Learning Goal: Conduct a chi-square test for independence between two categorical variables in a contingency table.

Introduction:

Previously, we learned the ANOVA test, which determines if 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. The response variable is quantitative.

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

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

In this Chi-Square test we select one sample and for each individual in the sample we collect two pieces of categorical data. The Chi-square test of independence determines whether explanatory variable impacts the distribution of values for the response variable.

Check your understanding:

1) For each research question, to 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?

   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 on average than men?

Stating hypotheses for Chi-Square Test for Independence

The null hypothesis can be stated in several equivalent ways:

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

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

Check your understanding:

2) In a study of marketing to children, researchers examine cereal placement on grocery store shelves. The explanatory variable is cereal type (child vs. adult). The response variable is shelf placement (bottom shelf, middle shelf, top shelf)

   State the null and alternative hypotheses.

Analyzing the data for a Chi-square test

Here is the two-way table for the 77 cereals in this random sample of cereals. In a Chi-Square setting, we call the data observed counts.

	Bottom	Middle	Top	Total
Adult	11	5	36	52
Child	9	16	0	25
Total	20	21	36	77

What does independent mean? Two categorical variables are independent if the distribution of the response variable does not change when we take the explanatory variable into account. In other words, Shelf and Target are independent if shelf placement looks the same for child and adult cereals and matches the distribution of cereals across shelves when we ignore the cereal type.

Check your understanding:

3) Fill in the missing percentages to complete the distribution of shelf placement for adult cereals, child cereals, and all cereals together.

	Bottom	Middle	Top	Total
Adult	11	5	36 36/52=69%	52
Child	9 9/25=36%	16	0 0/25=0%	25
Total	20 20/77=26%	21	36	77

If the Target variable and Shelf variable are independent, then the distribution of shelf placement will be the same (or very close to the same) for Adult and Child cereals. But we see differences in these distributions. For example, there is a larger percentage of adult cereals on the top shelf. Therefore, we need to determine if these differences are small enough to attribute to the variation we expect to see in random sampling, or not. Surprise, surprise, we need a test statistic and a P-value.

The Chi-Square test statistic

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 have you calculate the χ^2 test statistic by hand. However, we will spend a little time doing some calculations to help you understand the expected counts.

The expected counts are the counts we expect if the null hypothesis is true (the variables are not related; they are independent.)

If Shelf and Target are independent, then the placement distribution for Adult cereals and Child cereals will be the same as the placement distribution overall. Note that we are looking at the distribution of the response variable.

Previously, we found the overall placement distribution (bottom row of the previous table.) Record those percentages below:

Bottom shelf: 20/77 = 0.26 = 26%   Middle shelf: __________   Top shelf: _____________

Let's start by finding the expected counts for ADULT cereals. We use these overall placement percentages to calculate the number of adult cereals we expect to see on each shelf:

- Bottom shelf: If the target consumer is not affecting the location (which is what the null hypothesis says), we expect 26% of 52 adult cereals to be on the bottom Therefore, the expected count is 0.26(52) = 13.52

Check your understanding:

4) Find the expected counts for adult cereals for the middle shelf and for the top shelf.

   a) Middle shelf: What is the expected count of adult cereals for the middle shelf?

   b) Top shelf: What is the expected count of adult cereals for the top shelf?

(To check your work, the expected counts for adult cereals across the shelves should add to 52.)

   c) Now find the expected counts for CHILD cereals:

Bottom: Middle: Top:

We calculated the expected counts by hand in this example to help you understand where the expected counts come from. But, in general, we will use StatCrunch to find expected counts.

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 X^2 density curve. Here are the conditions:

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

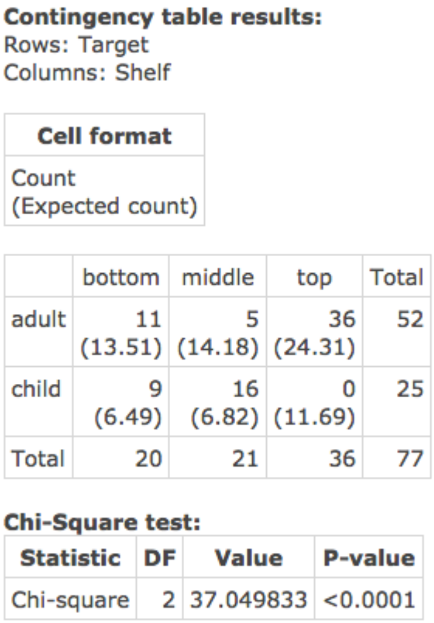
Check your understanding:

5) Verify that the conditions are met for use of the X^2 model for the cereal study.

Assessing the evidence

StatCrunch gives the results in an expanded two-way table. You can choose to show the expected counts or the distribution of the response variable (e.g. row percents if the explanatory variable defines the rows.)

When you have the data spreadsheet in StatCrunch, choose Stats, Tables, Contingency, with Data.



Contingency table results:
Rows: Target
Columns: Shelf

Cell format
Count (Expected count)

	bottom	middle	top	Total
adult	11 (13.51)	5 (14.18)	36 (24.31)	52
child	9 (6.49)	16 (6.82)	0 (11.69)	25
Total	20	21	36	77

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	2	37.049833	<0.0001

Check your understanding:

6) What is the X^2 statistic?

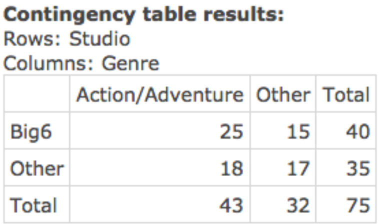
What is the P-value?

7) State a conclusion.

Group Work.

8) For movies, is there a relationship between studio type (Big 6, Other) and the movie genre (Action/Adventure, Other)? Or are they independent?

To investigate this question, students want to use the movie data that we analyzed earlier in the course. The movies in the data set are the top 75 USA box office sales earners of all time. Data was taken from IMDb.com.



Contingency table results:
Rows: Studio
Columns: Genre

	Action/Adventure	Other	Total
Big6	25	15	40
Other	18	17	35
Total	43	32	75

Can they use this data to conduct a chi-square test for independence? Why or why not?

9) We will use a study conducted during the 1980's to practice the chi-square test of independence.

Background: Clinical depression is a recurrent illness requiring treatment and often hospitalization. Nearly 50% of people who have an episode of major depression will have a recurrence within 2-3 years.

The Study: During the 1980's the federal government, through the National Institutes of Health (NIH), sponsored a multi-centered, randomized, controlled, clinical trial to evaluate two drugs to prevent the recurrence of depression in patients who have had at least one previous episode of the illness (Prien et al., *Archives of General Psychiatry*, 1984).

The Study Design: Patients suffering from depression were recruited from 5 medical clinics in 5 large cities. They were randomly assigned to one of the 3 treatment groups: Imipramine, Lithium, or a Placebo. Patients were followed for 2-4 years to see whether or not they had a recurrence of depression. If they did not have a recurrence within this time frame, then their treatment was considered a success. The study was double-blinded.

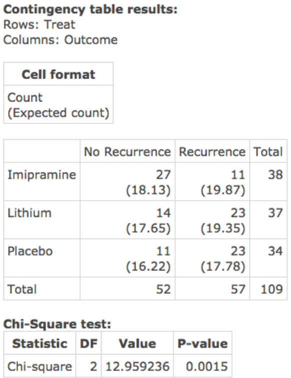
a) Select all the appropriate pairs of hypotheses to investigate the association between treatment and recurrence of depression.

H_0: There is a relationship between treatment and recurrence of depression.
H_a: There is no relationship between treatment and recurrence of depression.

H_0: Recurrence of depression is independent of treatment
H_a: Treatment and recurrence are dependent

H_0: There is no association between treatment and recurrence of depression.
H_a: There is an association between treatment and recurrence of depression.

b) In the StatCrunch print-out, what does the 18.13 tell us?



Contingency table results:
Rows: Treat
Columns: Outcome

Cell format
Count (Expected count)

	No Recurrence	Recurrence	Total
Imipramine	27 (18.13)	11 (19.87)	38
Lithium	14 (17.65)	23 (19.35)	37
Placebo	11 (16.22)	23 (17.78)	34
Total	52	57	109

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	2	12.959236	0.0015

c) Which of the following statements explains the expected counts? (Choose all of the statements that are correct)

- the number of patients expected to relapse or not if all 3 treatments are similarly effective
- the number of patients expected to to relapse or not if there is a relationship between treatment type and recurrence
- the number of patients expected to relapse or not if there is a no association between treatment type and recurrence
- the expected number to relapse or not if the null hypothesis of independence is true.

d) The P-value is 0.0015. Choose all interpretations that are correct.

a. The P-value is statistically significant at a significance level of 1%.
b. The P-value is not statistically significant at a significance level of 1%.
c. If outcome is independent of treatment type, we can expect a chi-square value of 12.96 or greater to occur less than 1% of the time.
d. The results are so rare we conclude that these results are due to fluctuations expected when randomly assigning subjects to treatments when the treatments have no association with the patient's outcome.
e. We can expect a chi-square value of 12.96 or greater about 0.15% of the time if there is no relationship between treatment type and patient outcome.

e) What can we conclude?

10) Is there a relationship between political affiliation and willingness to participate in political surveys?

	Survey Participation Yes	Survey Participation No
Democrat	49	47
Independent	15	27
Republican	32	30
None or will not say	8	10

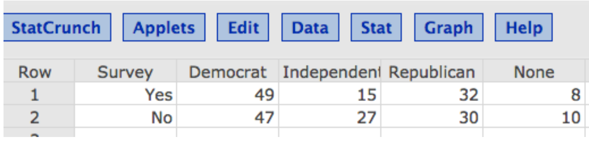
a) State your hypotheses.

b) Use StatCrunch to find expected counts and to test for independence (See instructions below.) Fill in the expected counts in the table.

c) Explain why we can use the X^2 model.

d) Give the P-value and state your conclusion.

StatCrunch instructions:
Open StatCrunch and enter the data as shown.
Choose **Stat, Tables, Contingency, With Summary**



Row	Survey	Democrat	Independen	Republican	None
1	Yes	49	15	32	8
2	No	47	27	30	10

Select Columns: select the response categories (Democrat, Independent, Republican, None)
Row labels: select the explanatory variable (Survey)
Display: choose expected count (if you want to see this)
Hypothesis tests: choose chi-square test for independence