Math 247: Chi-Square Hypothesis Test for Association (Sections 10.1 and 10.3)

Suppose you want to determine whether sugar consumption impacts focus in children.
Describe how you would do a study to determine whether sugar CAUSES a change in focus.

What are some possible confounding variables and how does the study design prevent them from influencing (confounding) the results?

Example: Suppose this is your study:

You take a random sample of 60 5-year-olds and randomly assign them to 2 groups. One group is the "Soda Group" and has 30 kids; the other group is the "No Soda Group" and has 30 kids. You give the soda kids a non-caffeinated 12-ounce can of soda to drink and the other kids get flavored water (same flavor as the soda). You then give all the kids a task of coloring in a picture and record how many have finished the task in 10 minutes.

What is the Independent Variable? _____

What is the Dependent Variable?_____

What is the population of interest (implied)?

You get the results in the table below.

		Sugar Consumption	
		Soda	No Soda
Focus	Finished task	15	21
	Didn't finish task	15	9

What is your impression about how sugar relates to focus for the kids in this sample, based on this data?

Since this is only a sample, how do we extend these results to the entire population?

Chi-Square Hypothesis for Association Test:

1. Hypothesize: Identify the two categorical variables in the study.

H_0: There is no association between _____ and _____

H_a: There is an association between _____ and _____

2. Prepare: State what test you will use. Also, choose a Level of Significance, α.

Check conditions for the test. For a Chi-Square Test for Association, the conditions are

1. Random Sample, Independent Observations

2. **All expected counts must be 5 or more.** If any of the expected counts is less than 5, then we shouldn't use this method. There is another method for this situation (Fisher's Exact Test) but we won't be covering that method in this course.

   To check this condition, you have to proceed to the "Compute" step.

3. **Compute: We will use StatCrunch to do the Compute Step, but we'll find the Expected Counts and the Chi-Square Value by hand, once, so you understand what's happening.**

Find the Expected Counts using the formula $E = \dfrac{\text{row total} \times \text{column total}}{\text{table total}}$

and fill in those values on the Two-Way table.

Find $\chi^2 = \sum \dfrac{(O-E)^2}{E}$ and the Degrees of Freedom,

$$\text{df} = (\text{number of rows} - 1) \cdot (\text{number of columns} - 1)$$

4. Interpret: *If the P-value is less than or equal to alpha, then we reject the null and accept the alternative hypothesis. "If P is small, the null will fall" (reject!).*

If P is larger than alpha then we are in limbo, neither rejecting nor accepting the null hypothesis.

- Begin with comparing the p-value to the alpha value (level of significance)
- Next, state whether you reject the null hypothesis and accept the alternative hypothesis or whether you fail to reject the null hypothesis. We do NOT say we "accept the null hypothesis"; instead state that there isn't enough evidence to reject it.

  If $P\text{-value} \le \alpha$, write "we reject H_0, and accept the alternative. There IS sufficient evidence from the data to conclude that there is a statistically significant association between _____ and _____."

  If $P\text{-value} > \alpha$, write "we reject H_0, and accept the alternative. There is NOT sufficient evidence from the data to conclude that there is a significant association between _____ and _____."

Example (continued): Okay, now let's apply these steps to our study on sugar and focus.

You take a random sample of 60 5-year-olds and randomly assign them to 2 groups. One group is the "Soda Group" and has 30 kids; the other group is the "No Soda Group" and has 30 kids. You give the soda kids a non-caffeinated 12-ounce can of soda to drink and the other kids get flavored water (same flavor as the soda). You then give all the kids a task of coloring in a picture and record how many have finished the task in 10 minutes.

You get the results in the table below.

		Sugar Consumption	
		Soda	No Soda
Focus	Finished task	15	21
	Didn't finish task	15	9

Conduct a Chi-Square hypothesis test to determine whether these data show that there is a statistically significant association between Focus and Sugar Consumption.

Step 1: _____

Step 2: _____

Step 3: _____ (In the homework, you'll do this using StatCrunch.)

The table (same as before) has all of the "Observed Values" in each category. O = Observed Value

Work for Expected Counts:

		Sugar Consumption	
		Soda	No Soda
Focus	Finished task	15	21
	Didn't finish task	15	9

Assuming there is NO ASSOCIATION between sugar and focus (the null hypothesis), determine how many kids in the soda group you would EXPECT to have finished the task, if they finished at the same proportion as the whole group. E = Expected Count.

Note that we can find E = Expected Count by using the following formula: $E = \frac{\text{row total} \times \text{column total}}{\text{table total}}$.

Find all the expected counts and put them in the table.

Once you have these values, be sure to CHECK that Condition #2 is satisfied (each cell has an expected value of at least 5).

Just as with standard deviation, we can construct a single value that summarizes all these differences. This new number is called "Chi-Square". **Find the Chi-Square value for this example.**

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

What does that Chi-Square number tell us?

Chi-Square value has a "probability distribution" which changes depending on the "degrees of freedom".

Degrees of Freedom: $\text{df} = (\text{number of rows} - 1) \cdot (\text{number of columns} - 1)$

$\chi^2 =$_____

df = _____

Once we have found χ^2 and df, by hand, we'll confirm our results and find the P-value using StatCrunch **(instructions on the following page).**

Cell format
Count (Expected count) (Contributions to Chi-Square)

	Soda	No Soda	Total
Finished	15 (18) (0.5)	21 (18) (0.5)	36
Didn't Finish	15 (12) (0.75)	9 (12) (0.75)	24
Total	30	30	60

Chi-Square test:

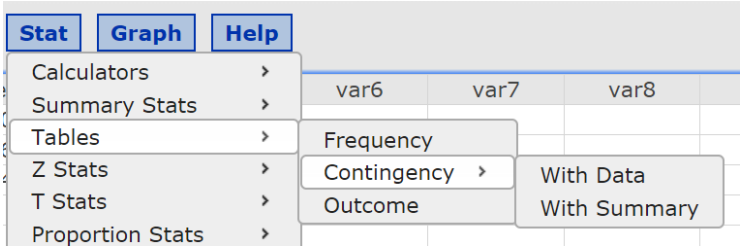
Statistic	DF	Value	P-value
Chi-square	1	2.5	0.1138

Step 4: _____

Using StatCrunch to do a Chi-Square Test for Independence

First, Open StatCrunch and type the Two-Way Table into rows and columns exactly as given in the problem.

Next, Click "Stat" then "Tables" then "Contingency" then "With Summary"



Next, make the following selections in the dialogue box:

For "Select columns", select the columns with the observed values in them.

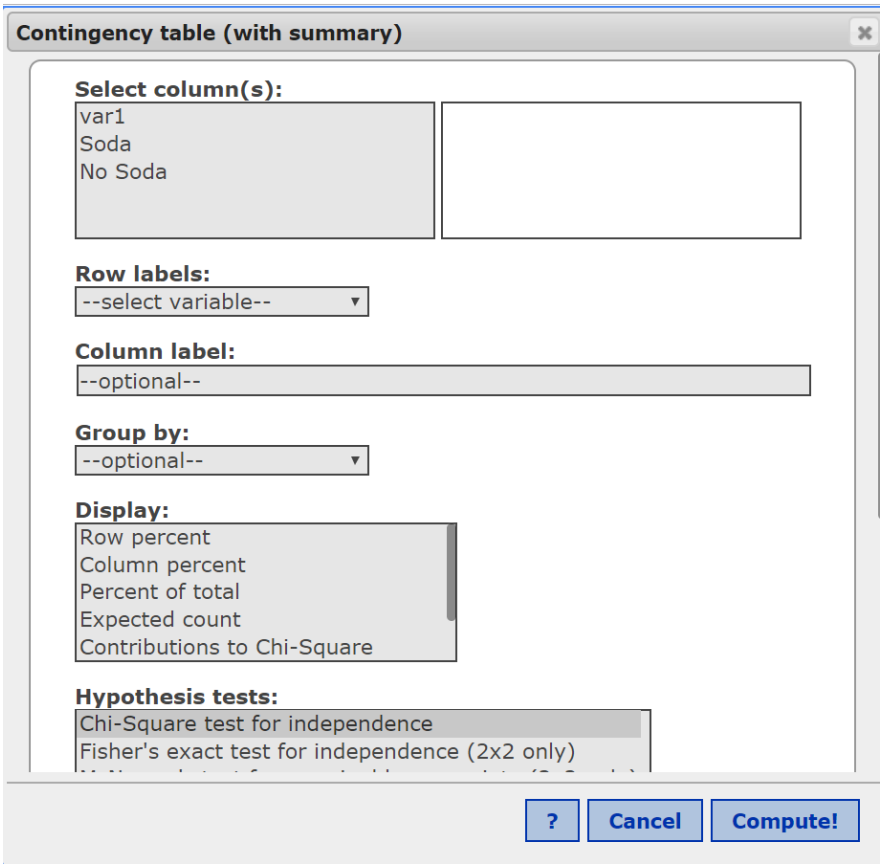
For "Row labels" select the first column (var 1)

Skip "Column label" and "Group by"

For "Display" select
- Expected count
- Contribution to Chi-Square

For "Hypothesis Test" the default is "Chi-Square test for independence" which is what you want.

Click "Compute"



Mediterranean Diet and Health (Source**:** http://onlinestatbook.com/2/case_studies/diet.html)

RESEARCH CONDUCTED BY: De Longerill et al.

CASE STUDY PREPARED BY: David Lane and Emily Zitek

OVERVIEW
Most doctors would probably agree that a Mediterranean diet, rich in vegetables, fruits, and grains, is healthier than a high-saturated fat diet. Indeed, previous research has found that the diet can lower risk of heart disease. However, there is still considerable uncertainty about whether the Mediterranean diet is superior to a low-fat diet recommended by the American Heart Association. This study is the first to compare these two diets.

The subjects, 605 survivors of a heart attack, were randomly assigned follow either (1) a diet close to the "prudent diet step 1" of the American Heart Association (control group) or (2) a Mediterranean-type diet consisting of more bread and cereals, more fresh fruit and vegetables, more grains, more fish, fewer delicatessen foods, less meat. An experimental canola-oil-based margarine was used instead of butter or cream. The oils recommended for salad and food preparation were canola and olive oils exclusively. Moderate red wine consumption was allowed.

Over a four-year period, patients in the experimental condition were initially seen by the dietician, two months later, and then once a year. Compliance with the dietary intervention was checked by a dietary survey and analyses of plasma fatty acids. Patients in the control group were expected to follow the dietary advice given by their physician.

The researchers collected information on number of deaths from cardiovascular causes e.g., heart attack, strokes, as well as number of nonfatal heart-related episodes. The occurrence of malignant and nonmalignant tumors was also carefully monitored.

QUESTIONS TO ANSWER
Is the Mediterranean diet superior to a low-fat diet recommended by the American Heart Association?

DESIGN ISSUES
The strength of the design is that subjects were randomly assigned to conditions. A possible weakness is that compliance rates depended on reports rather than observation since observation is impractical in this type of research.

DESCRIPTIONS OF VARIABLES

VARIABLE	DESCRIPTION

Data (summary of frequencies):

	Cancers	Deaths	Nonfatal illness	Healthy	Total
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

1. What percentage of people on the AHA diet had some sort of illness or death?

2. What percentage of people on the Mediterranean diet had some sort of illness or death?

3. Does the data suggest that type of diet is associated with health status?

4. Conduct a Pearson Chi-Square test to determine if there is a significant association between diet and outcome, using a significance level of .05. Write the "4 Steps" first. Use StatCrunch for the "Compute" step.

1.

2.

3.

(The data is on my website, Math 247, StatCrunch, Diet Study. Fill in the last table below from the StatCrunch results.)

Contingency table results:

Cell format
Count (Expected count) (Contributions to Chi-Square)

	Cancers	Deaths	Nonfatal illness	Healthy	Total
AHA	15 (11.02) (1.44)	24 (19.03) (1.3)	25 (16.53) (4.34)	239 (256.42) (1.18)	303
Mediterranean	7 (10.98) (1.44)	14 (18.97) (1.3)	8 (16.47) (4.36)	273 (255.58) (1.19)	302
Total	22	38	33	512	605

Statistic	DF	Value	P-value

4.