Chapter 4: Methods for Two Categorical Variables

Until this point, we have been working with **one categorical variable** with two or more levels (*toy* - helper/hinder; *font* - signet/salem; *season* - fall/winter/spring/summer). In this set of notes, we will introduce and investigate research questions, statistical analyses, and interpretation for **two categorical variables**.

Example 4.1: MythBusters and the Yawning Experiment

MythBusters, a popular television program on the Discovery Channel, once conducted an experiment to investigate whether or not yawning is contagious. The premise of the experiment was to invite a stranger to sit in a booth for an extended period of time. Fifty subjects were said to be tested in total, of which 34 were "seeded" with a yawn by the person conducting the experiment. The other 16 were not given a yawn seed. Using a two-way mirror and a hidden camera, the experimenters observed and recorded the data which is shown below:

```
library(tidyverse)                                          ①
myth_busters <- read_csv("data/myth_busters.csv")           ②
head(myth_busters)                                          ③
```

① Load the `tidyverse` package
② Read in the `myth_busters` data set
③ Print the top 6 rows of the `myth_busters` data set

```
# A tibble: 6 x 3
  seeding action age_years
  <chr>   <chr>      <dbl>
1 Control Yawned        45
2 Control NoYawn        49
3 Control NoYawn        41
4 Seeded  Yawned        27
```

```
5 Seeded  NoYawn        19
6 Seeded  Yawned        31
```

Research Question: Do these data provide statistical evidence there is a *relationship* between yawn seeding and yawn action?

When we analyze data on two variables, our first step is to distinguish between the *response variable* and the *explanatory* (or *predictor*) *variable*.

> **Note**
>
> - **Response variable:** The outcome variable on which comparisons are made.
> - **Explanatory (or predictor) variable:** This defines the groups to be compared.

1. What are the variables in the MythBusters Yawning experiment? Are they categorical or numerical?

2. Which is the response variable? Which is the explanatory variable?

> **Descriptive Methods for Two Categorical Variables:**
>
> - A **contingency table** shows the joint frequencies of two categorical variables. The rows of the table denote the categories of the explanatory variable, and the columns denote the categories of the response variable.
> - A **bar plot** gives a visual representation of the relationship between two categorical variables. A bar plot graphically presents the information given in the contingency table. We can create three types of bar plots – stacked, dodged, or filled.

Previously, we saw how we can summarize the counts from our data set using `count(VARIABLE)`. This is helpful, but notice we get one column for each count. It is easier to understand the relationship between the explanatory and response variables if we arrange our counts into a contingency table.

```
myth_busters |>                                                    ①
  count(seeding, action)                                           ②
```

① Start with the `myth_busters` data set
② Count the number of subjects in each seeding group and yawn action combination

```
# A tibble: 4 x 3
  seeding action     n
  <chr>   <chr>  <int>
1 Control NoYawn    13
2 Control Yawned     3
3 Seeded  NoYawn    23
4 Seeded  Yawned    11
```

```
library(janitor)
myth_busters |>
  count(seeding, action) |>
  pivot_wider(names_from = action,                                 ③
              values_from = n) |>
  adorn_totals(where = c("row", "col"))                            ④
```

③ Convert the summary table to a contingency table format with the response (`action`) being "pivoted" to the columns and the values for each cell coming from the summarized count (`n`)
④ Calculate the total counts across both the rows and the columns

```
 seeding NoYawn Yawned Total
 Control     13      3    16
  Seeded     23     11    34
   Total     36     14    50
```

3. Find the proportion that yawn in the Seeded group.

4. Find the proportion that yawn in the Not Seeded group.

We could instead obtain our observed proportion that yawn (and did not yawn) for each group using `adorn_percentages()`.

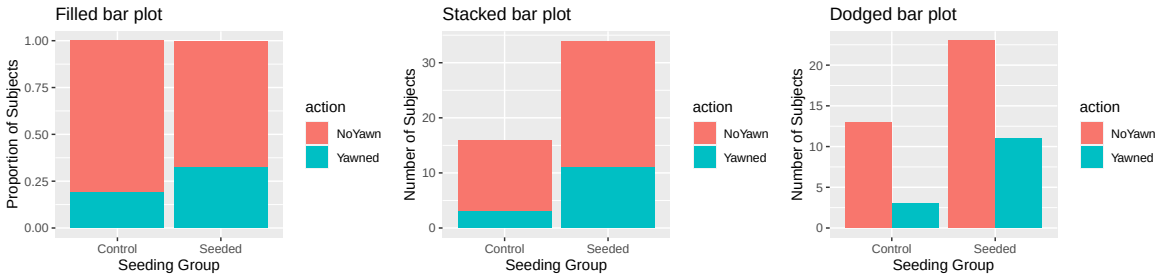
```
myth_busters |>
  count(seeding, action) |>
  pivot_wider(names_from = action,
              values_from = n) |>
  adorn_totals(where = c("row", "col")) |>
  adorn_percentages(denominator = "row")                    ⑤
```

⑤ Calculate the observed proportions based on the "row" totals (number of subjects in each seeding group).

```
seeding    NoYawn    Yawned Total
Control 0.8125000 0.1875000     1
 Seeded 0.6764706 0.3235294     1
  Total 0.7200000 0.2800000     1
```

```
ggplot(data = myth_busters,                                 ①
       mapping = aes(x = seeding,                           ②
                     fill = action)) +                      ③
  geom_bar(position = "fill") +                             ④
  labs(title = "Filled bar plot",                           ⑤
       x = "Seeding Group",
       y = "Proportion of Subjects"
       )
```

① Tell your plot which data set to get the information from.
② Tell the plot which variable you want on the x-axis (typically explanatory)
③ Tell the plot which variable you want to color by (typically response)
④ Create the bars and indicate "fill", "stack", or "dodge"
⑤ Provide appropriate plot titles and axis labels



Chi-square Test Statistic

The above descriptive analysis tells us what we have learned about the 50 subjects in the study. Can we make any inferences beyond what happened in the study (i.e., general statements about the population)? Similar to what we did last week with one categorical variable, we will be performing a Chi-Squared test to compare what we observed in the data to what we would have expected to see under the assumption that yawn seeding has no relationship with yawn action (if the null hypothesis is true).

5. Write the parameters of interest for this study.

6. Write the null and alternative hypotheses for this study. Recall, the Chi-square procedure can be used for ONLY TWO-SIDED TESTS!

Expected Counts under H_0

The Chi-Squared statistic (X^2) has exactly the same formula to what we used last week. The only aspect that changes is how we get each cell's expected count.
To find the expected count for each cell in our two variable table, we use three pieces of information:

- the row total for that cell (denoted with an i)
- the column total for that cell (denoted with a j)
- the total sample size

We find the expected value of a cell using the following formula:

$$\frac{\text{row}_i \text{ total} \times \text{column}_j \text{ total}}{\text{total sample size}}.$$

Once we have each of these, we can create a table of what frequencies we would have expected under the assumption there is not relationship with yawn action (if H_0 was true).

6. Complete the table of Expected counts.

Observed counts

seeding	NoYawn	Yawned	Total
Control	13	3	16
Seeded	23	11	34
Total	36	14	50

Expected counts under assumption null is true (no relationship between seeding and action)

	No yawn	Yawned	Total
Control			36
Seeded			34
Total	36	14	50

Chi-Square Test Statistic Next, we compare each of our observed counts to what we would have expected if H_0 was true. We compare how far "off" our observed frequencies are from what was expected using the same formula we saw in Chapter 3.

$$\text{Test Statistic} = \sum \frac{(\text{Observed - Expected})^2}{\text{Expected}}$$

Using the formula above, we can calculate how far "off" each of the cells in our observed table is from what we expected under the assumption there is no relationship between seeding and yawn action.

$$X^2 = \frac{(13-11.52)^2}{11.52} + \frac{(3-4.48)^2}{4.48} + \frac{(23-24.48)^2}{24.48} + \frac{(11-9.52)^2}{9.52} = 0.999$$

Conducting a Simulation Study

We will answer the research question by replicating the experiment over and over again, but in a situation where we assume that yawn seeding has no relationship with yawn action (the null model). We'll start with 14 yawners and 36 non-yawners, and we'll randomly assign 34 of these 50 subjects to the seeded group and the remaining 16 to the non-seeded group.

6. Note that we could use cards to replicate this experiment:

Step 1: Write the action (yawned/ did not yawn) and the seeding (seeded with a yawn / not seeded with a yawn) on _____ cards.

Step 2: Assume the null hypothesis is true (no relationship) and _____

Step 3: Create a new dataset that could have happened if H_0 was true by _____

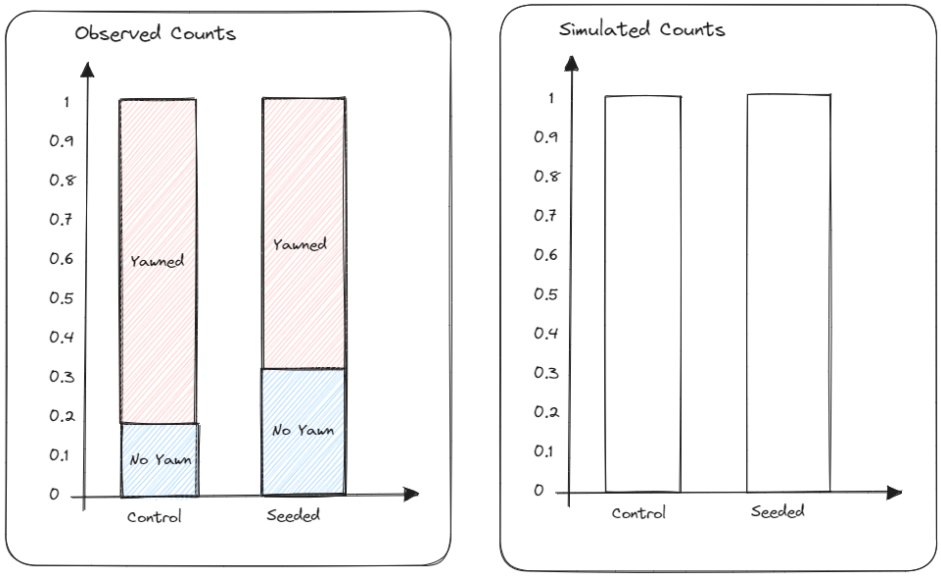
Step 4: Construct the contingency table of your simulated counts from the shuffled/randomly matched cards to show the number of yawners and non-yawners in each seeding group.

Keep in mind the column totals and the row totals will stay the same (there were only 14 people who yawned, and only 34 people who were seeded with a yawn). However, the cell values will change (we won't always have 10 individuals who Yawned from being seeded with a yawn).

Simulated Counts

	No yawn	Yawned	Total
Control			36
Seeded			34
Total	36	14	50

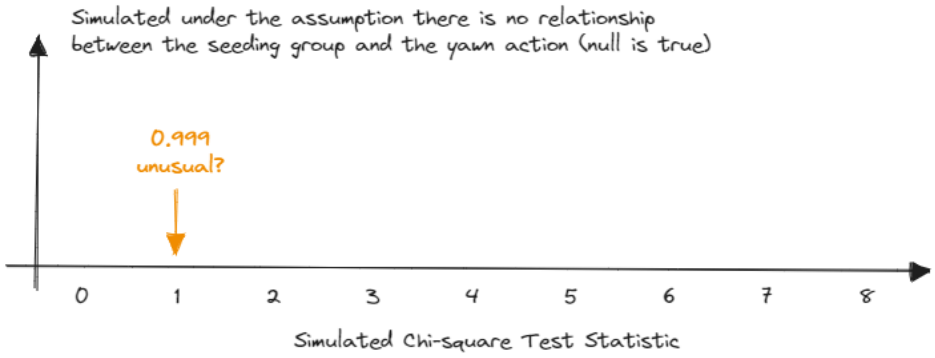
Divide your simulated counts by the row totals to give you the proportion of individuals who yawned and did not yawn for each seeding group. Sketch these in the second plot below. How do your simulated proportions compare to the observed proportions?



Step 5: Calculate the X^2 text statistic for your simulated counts under the assumption there is no relationship between seeding group and yawn action.

Recall our expected counts will be the same for every simulation (and the same as for our observed test statistic calculation.

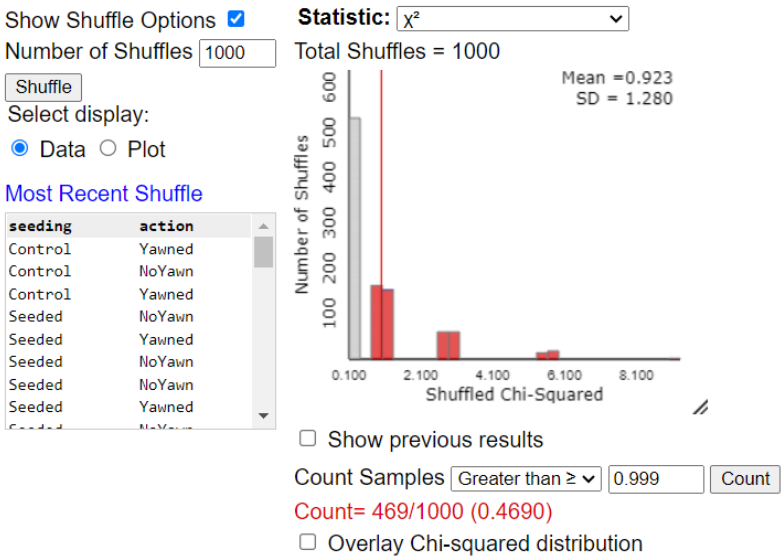
Step 6: Plot the simulated X^2 test statistic on the dot plot below. Add your group member's dots.



7. How does the observed test statistic compare to the simulated test statistics by your group members? Does this provide evidence for the alternative? Explain.

We can conduct a large scale simulation using `Online Simulation Applets` > `Two-way Tables`.

- Clear > Copy/Paste in the data <>
- Click `Use Data`
- Click `Show Shuffle Options`
- Change `Statistic` to X^2.
- Change `Number of Shuffles` to 1000, and hit `Shuffle`



10. What does each "dot" on the above plot represent?

11. How often did we see results at least as extreme as the observed test statistic under assumption the null is true? Estimate the p-value from the simulation. What decision would you make about the research question?

12. State your conclusion in terms of the research question.

Using The Chi-Square Distribution

Recall the Chi-square distribution can be used to approximate our simulated sampling distribution of the test statistic to test for a relationship between the explanatory and response variables. The Chi-square distribution takes on only positive numbers. In addition, this distribution is indexed by its degrees of freedom (or df).

> **i** Degrees of Freedom (df) for the Chi-Square Test:
>
> For this test, this is given by df = (rows - 1)(columns - 1).

In other words, when the null hypothesis is true, the test-statistic follows the chi-square distribution with df = (rows - 1)(columns - 1).

The R code below would be used to conduct a Chi-square Test for the Myth Busters study:

```
library(infer)
chisq_test(x = myth_busters,
           response = action,
           explanatory = seeding,
           correct = FALSE
           )
```

```
Warning in stats::chisq.test(table(x), ...): Chi-squared approximation may be
incorrect
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <int>   <dbl>
1     0.999        1   0.318
```

Conditions for conducting a Chi-square Test:

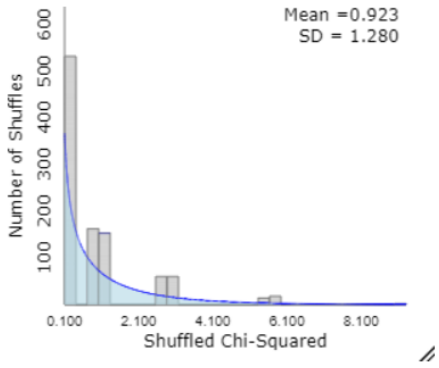
In order for the χ^2 distribution to be a good approximation of the sampling distribution under the assumption the null is true, we need to verify two conditions:

- The observations are independent
- We have a "large enough" sample size
  - This is checked by verifying there are at least 5 expected counts in each category

If the condition about expected cell counts is violated, we are forced to use the simulation-based method.

13. Check the conditions for using a Chi-square distribution to analyze the Myth Buster's study. Is it appropriate to conduct a Chi-square Test?

Notice the plot below shows the Chi-square distribution is not a very good approximation of our simulated sampling distribution of our test statistic. Therefore, we want to be cautious about using the theory-based Chi-square Test.



Example 4.2: Vested Interest and Task Performance

This example is from Investigating Statistical Concepts, Applications, and Methods by Beth Chance and Allan Rossman. 2006. Thomson-Brooks/Cole.

"A study published in the Journal of Personality and Social Psychology (Butler and Baumeister, 1998) investigated a conjecture that having an observer with a vested interest would

decrease subjects' performance on a skill-based task. Subjects were given time to practice playing a video game that required them to navigate an obstacle course as quickly as possible. They were then told to play the game one final time with an observer present. Subjects were randomly assigned to one of two groups. One group (A) was told that the participant and observer would each win $3 if the participant beat a certain threshold time, and the other group (B) was told only that the participant would win the prize if the threshold were beaten. The threshold was chosen to be a time that they beat in 30% of their practice turns. The following results are very similar to those found in the experiment: 3 of the 12 subjects in group A beat the threshold, and 8 of 12 subjects in group B achieved success."

	A: Vested Interest	B. No Vested Interest	Total
Achieved Success	3	8	11
Did not achieve success	9	4	13
Total	12	12	24

Research Question: Does whether the observer has a vested interest or not have an impact on performance on a skill-based task?

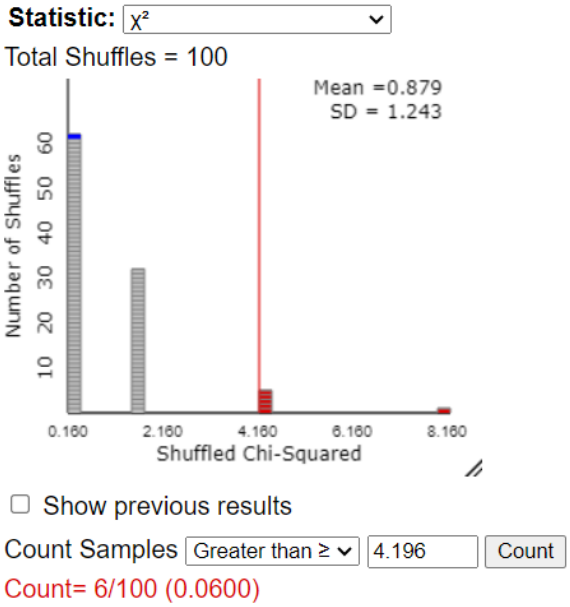
1. What are the variables in the study? Are they categorical or numerical?

2. Which is the response variable? Which is the explanatory variable?

3. Convert the research question into your null and alternative hypotheses.

4. Is it appropriate to conduct a Chi-square Test or do we need to use simulation? *Hint: For the large sample size condition, you only need to calculate the Expected Count for the row and column combination with the lowest totals, not all cells.*

5. Using the simulation output below, determine the p-value. Make a decision concerning the null hypothesis.



6. Write a conclusion in terms of the original research question.

Example 4.3 High-Salt Diets and Cardiovascular Disease

Suppose a retrospective study is carried out among men aged 50-54 in a specific county who died over a one-month period. The investigators try to include approximately an equal number of men who died from CVD (the cases) and men who died from other causes (the controls). Of 15 people who died from CVD, 4 were on a high-salt diet before they died. In contrast, of 15 people who died from other causes, only 1 was on a high-salt diet. The data are shown in the following contingency table:

	High-Salt	Low-Salt	Total
Non-CVD	1	14	15
CVD	4	11	15
Total	5	25	30

Research Question: Is there an association between the cause of death and salt diet? In other words, is the proprtion of the CVD group which has a high-salt diet different from the proportion of the Non-CVD group which has a high-salt diet?

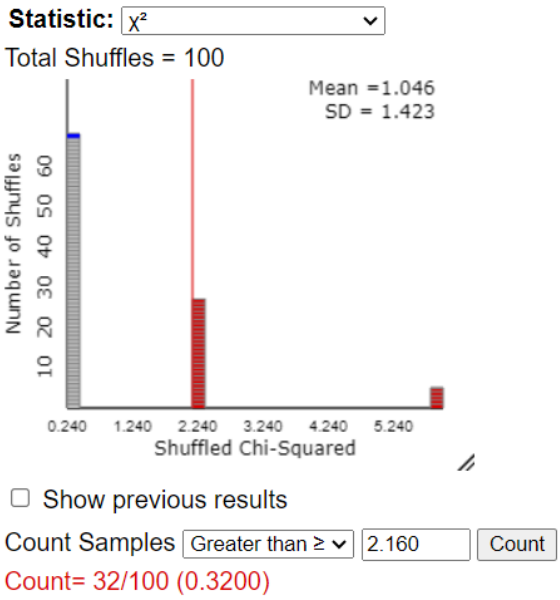
1. What are the variables in the study? Are they categorical or numerical?

2. Which is the response variable? Which is the explanatory variable?

3. Convert the research question into your null and alternative hypotheses.

4. Is it appropriate to conduct a Chi-square Test or do we need to use simulation?

5. Using the simulation output below, determine the p-value. Make a decision concerning the null hypothesis.



6. Write a conclusion in terms of the original research question.

Observational Studies vs. Designed Experiments

Reconsider the "Vested Interest and Task Performance" example. Fisher's exact test provided evidence that the proportion of successes was in fact smaller for the vested interest group (p-value = .0498). Now, the question is this: can we conclude that having a vested interest really is the cause of the decreased performance? Similarly, consider the "High-Salt Diet and Cardiovascular Disease" example. If this would have yielded a statistically significant result, would it have been fair to conclude that high-salt diets cause cardiovascular disease based on this study? The answer to these questions lies in how the data were collected; i.e., we must first determine whether the study was a designed experiment or an observational study.

Observational Study vs Designed Experiment

- An **observational study** involves collecting and analyzing data *without randomly assigning treatments* to observations.
- On the other hand, in a **designed experiment**, a *treatment is randomly assigned* on individual observations in order to observe whether the treatment *causes* a change in the response.

Key statistical idea:

Observational studies establish only that an **association exists** between the variables under study. With observational studies, it is always possible that there are other lurking variables not controlled for in the study that may be impacting the response. Since we can't be certain these other factors are balanced out between treatment groups, it is possible that these other factors could explain the difference between treatment groups. With **designed experiments**, there still may be lurking variables not specifically controlled for in the study; however, the **random assignment** of treatments used by researchers in a designed experiment should balance out between the treatment groups any other factors that might be related to the response variable. In a sense, we control for these other lurking variables by balancing their effects between the two groups via random assignment. Therefore, designed experiments can be used to establish a **cause-and-effect** relationship if the results are significant.

Note that the *"Vested Interest and Task Performance"* study is an example of a designed experiment since participants were randomly assigned to the two groups. We were trying to show that having a vested interest caused a decreased task performance. The small p-value rules out observing the decreased performance in the vested interest group simply by chance, and the randomization of subjects to treatment groups should have balanced out any other factors that might explain the difference. So, the most likely explanation left is that having a vested interest really does decrease task performance.

In the *"High-salt and Cardiovascular Disease"* study, however, there could be other factors that explain the outcome. For example, it's reasonable to assume that those with high-salt diets are less health conscious and tend to exercise less. Maybe the proportion with CVD is higher in this group because they exercise less!

Example 4.4: Alcoholism and Depression

Past research has suggested a high rate of alcoholism in families among patients with primary unipolar depression. A study of 210 families of females with primary unipolar depression found that 89 families had alcoholism present. A set of 299 control families found 94 present.

Research Question: Does the proportion of families afflicted by alcoholism differ between those families in which the female has primary unipolar depression and the control group? That is, is there a relationship between unipolar depression in females and alocholism in the family?

group	No	Yes	Total
Control	205	94	299
Depression	121	89	210
Total	326	183	509

1. What are the variables in the study? Are they categorical or numerical?

2. Which is the response variable? Which is the explanatory variable?

3. Identify the proportion of families afflicted by Alcoholism in both groups.

4. Convert the research question into your null and alternative hypotheses.

5. Is it appropriate to conduct a Chi-square Test or do we need to use simulation?

6. Is there evidence there a relationship between unipolar depression in females and alcholism in the family? Use the R output to answer this question. Make sure to use the observed test statistic, df, and p-value in your conclusion.

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <int>   <dbl>
1      6.42        1  0.0113
```

7. Can we say that having unipolar depression *causes* alcoholism? Explain your reasoning.