01)	13 (1.98%) (17.03)	655 (100%)
Abled	14 (0.6%) (20.32)	24 (1.02%) (28.92)	54 (2.3%) (79.73)	1616 (68.91%) (1614.92)	409 (17.44%) (386.14)	163 (6.95%) (153.99)	65 (2.77%) (60.97)	2345 (100%)
Total	26 (0.87%)	37 (1.23%)	102 (3.4%)	2066 (68.87%)	494 (16.47%)	197 (6.57%)	78 (2.6%)	3000 (100%)

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	6	60.713564	<0.0001

Interpret the results

7) Use the information at the bottom of the previous page to respond to each of the following.

   a) What is the χ^2 statistic?

   b) What is the P-value? What does the P-value represent?

   c) State a conclusion in context.