Lecture 19

Categorical Data Analysis II

STAT 8010 Statistical Methods I
June 13, 2023

Whitney Huang
Clemson University

Agenda

1. Inference for Multi-Category Data

2. Analyzing Bivariate Categorical Data

Binomial Experiments and Inference for p

- Fixed number of n trials (sample size), each trial is an independent event (simple random sample)
- Binary outcomes ("success/failure"), where the probability of success, p, for each trial is constant
- The number of successes $X \sim \text{Bin}(n, p)$

We use a random sample x to infer p, the population proportion, using $\hat{p} = \frac{x}{n}$

Multinomial Experiments and Inference for $\boldsymbol{p} = (p_1, \cdots, p_K)$

- Fixed number of n trials, each trial is an independent event
- K possible outcomes, each with probability $p_k, k = 1, \cdots, K$ where $\sum_{k=1}^{K} p_k = 1$
- $(X_1, X_2, \cdots, X_K) \sim$ Multi$(n, p_1, p_2, \cdots, p_K)$

> We use a random sample $\boldsymbol{x} = (x_1, x_2, \cdots, x_K)$ to infer $\{p_k\}_{k=1}^{K}$, the event probabilities

Question: How many parameters here?

Example: Multinomial Probability

Suppose that in a three-way election for a large country, candidate 1 received 20% of the votes, candidate 2 received 35% of the votes, and candidate 3 received 45% of the votes. If ten voters are **selected randomly**, what is the probability that there will be exactly two supporter for candidate 1, three supporters for candidate 2 and five supporters for candidate 3 in the sample?

$$\mathrm{P}(X_1 = 2, X_2 = 3, X_3 = 5) = \frac{10!}{2!3!5!}(0.2)^2(0.35)^3(0.45)^5 \approx 0.08$$

Example: Estimating Multinomial Parameters

If we **randomly select** ten voters, two supporter for candidate 1, three supporters for candidate 2 and five supporters for candidate 3 in the sample. What would our best guess for the population proportion each candidate would received?

Pearson's χ^2 Test

- The Hypotheses:
  $H_0 : p_1 = p_{1,0}; p_2 = p_{2,0}; \cdots, p_K = p_{K,0}$
  H_a : At least one is different

- The Test Statistic:

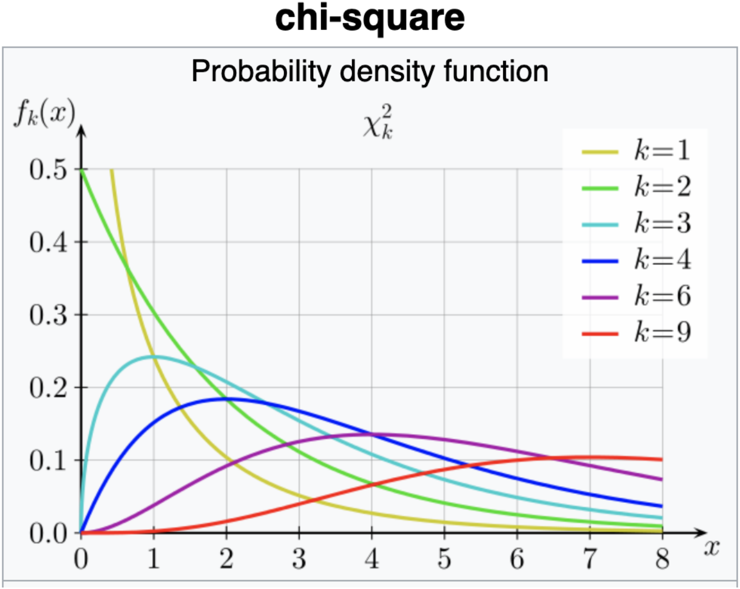
$$\chi^2_* = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k},$$

  where O_k is the observed frequency for the k_{th} event and E_k is the expected frequency under H_0

- The Null Distribution: $\chi^2_* \sim \chi^2_{df=K-1}$

- Assumption: $np_k > 5, k = 1, \cdots, K$

χ^2-Distribution



Example: Testing Mendel's Theories (pp 22–23, "Categorical Data Analysis" 2_{nd} Ed by Alan Agresti)

"Among its many applications, Pearson's test was used in genetics to test Mendel's theories of natural inheritance. Mendel crossed pea plants of pure yellow strain (dominant strain) plants of pure green strain. He predicted that second generation hybrid seeds would be 75% yellow and 25% green. One experiment produced $n = 8023$ seeds, of which $X_1 = 6022$ were yellow and $X_2 = 2001$ were green."

Use Pearson's χ^2 test to assess Mendel's hypothesis.

Color Preference Example

In Child Psychology, color preference by young children is used as an indicator of emotional state. In a study of 112 children, each was asked to choose "favorite" color from the 7 colors indicated below. Test if there is evidence of a preference at the 5% level.

Color	Blue	Red	Green	White	Purple	Black	Other
Frequency	13	14	8	17	25	15	20

An Example of Bivariate Categorical Data

A psychologist is interested in whether or not handedness is related to gender. She collected data on handedness for 100 individuals and the data set is summarized in the table below

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

An Example of Bivariate Categorical Data

A psychologist is interested in whether or not handedness is related to gender. She collected data on handedness for 100 individuals and the data set is summarized in the table below

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

- Grand total: 100

An Example of Bivariate Categorical Data

A psychologist is interested in whether or not handedness is related to gender. She collected data on handedness for 100 individuals and the data set is summarized in the table below

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

- Grand total: 100
- Marginal total for males: 52

An Example of Bivariate Categorical Data

A psychologist is interested in whether or not handedness is related to gender. She collected data on handedness for 100 individuals and the data set is summarized in the table below

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

- Grand total: 100
- Marginal total for males: 52
- Marginal total for females: 48

An Example of Bivariate Categorical Data

A psychologist is interested in whether or not handedness is related to gender. She collected data on handedness for 100 individuals and the data set is summarized in the table below

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

- Grand total: 100
- Marginal total for males: 52
- Marginal total for females: 48
- Marginal total for right-handed: 87

An Example of Bivariate Categorical Data

A psychologist is interested in whether or not handedness is related to gender. She collected data on handedness for 100 individuals and the data set is summarized in the table below

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

- Grand total: 100
- Marginal total for males: 52
- Marginal total for females: 48
- Marginal total for right-handed: 87
- Marginal total for left-handed: 13

An Example of Bivariate Categorical Data

A psychologist is interested in whether or not handedness is related to gender. She collected data on handedness for 100 individuals and the data set is summarized in the table below

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

- Grand total: 100
- Marginal total for males: 52
- Marginal total for females: 48
- Marginal total for right-handed: 87
- Marginal total for left-handed: 13

An Example of Bivariate Categorical Data

A psychologist is interested in whether or not handedness is related to gender. She collected data on handedness for 100 individuals and the data set is summarized in the table below

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

- Grand total: 100
- Marginal total for males: 52
- Marginal total for females: 48
- Marginal total for right-handed: 87
- Marginal total for left-handed: 13

This is an example of a contingency table

Contingency Tables

- Bivariate categorical data is typically displayed in a contingency table
- The number in each cell is the frequency for each category level combination
- Contingency table for the previous example:

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

For a given contingency table, we want to test **if two variables have a relationship or not?** ⇒ χ^2-Test

χ^2-Test for Independence

1. Define the null and alternative hypotheses:

   H_0 : there is no relationship between the 2 variables

   H_a : there is a relationship between the 2 variables

2. (If necessary) Calculate the marginal totals, and the grand total

3. Calculate the expected cell frequencies:

   $$\text{Expected cell frequency} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

4. Calculate the partial χ^2 values (χ^2 value for each cell of the table):

   $$\text{Partial } \chi^2 \text{ value } = \frac{(\text{observed - expected})^2}{\text{expected}}$$

χ^2-Test for Independence Cont'd

5. Calculate the χ^2 statistic:

$$\chi^2_{obs} = \sum \text{partial } \chi^2 \text{ value}$$

6. Calculate the degrees of freedom (df)

$$df = (\#\text{of rows} - 1) \times (\#\text{of columns} - 1)$$

7. Find the χ^2 critical value with respect to α

8. Draw the conclusion:

Reject H_0 if χ^2_{obs} is bigger than the χ^2 critical value ⇒ There is an statistical evidence that there is a relationship between the two variables at α level

Handedness/Gender Example Revisited

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

Is the percentage left-handed men in the population different from the percentage of left-handed women?

Summary

In this lecture, we learned

- Inference for multi-category data
- Inference for bivariate categorical data