Day 22 - Intro to inference for multiple proportions

Learning outcomes

In this section, we explore how to assess the relationship between two categorical variables, regardless of how many levels each variable has.

Motivating example

How crucial is it to ask specific, pointed questions? In 2012, psychologists conducted an experiment to explore how the phrasing of questions influences the honesty of the responses.[1] In this study, 219 volunteers were tasked with selling an iPod that had shut off inexplicably twice in the past but otherwise functioned well. Each participant was told they would receive $10 plus 5% of the sale price. Participants were randomly assigned to receive a buyer from one of three groups: positive assumption, general, or negative assumption.

Buyers in the positive assumption group asked, "It doesn't have any problems, does it?" Those in the general group inquired, "What can you tell me about it?" Buyers in the negative assumption group asked, "What problems does it have?" For each sale, researchers recorded whether the seller disclosed the issues with the iPod.

The data for this study can be found in the `ask` data set within the `openintro` R package. The code below reads in the data.

```
library(openintro)
data(ask)
head(ask)
```

```
# A tibble: 6 x 3
  question_class question                                response
  <chr>          <chr>                                  <chr>
1 general        What can you tell me about it?         hide
2 pos_assumption It doesn't have any problems, does it? hide
3 pos_assumption It doesn't have any problems, does it? disclose
4 neg_assumption What problems does it have?            disclose
5 general        What can you tell me about it?         hide
6 neg_assumption What problems does it have?            disclose
```

[1] Link to a paper describing the study.

Visualizing the data

Create some tabular and visual summaries of the data and comment on whether they suggest there is an association between the response and the type of question asked.

The Chi-Square statistic

Let's focus on the contingency table produced above. What would you expect the table to look like if there was no association?

The X^2 statistic

Given some contingency table, the X^2 statistic may be computed in the following way:

$$X^2 = \sum_i \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2$$

where i indexes the elements in the table, O_i is the observed count, and E_i is the expected count. The quantity inside the parentheses,

$$r_i = \frac{O_i - E_i}{\sqrt{E_i}}$$

is called the **Pearson residual**, and describes how far away the observed count is from the expected, standardized by the expected count. To obtain the expected counts, we compute

$$E_i = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

We can use the `chisq.test` function to produce each component of the test statistic.

The Chi-square Test of Independence

The Chi-square test of independence

The _____ is used to determine whether two categorical variables are associated with one another when sampling from a single population. For this example, the hypotheses are as follows:

- H_0: there is no association between the type of question asked of the buyer and whether the buyer discloses the issues with the iPod at the population level
- H_a: there is an association between the type of question asked of the buyer and whether the buyer discloses the issues with the iPod at the population level

The data are summarized via the X^2 statistic.

$$X^2 = \sum_i \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2$$

Review: what are the steps for hypothesis test?

Simulating the variability of X^2 under H_0

How can we simulate the null hypothesis being true?

Computing the p-value and interpretting

If the null hypothesis is true, do we expect X^2 to be large or small?

Let's compute the p-value.

Interpret the p-value as a conditional probability in the context of the problem.

Interpret the p-value as strength of evidence in the context of the problem.

Now that we have determined that there is evidence of an association between the type of question asked and whether the issue is disclosed, what is a natural follow-up question?

Post-hoc visualizations

Let's explore the contribution of each category to the X^2 statistic.