Describing Data: Categorical Variables

Sections 2.1

- One categorical variable
- Two categorical variables

Question of the Day

Is cat ownership related to Schizophrenia?

Toxoplasmosis

- Toxoplasmosis is a disease caused by the protozoan parasite Toxoplasma gondii ("toxo"), one of the world's most common parasites
- Humans can get toxoplasmosis in several ways:
  1. Food (undercooked and contaminated meat)
  2. Animals (contact with infected cat feces)
  3. Mother-to-child transmission (placenta)
  4. Blood transfusion/organ transplantation (rare)
- Toxoplasmosis may cause flu-like symptoms and is dangerous in compromised immune systems, but most infected people are asymptomatic (physically)

Data

- How prevalent in Toxoplasmosis?
- Data from NHANES 1999-2000
- 4234 people aged 6 – 49 were tested for Toxoplasmosis.
- Of these, 605 people were infected and 3629 were not.
- One categorical variable.

Frequency Table

- A frequency table shows the number of cases that fall in each category:

<table>
<thead>
<tr>
<th>Infected</th>
<th>Not Infected</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>605</td>
<td>3629</td>
<td>4234</td>
</tr>
</tbody>
</table>

- For one categorical variable, this summarizes all the information in the data

Bar Chart/Plot/Graph

- In a bar chart, the height of the bar is the number of cases falling in each category

**Histogram vs Bar Chart**

- A bar chart is for categorical data, and the x-axis has no numeric scale.
- A histogram is for quantitative data, and the x-axis is numeric.
- For a categorical variable, the number of bars equals the number of categories, and the number in each category is fixed.
- For a quantitative variable, the number of bars in a histogram is up to you (or your software), and the appearance can differ with different number of bars.

**Proportion**

- The proportion in a category is found by the number in that category divided by the sample size.

- Proportion for a sample: \( \hat{p} \) ("p-hat")
- Proportion for a population: \( p \)

**Pie Chart**

- In a pie chart, the relative area of each slice of the pie corresponds to the proportion in each category.

**Proportion**

- What proportion of people in the sample were infected with Toxoplasmosis?

<table>
<thead>
<tr>
<th>Infected</th>
<th>Not Infected</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>605</td>
<td>3629</td>
<td>4234</td>
</tr>
</tbody>
</table>

- \( \hat{p} = \frac{605}{4234} = 0.143 \)
- A proportion of 0.14 is the same as 14%

**Relative Frequency Table**

- A relative frequency table shows the proportion of cases that fall in each category.

<table>
<thead>
<tr>
<th>Infected</th>
<th>Not Infected</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.143</td>
<td>0.857</td>
<td></td>
</tr>
</tbody>
</table>

- All the numbers in a relative frequency table sum to 1

**Toxoplasmosis**

- In the United States, the CDC estimates that 22.5% of the population 12 years and older have been infected with Toxoplasma. (CDC)

- Why the difference?
  - Adults only (12 and older different from 6 – 49)
  - Different sample, different years
  - Random chance

- In other places in the world, prevalence is as high as 95% in some populations
Summary: One Categorical Variable

- Summary Statistics
  - Frequency table
  - Relative frequency table
  - Proportion

- Visualization
  - Bar chart
  - Pie chart

Mind Controlling Parasite?

- Normal rats are terrified of cat pee, toxo-infected rats are drawn to the smell of it
- Toxo-infected rats are more active, and "less wary of predators in exposed spaces"
- Toxo-infected humans like cats more
- Toxoplasmosis is linked to delayed reaction times, and infected people are 2.5 times more likely to get in a car accident
- Toxoplasmosis is linked to Schizophrenia
- Lots more... see [this article](#) or google the topic.

Question of the Day

Is cat ownership related to Schizophrenia?

Case-Control Study

- A case-control study is an observational study in which cases are matched with controls
  - Cases are people with a specific disease or trait
  - Controls are people similar to the cases that do not have the disease or trait

- Case-control studies are useful for studying/identifying risk factors for rare diseases
- Can a case-control study be used to make conclusions about causality?
  a) Yes
  b) No

Cat Ownership and Schizophrenia

- Multiple case-control studies have been conducted to study the association between cat ownership and schizophrenia
- Cases were randomly selected from NAMI (National Alliance for the Mentally Ill), almost all of whom had Schizophrenia
- Controls (from families without mental illness) chosen:
  - 1992 Data: a family friend
  - 1997 Data: matched for age, sex, and socioeconomic status
  - 1982 Data (analyzed 2015): similar families from a survey conducted by the American Veterinary Medical Association


Two Categorical Variables

- Each of these studies recorded (among many others) two different variables:
  1. Case or control
  2. Whether or not the person had a cat in their house during childhood (birth to age 10 or 13)
- Is there a relationship between these two variables?
Side-by-Side Bar Chart (1992)

Side-by-Side Bar Chart (1997)

Side-by-Side Bar Chart (1982)

Segmented/Stacked Bar Chart (1982)

Two-Way Table

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat in house as child</td>
<td>84</td>
<td>65</td>
<td>149</td>
</tr>
<tr>
<td>No cat in house as child</td>
<td>81</td>
<td>100</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>165</td>
<td>330</td>
</tr>
</tbody>
</table>

- It doesn’t matter which variable is displayed in the rows and which in the columns
Cats and Schizophrenia

What proportion of people in this sample had a cat in the house during their childhood?

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat in house as child</td>
<td>84</td>
<td>65</td>
<td>149</td>
</tr>
<tr>
<td>No cat in house as child</td>
<td>81</td>
<td>100</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>165</td>
<td>330</td>
</tr>
</tbody>
</table>

What proportion of controls had a cat in the house during their childhood?

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat in house as child</td>
<td>84</td>
<td>65</td>
<td>149</td>
</tr>
<tr>
<td>No cat in house as child</td>
<td>81</td>
<td>100</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>165</td>
<td>330</td>
</tr>
</tbody>
</table>

1992 Data

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat in house as child</td>
<td>84</td>
<td>65</td>
<td>149</td>
</tr>
<tr>
<td>No cat in house as child</td>
<td>81</td>
<td>100</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>165</td>
<td>330</td>
</tr>
</tbody>
</table>

What proportion of cases had a cat in the house during their childhood?

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat in house as child</td>
<td>84</td>
<td>65</td>
<td>149</td>
</tr>
<tr>
<td>No cat in house as child</td>
<td>81</td>
<td>100</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>165</td>
<td>330</td>
</tr>
</tbody>
</table>

Difference in Proportions

A *difference in proportions*, is the difference in proportions for one categorical variable calculated for different levels of another categorical variable.

Example: proportion of cases who were cat owners as children – proportion of controls who had a cat in the house as children

\[
\hat{p}_{\text{cases}} - \hat{p}_{\text{controls}} = 0.51 - 0.34 = 0.17
\]

1997 Data

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat in house as child</td>
<td>136</td>
<td>220</td>
<td>356</td>
</tr>
<tr>
<td>No cat in house as child</td>
<td>126</td>
<td>302</td>
<td>428</td>
</tr>
<tr>
<td>Total</td>
<td>262</td>
<td>522</td>
<td>784</td>
</tr>
</tbody>
</table>

What is the difference in proportions (cases – controls) of having a cat in the house as a child?

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat in house as child</td>
<td>136</td>
<td>220</td>
<td>356</td>
</tr>
<tr>
<td>No cat in house as child</td>
<td>126</td>
<td>302</td>
<td>428</td>
</tr>
<tr>
<td>Total</td>
<td>262</td>
<td>522</td>
<td>784</td>
</tr>
</tbody>
</table>

Creating a Two-Way Table

In the 2015 study (1982 data), 1075 out of 2125 cases owned cats as children, and 2065 out of 4847 controls owned cats as children.

1. Create the two-way table.
2. Calculate the difference in proportions.
**Odds**

If \( p \) denotes the proportion, the **odds** are defined as

\[
\text{odds} = \frac{p}{1 - p}
\]

- **Interpreting odds**
  - Odds of 1 indicate 50/50
  - \( p < 0.5 \) yield odds < 1
  - \( p > 0.5 \) yield odds > 1
- Odds of 3, or 3:1, mean that out of 4 times, we would expect the variable to be in that category 3 times and out of that category 1 time

**Odds Ratio**

The **odds ratio (OR)** is the ratio of the odds in one group to the odds in the other group:

\[
OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}
\]

- Odds ratios of 1 indicate no difference between the groups (no relationship between the two variables)

**OR for Cats and Schizophrenia**

- Odds ratio for having a cat in the house as a child, comparing cases to controls (1982):

\[
OR = \frac{\text{odds of having a cat in the house for cases}}{\text{odds of having a cat in the house for controls}}
\]

\[
= \frac{0.506 / (1 - 0.506)}{0.426 / (1 - 0.426)} = 1.38
\]

**From the Paper**

<table>
<thead>
<tr>
<th>Year</th>
<th>Case</th>
<th>Controls</th>
<th>( p )</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>94/165</td>
<td>65/165</td>
<td>0.03</td>
<td>1.60 (1.03-2.53)</td>
</tr>
<tr>
<td>1997 survey</td>
<td>136/282</td>
<td>220/282</td>
<td>0.01</td>
<td>1.48 (1.08-2.02)</td>
</tr>
<tr>
<td>1992</td>
<td>1075/2125</td>
<td>2005/4847</td>
<td>0.001</td>
<td>1.36 (1.25-1.53)</td>
</tr>
</tbody>
</table>

\( p \) values are derived from chi square, 2 tailed. Odds shown as mean (95% CI).

**Summary: Two Categorical Variables**

- **Visualization**
  - Side-by-side bar chart
  - Segmented bar chart
- **Summary Statistics**
  - Two-way table
  - Difference in proportions
  - Odds ratio

**Real-Life Takeaways**

- Toxoplasmosis can have serious consequences
- Children and pregnant women should not be exposed to cat feces
- Only cats that hunt are susceptible, and can only transmit for the first three weeks
- Worried? Most people with Toxoplasmosis never develop Schizophrenia or mental illness
To Do

- Read Section 2.1
- Do HW 2.1 (due Friday, 9/25)