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*.

i Note

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

The *explanatory variable* is the variable we think is "explaining" the change in the response variable and the *response variable* is the variable we think is being impacted or changed by the explanatory variable. The explanatory variable is sometimes called the independent variable and the response variable is sometimes called the dependent variable.

1. What are the variables in the MythBusters Yawning experiment? Are they categorical or numerical? Which is the response variable? Which is the explanatory variable?

i Descriptive Methods for Two Categorical Variables:

- 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. The x-axis of the graph denotes the categories of the *explanatory variable*, and the *fill* color denotes the categories of the *response variable*. We can create three types of bar plots – filled (my fav!), stacked, or dodged.

- A **contingency table** shows the joint counts (aka frequencies) of two categorical variables. The *rows* of the table denote the categories of the *response variable*, and the *columns* denote the categories of the *explanatory variable*.

```
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 `position =` – "fill", "stack", or "dodge".
⑤ Provide appropriate plot titles and axis labels.

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.

```
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
```

It is easier to understand the relationship between the explanatory and response variables if we arrange our counts into a contingency table. This displays our *observed counts*:

```
library(janitor)                                                    ①

myth_busters |>                                                     ②
  tabyl(action, seeding) |>                                         ③
  adorn_totals(where = c("row", "col"))                             ④
```

① Load the `janitor` package to access specific functions.
② Start with the `myth_busters` data set.
③ Use the `tabyl` function to count the number of subjects in each yawn action and seeding group combination, and arrange in a contingency table format.
④ Use the `adorn_totals()` function to add total counts to the contingency table.

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

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

$$\hat{p}_{yawned|control} =$$

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

$$\hat{p}_{yawned|seeded} =$$

> **Note**
>
> Two variables are associated or related if the value of one variable gives you information about the value of the other variable. When comparing two groups this means that the proportions or means take different values in the two groups.

3. Do the two variables (seeded group and yawn action) seem to be related?

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

```
myth_busters |>
  tabyl(action, seeding) |>
  adorn_totals(where = c("row", "col")) |>
  adorn_percentages(denominator = "col")                    ⑤
```

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

```
 action Control    Seeded Total
 NoYawn  0.8125 0.6764706  0.72
 Yawned  0.1875 0.3235294  0.28
  Total  1.0000 1.0000000  1.00
```

Chi-square Test of Independence

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 in Chapter 3 with the Chi-square Goodness of Fit Test for one categorical variable with more than two groups, we will be simulating and calculating a Chi-Squared Test Statistic 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 (i.e., if the null hypothesis is true).

4. Write the null and alternative hypotheses for this study.

Expected Counts under the assumption the null is true

The Chi-Squared Test Statistic (X^2) has exactly the same formula to what we used last week.

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

The only aspect that changes is how we get each cell's expected count. To find the *expected count* for each cell in our contingency table:

$$\frac{\text{row total}}{\text{total sample size}} \times \text{col total.}$$

The $\frac{\text{row total}}{\text{total sample size}}$ gives you the proportion of Response 1 (e.g., No yawn) in the entire sample regardless of explanatory variable group (e.g., regardless of Control or Seeded). Then assuming that is the expected probability of Outcome 1, we can distribute out the n in each group to get the expected count (under the null) by multiplying by the col total.

Once we have each of these, we can create a table of what counts we would have expected *under* the assumption there is no relationship between seeding group and yawn action (i.e., if the null is true).

5. Complete the table of Expected counts.

Observed counts

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

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

	Control	Seeded	Total
No yawn			36
Yawned			14
Total	16	34	50

Chi-Square Test Statistic Next, we compare each of our observed counts to what we would have 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 with Two Variables

We will answer the research question by replicating the experiment over and over again, but *under* the assumption that yawn seeding has no relationship with yawn action (i.e., the null is true). 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.

Step 1: Write the action (yawned / no yawn) and the seeding (seeded / control) on _____ cards.

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

_____.

Step 3: Create a new data set that could have happened if H_0 was true by

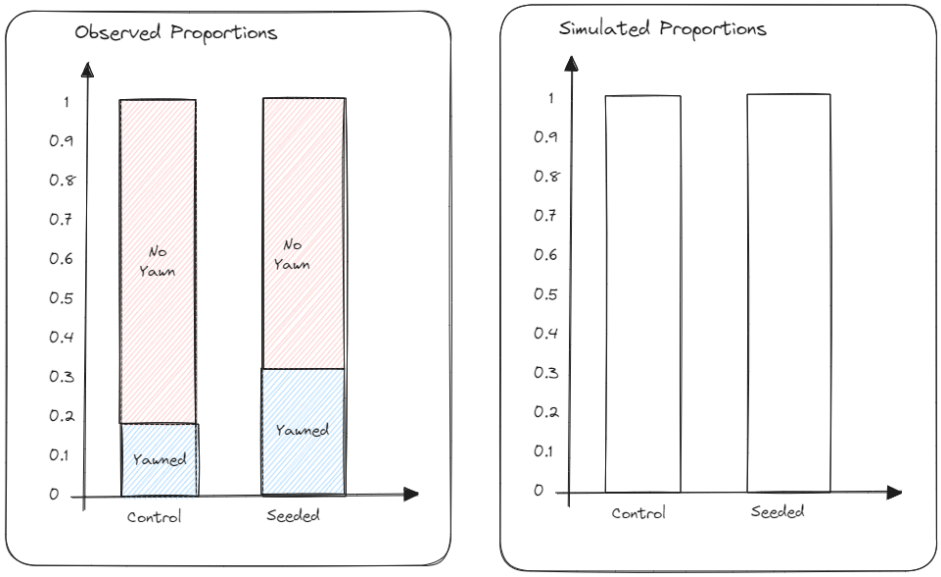
_____.

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

Simulated Counts

	Control	Seeded	Total
No yawn			36
Yawned			14
Total	16	34	50

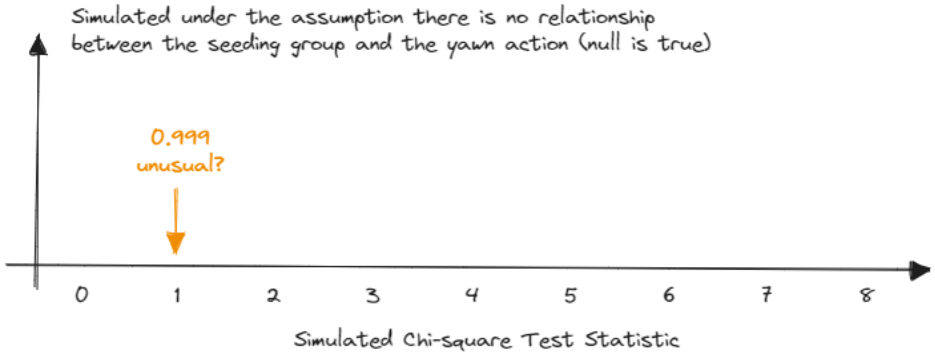
6. Divide your simulated counts by the column totals to give you the proportion of individuals who yawned and did not yawn for each seeding group. Sketch these in the second plot below. How do your simulated proportions compare to the observed proportions?



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).

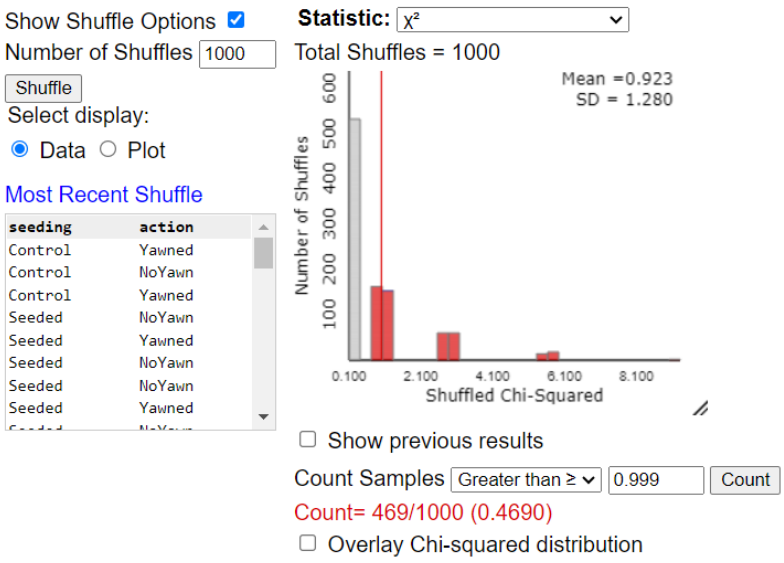
Step 6: Plot the simulated X^2 test statistic on the dot plot below.



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

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

- Clear > Copy/Paste in the data https://raw.githubusercontent.com/earobinson95/stat218-calpoly/refs/heads/main/01-course-notes/data/myth-yawns.csv
- Click `Use Data`
- Click `Show Shuffle Options`
- Change `Statistic` to X^2.
- Change `Number of Shuffles` to 1000, and hit `Shuffle`



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

9. 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?

10. 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 null 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 of Independence:
>
> 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)                                    ①
chisq_test(x = myth_busters,                      ②
           explanatory = seeding,                 ③
           response = action,                     ④
           correct = FALSE                        ⑤
           )
```

① Load the `infer` package to access specific functions.
② Use the `chisq_test()` function and denote the data set `x =`.
③ Specify the `explanatory =` variable name,
④ and the `response =` variable name.
⑤ We will not talk about corrections, but the default is `TRUE` and we want `FALSE`.

```
# 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:

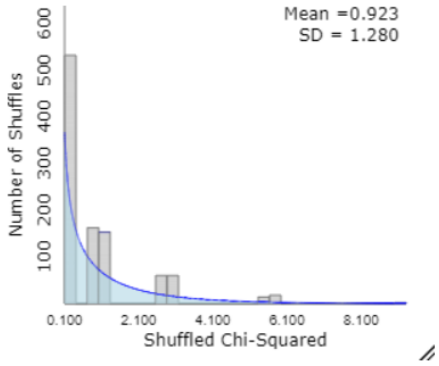
In order for the χ^2 distribution to be a good approximation of the null sampling distribution, we need to verify two conditions:

- Independent observations
- 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.

11. 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.




```
set.seed(93401)
library(infer)
chisq_test(x = myth_busters,
           explanatory = seeding,
```

```
            response = action,
            correct = FALSE,
            simulate.p.value = TRUE                                    ⑥
            )
```

⑥ In order to conduct a *simulation* instead, specify `simulate.p.value = TRUE`. Notice we no longer have `chisq_df`.

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl> <lgl>      <dbl>
1     0.999 NA         0.498
```

Observational Studies vs. Designed Experiments

Recall the concept of a *confounding variable*, a characteristic other than the variable of interest that may be related to the outcome.

Definitions

- An **observational study** involves collecting and analyzing data without manipulating or randomly assigning treatments. The data exists naturally, and we can only identify *associations* between variables.
- A **designed experiment**, however, involves randomly assigning treatments or groups to individuals to investigate whether the treatment *causes* a change in the response variable. The study is manipulated to control other variables.

Key statistical idea:

In **observational studies**, we can only establish that an *association* exists between the explanatory and response variables. This is because other uncontrolled factors, or confounding variables, could influence the response. Since these factors aren't balanced between groups, they might explain the observed differences.

In **designed experiments**, *random assignment* helps balance confounding variables across treatment groups, controlling for their influence. Therefore, designed experiments can provide evidence of a *cause-and-effect* relationship.

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 (Vested Interest) was told that the participant and observer would each win $3 if the participant beat a certain threshold time, and the other group (No Vested Interest) 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."

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

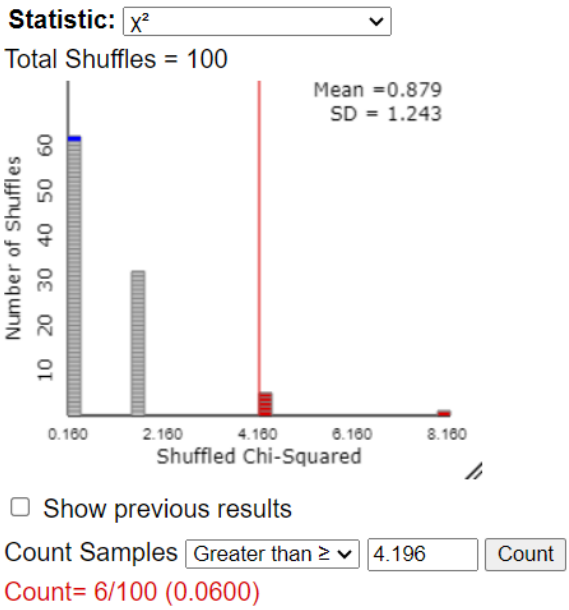
Research Question: Does whether the observer has a vested interest or not have an impact on performance on a skill-based task?

1. What are the variables in the study? Are they categorical or numerical? Which is the explanatory variable? Which is the response variable?

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

3. Is it appropriate to use the Chi-square distribution or do we need to use simulation?

4. Using the observed $X^2 = 4.196$ and the simulation output below, determine the p-value. Make a decision concerning the null hypothesis.



5. At an $\alpha = 0.10$, write a conclusion in terms of the original research question. Make sure to include your evidence in your conclusion.

6. Suppose all 12 individuals in the Vested Interest group hold high scores on Mario Cart and all 12 individuals in the No Vested Interest group are new to playing video games. Why would this be problematic?

7. The scenario states that subject's were *randomly* assigned to one of two groups. How does this ensure we are actually evaluating how vested interest is related to the success of the video game rather than say skill?

8. If we determined that having a vested interest does have a relationship with achieving an outcome or not, does this allow us to draw a cause-and-effect conclusion between vested interest and success?

Example 4.3: Home Court Advantage

Sports teams prefer to play in front of their own fans rather than at the opposing team's site. Having a sell-out crowd should provide even more excitement and lead to an even better performance, right? Well, consider the Oklahoma City Thunder, a National Basketball Association (NBA) team, in its second season (2008–2009) after moving from Seattle. This team had a win–loss record that was actually worse for home games with a sell-out crowd (3 wins and 15 losses) than for home games without a sell-out crowd (12 wins and 11 losses). (These data were noted in the April 20, 2009, issue of Sports Illustrated in the Go Figure column.)

1. Identify the observational unit and variables in this study. Also classify each variable as categorical (specify levels) or numeric.

- Observational Unit:

- Explanatory Variable:

- Response Variable:

2. Complete the observed counts in the contingency table below:

	Sell-out Crowd	Smaller Crowd	Total
Win			
Loss			
Total			

3. When did the Thunder have a higher winning percentage: in front of a sell-out crowd or a smaller crowd? Support your answer by calculating the proportion of sell-out games that they won and also the proportion of non-sell-out games that they won.

$$\hat{p}_{win|sell-out} =$$

$$\hat{p}_{win|smaller} =$$

4. Do the two variables appear to be associated? Explain.

There are two possible explanations for this odd finding that the team had a better winning percentage with smaller crowds: + The sell-out crowd caused the Thunder to play worse, perhaps because of pressure or nervousness. + The sell-out crowd did not cause a worse performance, and some other issue (variable) explains why they had a worse winning percentage with sell-out crowds. In other words, a third variable is at play, which is related to both the crowd size and the game outcome.

(Of course, another explanation is random chance, but as we saw with the Chi-Square Test of Independence, we can test for this and will later rule out random chance in this case.)

5. Consider the second explanation. Suggest a plausible alternative variable that would explain why the team would be less likely to win in front of a sell-out crowd than in front of a smaller crowd. (Make sure it's clear not just that your explanation would affect the team's likelihood of winning but also that the team would be less likely to win in front of a sell-out crowd compared to a smaller crowd.)

Definition

A **confounding variable** is a variable that is related both to the explanatory and to the response variable in such a way that its effects on the response variable cannot be separated from the effects of the explanatory variable.

6. Identify the confounding variable based on your suggested explanation above. Explain how it is confounding: what is the link between this third variable and the response variable, and what is the link between this third variable and the explanatory variable? (Hint: Remember that this variable has to be recorded on the observational units: home games for the Thunder (i.e., legit variable).)

Another variable recorded for these data was whether the opponent had a winning record the previous season. Of the Thunder's 41 home games, 22 were against teams that won more than half of their games. Let's refer to those 22 teams as "strong opponents". Of these 22 games, 13 were sell-outs. Of the 19 games against opponents that won less than half of their games that season (weak opponents), only 5 of those games were sell-outs.

7. Was the Thunder more likely to have a sell-out crowd against a strong opponent or a weak opponent? Calculate the relevant (conditional) proportions to support your answer.

$$\hat{p}_{sell-out|strong} =$$

$$\hat{p}_{sell-out|weak} =$$

When the Thunder played a strong opponent, they won only 4 of 22 games. When they played a weak opponent, the Thunder won 11 of 19 games.

8. Was the Thunder less likely to win against a strong opponent than a weak one? Again calculate the relevant (conditional) proportions to support your answer.

$$\hat{p}_{win|strong} =$$

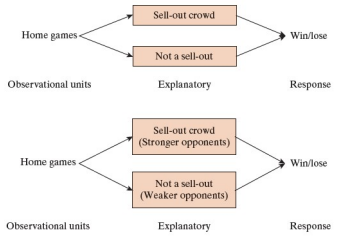
$$\hat{p}_{win|weak} =$$

9. Explain how your answers to the previous two questions establish that strength of opponent is a confounding variable that prevents drawing a cause-and-effect conclusion between crowd size and game outcome.

10. Although we did not carry out a Chi-square test of independence, summarize your conclusion about whether these data provide evidence that a sell-out crowd caused the Thunder to play worse. Write as if to a friend who has never studied statistics. Be sure to address the fact that the Thunder had a much smaller winning percentage in front of a sell-out crowd.

Key statistical idea

Confounding explains why you cannot draw a cause-and-effect conclusion from association alone: The groups defined by the explanatory variable could differ in more ways than just the explanatory variable when confounding is present. The diagram below illustrates the confounding in the Thunder study. The top panel shows the study design: Observational units (home games) are sorted into groups according to the explanatory variable (whether the arena was sold out). Then the response (win/lose) was observed. The bottom panel shows the confounding: sell-out crowds tended to be against stronger opponents; weaker opponents tended not to sell out the arena.



Key statistical idea: Scope of Inference

The **scope of inference** in a study refers to what conclusions can be drawn based on how the data were collected. If a study uses random assignment, it allows for causal conclusions because differences in outcomes can be attributed to the treatment or group. However, random sampling allows for results to be generalized to the broader population. Without random assignment, we can only infer associations, not causation, and without random sampling, the findings can only be applied to the specific sample studied, limiting generalizability.