Elementary Statistics – Inference for Categorical Data Pt. 2

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Kaitlyn Cook (https://www.smith.edu/people/kaitlyn-cook)

Plan for Today

- Inference for categorical variables

Warm Up: Inference for Two Proportions

Confidence Interval for Difference Between Two Proportions

Conditions for $\hat{p}_1 - \hat{p}_2$ to be approximated with the normal distribution:

1. Data are **independent within and between** the two **groups**
2. The **success-failure condition** is met in each group. To check, verify $(best\ guess\ for\ p_1)n_1 \geq 10, \big(1 - (best\ guess\ for\ p_1)\big)n_1 \geq 10,$ $(best\ guess\ for\ p_2)n_2 \geq 10, \big(1 - (best\ guess\ for\ p_2)\big)n_2 \geq 10$

When the conditions are met so that the distribution of $\hat{p}_1 - \hat{p}_2$ is nearly normal, variability of $\hat{p}_1 - \hat{p}_2$ is well described by:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Use SE to compute margin of error for our confidence interval:

$$\big((\hat{p}_1 - \hat{p}_2) - z^* SE, (\hat{p}_1 - \hat{p}_2) + z^* \mathrm{SE}\big)$$

You suspect different proportions of STEM and Humanities majors play sports. You perform an experiment to statistically test this suspicion. You sample 160 STEM majors and find 40 of them play sports. You sample 150 Humanities majors and find 70 of them play sports. Calculate a 95% CI for $p_{STEM} - P_{Humanities}$ from your $\hat{p}_{STEM}$ and $\hat{p}_{Humanities}$.

Warm Up: Inference for Two Proportions

Hypothesis Test for Difference Between Two Proportions

Conditions for $\hat{p}_1 - \hat{p}_2$ to be approximated with the normal distribution:

1. Data are **independent within and between** the two **groups**
2. The **success-failure condition** is met in each group. To check, verify $(best\ guess\ for\ p_1)n_1 \geq 10, \big(1 - (best\ guess\ for\ p_1)\big)n_1 \geq 10,$ $(best\ guess\ for\ p_2)n_2 \geq 10, \big(1 - (best\ guess\ for\ p_2)\big)n_2 \geq 10$

For a hypothesis test, we use $\hat{p}_{pool}$ as the best guess for p

$$\hat{p}_{pool} = \frac{number\ of\ successes}{number\ of\ cases} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

You suspect different proportions of STEM and Humanities majors play sports. You perform an experiment to statistically test this suspicion. You sample 160 STEM majors and find 40 of them play sports. You sample 150 Humanities majors and find 70 of them play sports. Use a hypothesis test with $\alpha = 0.05$ to see if there is a significant difference.

Inference for Categorical Variables

So far we have focused on questions like

- Is the proportion of WSU students who major in math more than 50%?
- Is the proportion of WSU students who major in math different than the proportion of Springfield College students who major in math?

Inference for Categorical Variables

So far we have focused on questions like

- Is the proportion of WSU students who major in math more than 50%?
- Is the proportion of WSU students who major in math different than the proportion of Springfield College students who major in math?

However, many research questions are posed in terms of categorical variables that have more than 2 levels. Ex.

- What is the comparative effectiveness of three drug treatments (drug A, drug B, drug C)?
- Has the distribution of educational attainment in the US changed since 2000?

Inference for Categorical Variables

Many research questions are posed in terms of categorical variables that have more than 2 levels. Ex.

- What is the comparative effectiveness of three drug treatments (drug A, drug B, drug C)?
- Has the distribution of educational attainment in the US changed since 2000?

In these cases, we want to understand joint behavior of all possible levels of our categorical variables. There is no longer a single population parameter of interest.

Inference for Categorical Variables

Many research questions are posed in terms of categorical variables that have more than 2 levels. Ex.

- What is the comparative effectiveness of three drug treatments (drug A, drug B, drug C)?
- Has the distribution of educational attainment in the US changed since 2000?

In these cases, we want to understand joint behavior of all possible levels of our categorical variables. There is no longer a single population parameter of interest.

- We typically do not construct CI's in this context
- We can still conduct hypothesis tests

Inference for Categorical Variables

Hypothesis Tests for Categorical Variables with more than 2 levels follow the same recipe as the binary examples we've looked at:

Pieces of a Hypothesis test:

1. ***Null and Alternative Hypotheses***
2. ***Test Statistic***
3. ***Null Distribution***
4. ***P-value***

Inference for Categorical Variables

Hypothesis Tests for Categorical Variables with more than 2 levels follow the same recipe as the binary examples we've looked at:

Pieces of a Hypothesis test:

1. ***Null and Alternative Hypotheses***
2. ***Test Statistic***
3. ***Null Distribution***
4. ***P-value***

However now, our test statistic is the Chi-square statistic (instead of $\hat{p}$)

and our null distribution will be modeled by the Chi distribution (instead of the normal distribution).

Inference for Categorical Variables

Motivating Example

As part of their 2017 "Pulse of the Nation" project, Cards Against Humanity conducted monthly polls examining American's social and political views. The following contingency table summarizes the political part affiliation and climate change beliefs of 1000 participants in the September 2017 poll:

	Views on Climate Change			
Political Party	Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat	16	276	35	327
Independent	58	287	80	425
Republican	76	92	80	248
Total	150	655	195	1000

Inference for Categorical Variables

Motivating Example

As part of their 2017 "Pulse of the Nation" project, Cards Against Humanity conducted monthly polls examining American's social and political views. The following contingency table summarizes the political part affiliation and climate change beliefs of 1000 participants in the September 2017 poll:

Observed

Political Party	Views on Climate Change: Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat	16	276	35	327
Independent	58	287	80	425
Republican	76	92	80	248
Total	150	655	195	1000

Suppose our research question is: Are political affiliation and belief in climate change independent of one another?

Inference for Categorical Variables

Motivating Example

Observed

Political Party	Views on Climate Change: Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat	16	276	35	327
Independent	58	287	80	425
Republican	76	92	80	248
Total	150	655	195	1000

Suppose our research question is: Are political affiliation and belief in climate change independent of one another?

Our **hypotheses** are:

H_0: individuals' climate change beliefs **are independent** of their political affiliation

H_A: individuals' climate change beliefs **are not independent** of their political affiliation.

Inference for Categorical Variables

Motivating Example

Observed

Political Party	Views on Climate Change: Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat	16	276	35	327
Independent	58	287	80	425
Republican	76	92	80	248
Total	150	655	195	1000

The **Test Statistic** will capture the degree of discrepancy between observed counts, and the counts we would expect if the null hypothesis (independence between variables) were true.

Inference for Categorical Variables

Motivating Example

Observed

Political Party	Views on Climate Change: Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat	16	276	35	327
Independent	58	287	80	425
Republican	76	92	80	248
Total	150	655	195	1000

If H_0 were true, how many of the 237 Democratic respondents would we expect to not believe in climate change?

Our **hypotheses** are:

H_0: individuals' climate change beliefs **are independent** of their political affiliation

H_A: individuals' climate change beliefs **are not independent** of their political affiliation.

Inference for Categorical Variables

Motivating Example

Observed

Political Party	Views on Climate Change: Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat	16	276	35	327
Independent	58	287	80	425
Republican	76	92	80	248
Total	150	655	195	1000

The **Test Statistic** will capture the degree of discrepancy between observed counts, and the counts we would expect if the null hypothesis (independence between variables) were true.

Inference for Categorical Variables

Motivating Example

Expected	Views on Climate Change			
Political Party	Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat				327
Independent				425
Republican				248
Total	150	655	195	1000

Test Statistic calculation

1. Generate expected number of observations for each cell of the contingency table if H_0 is true

$$E_{ij} = \frac{row\ i\ total \times column\ j\ total}{table\ total}$$

Inference for Categorical Variables

Motivating Example

Expected **Political Party**	**Views on Climate Change** Not Real at All	 Real and Caused by People	 Real and Not Caused by People	 Total
Democrat				327
Independent				425
Republican				248
Total	150	655	195	1000

Test Statistic calculation

1. Generate expected number of observations for each cell of the contingency table if H_0 is true

$$E_{ij} = \frac{row\ i\ total \times column\ j\ total}{table\ total}$$

Fill in the expected table.

Inference for Categorical Variables

Motivating Example

Observed

Political Party	Views on Climate Change: Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat	16	276	35	327
Independent	58	287	80	425
Republican	76	92	80	248
Total	150	655	195	1000

Test Statistic calculation

1. Generate expected number of observations for each cell of the contingency table if H_0 is true

$$E_{ij} = \frac{row\ i\ total \times column\ j\ total}{table\ total}$$

2. Calculate the chi-square statistic, X^2

$$X^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Inference for Categorical Variables

Motivating Example

Observed

Political Party	Views on Climate Change: Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat	16	276	35	327
Independent	58	287	80	425
Republican	76	92	80	248
Total	150	655	195	1000

Test Statistic calculation

1. Generate expected number of observations for each cell of the contingency table if H_0 is true

$$E_{ij} = \frac{row\ i\ total \times column\ j\ total}{table\ total}$$

2. Calculate the chi-square statistic, X^2

$$X^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Calculate the chi-square statistic

Inference for Categorical Variables

Motivating Example

Observed

Political Party	Views on Climate Change: Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat	16	276	35	327
Independent	58	287	80	425
Republican	76	92	80	248
Total	150	655	195	1000

Null Distribution

If these conditions are met:

1. Observations are independent
2. Each cell in the contingency table has an expected count ≥ 5

Then we use the Chi-square distribution with $(R-1)(C-1)$ degrees of freedom to model the null distribution of X^2

Inference for Categorical Variables

Motivating Example

Observed	Views on Climate Change			
Political Party	Not Real at All	Real and Caused by People	Real and Not Caused by People	Total
Democrat	16	276	35	327
Independent	58	287	80	425
Republican	76	92	80	248
Total	150	655	195	1000

Null Distribution

If these conditions are met:

1. Observations are independent
2. Each cell in the contingency table has an expected count ≥ 5

Then we use the Chi-square distribution with $(R-1)(C-1)$ degrees of freedom to model the null distribution of X^2

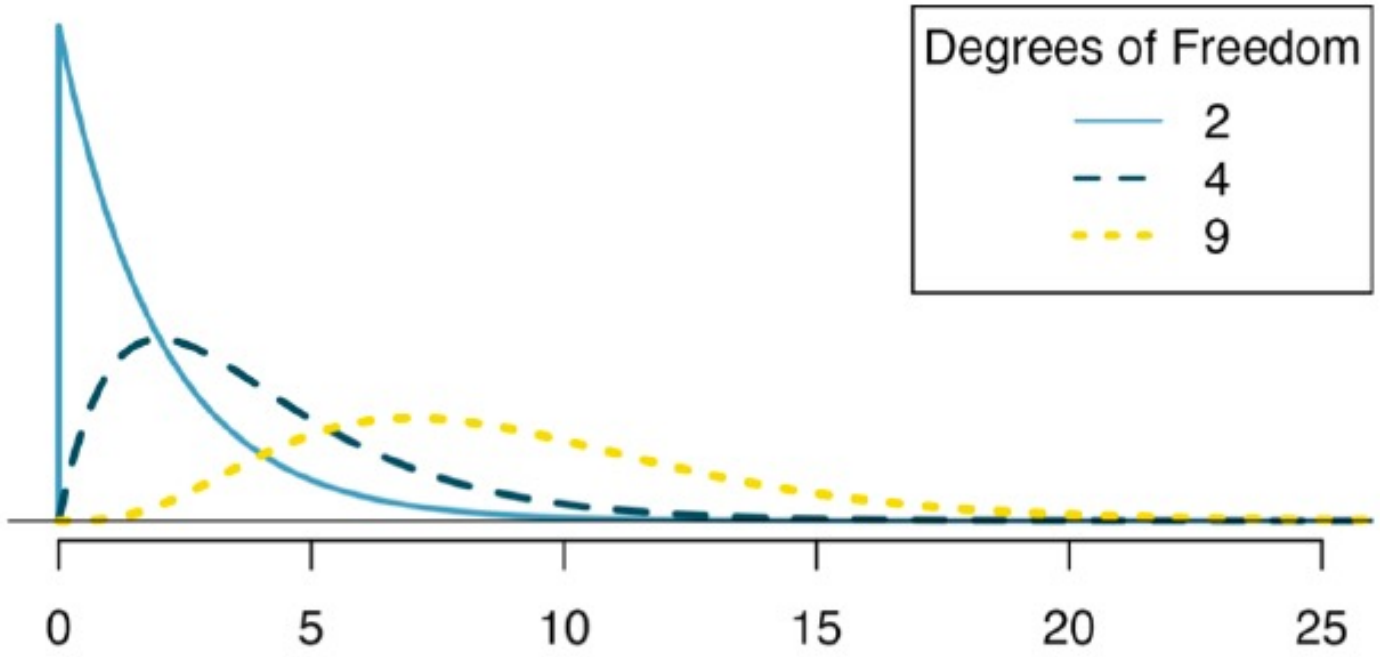
How many degrees of freedom do we have in this example?

Inference for Categorical Variables

Motivating Example

Null Distribution

The χ^2_k distribution is a right-skewed distribution whose shape and degree of skew are controlled by the degrees of freedom, k:

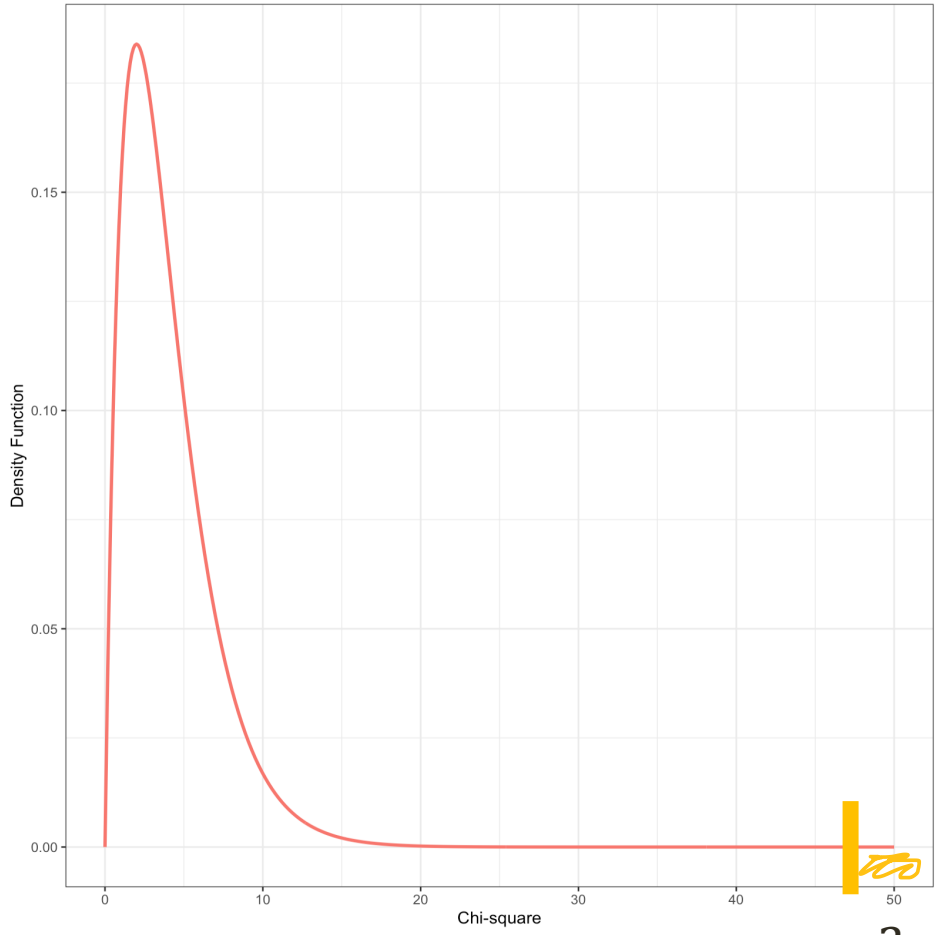


Inference for Categorical Variables

Motivating Example

P-value

We will measure how unusual our X^2 is, using the appropriate χ^2 distribution.



What is the p-value for the X^2 we calculated earlier?

If $\alpha = 0.05$, what is the conclusion of our test?

Inference for Categorical Variables

Practice

The table below summarizes the results of an experiment evaluating three treatments for Type 2 Diabetes in patients aged 10-17 who were being treated with metformin. The three treatments considered were continued treatment with metformin (met), treatment with metformin combined with rosiglitazone (rosi), or a lifestyle intervention program. Each patient have a primary outcome, which either lacked glycemic control (failure) or did not lack control (success). Perform a hypothesis test with $\alpha = 0.05$ to answer the question: Does treatment effect outcome?

Treatment	Failure	Success	Total
lifestyle	109	125	234
met	120	112	232
rosi	90	143	233
Total	319	380	699

Inference for Categorical Variables

Practice

A county health department enrolled 300 smokers in a randomized experiment. 150 participants were randomly assigned to a group that used a nicotine patch and met weekly with a support group; the other 150 received the patch and did not meet with a support group. At the end of the study, 40 of the participants in the patch plus support group had quit smoking while only 30 smokers had quit in the other group. Perform a hypothesis test with $\alpha = 0.05$ to answer the question: Does being part of a support group affect the ability of people to quit smoking?