This class is being conducted over Zoom. As the instructor, I will be recording this session. I have disabled the recording feature for others so that no one else will be able to record this session. I will be posting this session to the course's website.

If you have privacy concerns and do not wish to appear in the recording, you may turn video off (click "stop video") so that Zoom does not record you.

The chat box is always open for discussion and questions to the entire class. You may also send messages privately to the instructor or the TAs. Please note that Zoom saves all chat transcripts.

I create a live transcription of each session using Otter.ai. This means that Otter.ai will transcribe anything spoken over the Zoom audio. The transcript will be posted with the session video on the course website.

Some Notes About Week 5 & Finals Week

- Final exam is next **Tuesday, Sept. 8 @ 1 p.m. PST**
  - If you **have not** heard from me this week and cannot make the original time, **email me as soon as possible**
- Discussion Lab 5 this Friday will be a review. The discussion section will be recorded.
- Deadline to change grade option to pass/no pass is Friday, Sept. 4
- All extra credit work due by Tuesday, Sept. 8 @ 1 p.m. PST
- Finals week OH:
  - Mary
    - Sat., Sept. 5 @ 6 - 8 p.m. PST
  - Kyle
    - Sun., Sept. 6 @ 7:30 - 8:30 p.m. PST
    - Mon., Sept. 7 @ 7 - 9 p.m. PST
  - Jenifer
    - Fri., Sept. 5 @ 2 - 4 p.m. PST
    - Mon., Sept. 7 @ 2 - 4 p.m. PST
- Please fill out class evals! (They close tonight at 11:45 p.m. PST)

Inference for Categorical Data

Stats 7

Mary Ryan

Sept. 3, 2020


Course website:

https://canvas.eee.uci.edu/courses/28451


Slides can be found at:

https://maryryan.github.io/stats7-SS2-2020-slides/stats7-SS2-2020-chiSquared/stats7-SS2-2020-chiSquared

Learning Objectives

By the end of today's lecture, you should be able to:

- organize data from two categorical variables into a two-way table
- understand how to find expected counts under the assumption of independence between variables
- perform chi-squared hypothesis tests for indepdence
- draw appropriate conclusions for chi-squared tests in the context of the problem

Organizing Categorical Data for Comparison

- In the last few lecture we've talked about how draw inference about populations using continuous quantitative data or a single binary variable
- What if we wanted to compare categorical variables, though?
  - For each variable, we can look at the number of observations (counts) in each category
  - If we want to look at two variables at once, we can look at the number of observations that belong to one category of the first variabe and another category of the second variable
  - Can arrange counts into a two-way table (AKA, contingency table)

		Variable X	
		Category A	Category B
Variable Y	Category C	# in C AND # in A	# in C AND # in B
	Category D	# in D AND # in A	# in D AND # in B

Comparing Two Categorical Variables

- When we compare categorical variables, we often want to study the **relationship** between them
  - Does being in one category of one variable influence what category from the second variable you're in?
  - Are the variables **independent** or **dependent**?
- How do we tell if variables are independent?
- Recall basic probability:
  - Events *A* and *B* are independent if:

$$P(A \cap C) = P(A) \times P(C)$$

Independence of Two Categorical Variables

Events *A* and *B* are independent if $P(A \cap C) = P(A) \times P(C)$

- **P(category A of variable X)?**

$$P(A) = \frac{(\#\text{ in C AND }\#\text{ in A}) + (\#\text{ in D AND }\#\text{ in A})}{\text{Total }\#\text{ observed in all categories}}$$
$$= \frac{\text{Total in A}}{\text{Total observed}}$$

- **P(category C of variable Y)?**

$$P(C) = \frac{(\#\text{ in C AND }\#\text{ in A}) + (\#\text{ in C AND }\#\text{ in B})}{\text{Total }\#\text{ observed in all categories}}$$
$$= \frac{\text{Total in C}}{\text{Total observed}}$$

		Variable X	
		Category A	Category B
Variable Y	Category C	# in C AND # in A	# in C AND # in B
	Category D	# in D AND # in A	# in D AND # in B

Independence of Two Categorical Variables

Events *A* and *B* are independent if $P(A \cap C) = P(A) \times P(C)$

P(category A AND category C) if independent?

$$P(A \cap C) = P(A) \times P(C)$$

$$= \frac{(\text{Total in A})}{\text{Total \# observed in all categories}} \times \frac{(\text{Total in C})}{\text{Total \# observed in all categories}}$$

$$= \frac{(\text{Total in A}) \times (\text{Total in C})}{(\text{Total observed})^2}$$

Variable Y	Variable X: Category A	Variable X: Category B
Category C	# in C AND # in A	# in C AND # in B
Category D	# in D AND # in A	# in D AND # in B

Number of observations in A and C if independent?

$$P(A \cap C) \times (\text{Total observed}) = \frac{(\text{Total in A}) \times (\text{Total in C})}{(\text{Total observed})^2} \times (\text{Total observed})$$

$$= \frac{(\text{Total in A}) \times (\text{Total in C})}{(\text{Total observed})}$$

Variable Y	Variable X: A	Variable X: B
C	$\frac{(\text{Total C}) \times (\text{Total A})}{\text{TOTAL}}$	$\frac{(\text{Total C}) \times (\text{Total B})}{\text{TOTAL}}$
D	$\frac{(\text{Total D}) \times (\text{Total A})}{\text{TOTAL}}$	$\frac{(\text{Total D}) \times (\text{Total B})}{\text{TOTAL}}$

Independence of Two Categorical Variables

- So if your table of data matches the table of counts we would expect to get if two variables are independent, then can conclude variables are independent!

Variable Y \ Variable X	Category A	Category B
Category C	# in C AND # in A	# in C AND # in B
Category D	# in D AND # in A	# in D AND # in B

Variable Y \ Variable X	A	B
C	$\frac{(\text{Total C})\times(\text{Total A})}{\text{TOTAL}}$	$\frac{(\text{Total C})\times(\text{Total B})}{\text{TOTAL}}$
D	$\frac{(\text{Total D})\times(\text{Total A})}{\text{TOTAL}}$	$\frac{(\text{Total D})\times(\text{Total B})}{\text{TOTAL}}$

- But we're sampling, and we rarely get *exactly* our expected values. And what if some counts match but others don't?

Independence of Two Categorical Variables

- So if your table of data matches the table of counts we would expect to get if two variables are independent, then can conclude variables are independent!

Variable Y	Variable X: Category A	Variable X: Category B
Category C	# in C AND # in A	# in C AND # in B
Category D	# in D AND # in A	# in D AND # in B

Variable Y	Variable X: A	Variable X: B
C	$\frac{(\text{Total C})\times(\text{Total A})}{\text{TOTAL}}$	$\frac{(\text{Total C})\times(\text{Total B})}{\text{TOTAL}}$
D	$\frac{(\text{Total D})\times(\text{Total A})}{\text{TOTAL}}$	$\frac{(\text{Total D})\times(\text{Total B})}{\text{TOTAL}}$

- But we're sampling, and we rarely get *exactly* our expected values. And what if some counts match but others don't?
- How do we know when counts are close enough to expected for the variables to be considered independent?
- How do we when counts are far away enough from expected for the variables to be considered not independent?

Independence of Two Categorical Variables

- So if your table of data matches the table of counts we would expect to get if two variables are independent, then can conclude variables are independent!

Variable Y	Variable X: Category A	Variable X: Category B
Category C	# in C AND # in A	# in C AND # in B
Category D	# in D AND # in A	# in D AND # in B

Variable Y	Variable X: A	Variable X: B
C	$\frac{(\text{Total C})\times(\text{Total A})}{\text{TOTAL}}$	$\frac{(\text{Total C})\times(\text{Total B})}{\text{TOTAL}}$
D	$\frac{(\text{Total D})\times(\text{Total A})}{\text{TOTAL}}$	$\frac{(\text{Total D})\times(\text{Total B})}{\text{TOTAL}}$

- But we're sampling, and we rarely get *exactly* our expected values. And what if some counts match but others don't?
- How do we know when counts are close enough to expected for the variables to be considered independent?
- How do we when counts are far away enough from expected for the variables to be considered not independent?
- We formally test this relationship through a Chi-Squared Test for Independence

Chi-Squared Test for Independence: Hypotheses

- We are looking to see if there is a **non-independent relationship** between the two variables
  - For **null hypothesis** (H_0), we will assume there is no relationship, or the variables are **independent**
  - For **alternative hypothesis** (H_A), we will find evidence that there is some kind of relationship, or the variables are **not independent**

$$H_0: \text{ variable X and variable Y are independent}$$

$$H_A: \text{ variable X and variable Y are not independent}$$

- In order to perform a hypothesis test, we need to make sure that all the cells in the expected count table are **greater than 5**
  - This is like the "large enough sample size" requirement for CLT

Chi-Squared Test for Independence: Test Statistic

- In other hypothesis testing scenarios, we used a **test statistic** to help us determine whether our data would be too weird to see if H_0 was true
    - Z-statistics, T-statistics
- Here we use a **chi-squared statistic** (χ^2)

$$\chi^2 = \frac{[(\text{observed counts for being in category A and category C}) - (\text{expected counts for category A and C})]^2}{(\text{observed counts for category A and C})}$$

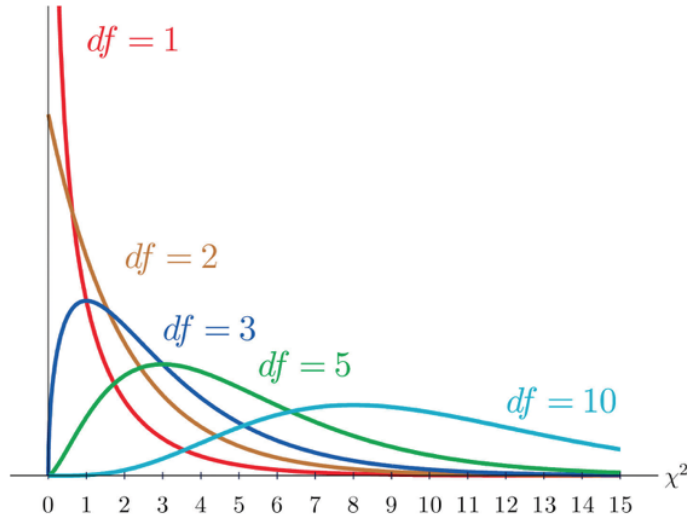
$$+ \ldots +$$

$$+ \frac{[(\text{observed counts for category B and D}) - (\text{expected counts for category B and D})]^2}{(\text{observed counts for category B and D})}$$

- If $\chi^2 = 0$, variables X and Y are perfectly **independent**
- As χ^2 gets bigger, we are accruing more and more evidence that X and Y are **not independent**
- χ^2 will also be bigger if we have more categories to compare

Chi-Squared Test for Independence: Chi-Squared Distribution

- χ^2 follows a Chi-squared distribution
  - Shape of the distribution controlled by degrees of freedom



$$\begin{aligned}\text{degrees of freedom} &= (\#\text{ rows in two-way table} - 1) \times (\#\text{ columns in two-way table} - 1) \\ &= (\#\text{ categories in variable X} - 1) \times (\#\text{ categories in variable Y} - 1)\end{aligned}$$

Chi-Squared Test for Independence: Critical Value

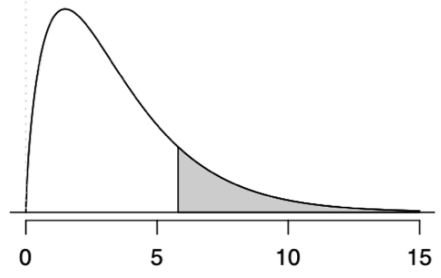
How do we know when χ^2 is big enough to reject H_0 ?

- Look up **critical value**, χ^{2*}_{df} , in chi-squared table

1. Find row with correct degrees of freedom (df)
2. Find column with desired probability of statistics being more extreme
    - Want this to be small, so usually choose 0.05

- Can also use **p-value** as justification, if using calculator function (more on this in a few slides)

Chi-square probability table



Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
	12	14.01	15.81	18.55	21.03	24.05	26.22	28.30	32.91
	13	15.12	16.98	19.81	22.36	25.47	27.69	29.82	34.53
	14	16.22	18.15	21.06	23.68	26.87	29.14	31.32	36.12
	15	17.32	19.31	22.31	25.00	28.26	30.58	32.80	37.70
	16	18.42	20.47	23.54	26.30	29.63	32.00	34.27	39.25

Chi-Squared Test for Independence: Decision

- If your test statistic χ^2 > critical value χ^{2*}_{df}, **reject the null hypothesis**
  - We have sufficient evidence to conclude that variables X and Y are **not independent**
  - The category you are in for variable X will influence what category you're in for variable Y
- If your test statistic χ^2 < critical value χ^{2*}_{df}, **fail-to-reject the null hypothesis**
  - We **do not** have sufficient evidence to conclude that variables X and Y are **not independent**
  - The category you are in for variable X will likely not influence what category you're in for variable Y
- Unlike other hypothesis tests we went over, chi-squared tests don't have "sided" tests so we are always rejecting if our test statistic is **greater than** our critical value

Calculator: χ^2 -Test

Before finding the function:

- 2nd MATRIX > EDIT > [A]
- Change number of rows and columns to matrix to match your two-way table
- Copy values from your two-way table into the cells of the matrix

To get to the calculator function on a TI-84:

- STAT > TESTS > C: χ^2 -Test

To calculate perform a χ^2 test for independence:

- input:
  - Observed: [A]
  - Expected: [B] (this is where the calculator will save the expected counts table)
- then arrow down to CALCULATE and press ENTER

Example: Alcohol & Smoking

A social study was conducted to see if hard alcohol consumption (yes or no) and cigarette (traditional or electronic) smoking (yes or no) were linked. 400 subjects were sampled.

- What are our hypotheses?

	Smoking: Yes	Smoking: No
Alcohol: Yes	120	102
Alcohol: No	80	98

Example: Alcohol & Smoking

A social study was conducted to see if hard alcohol consumption (yes or no) and cigarette (traditional or electronic) smoking (yes or no) were linked. 400 subjects were sampled.

- Find the expected count table.

	Smoking: Yes	Smoking: No
Alcohol: Yes	120	102
Alcohol: No	80	98

Example: Alcohol & Smoking

A social study was conducted to see if hard alcohol consumption (yes or no) and cigarette (traditional or electronic) smoking (yes or no) were linked. 400 subjects were sampled.

- Calculate the test statistic. Find the critical value.

	Smoking: Yes	Smoking: No
Alcohol: Yes	120	102
Alcohol: No	80	98

Example: Alcohol & Smoking

A social study was conducted to see if hard alcohol consumption (yes or no) and cigarette (traditional or electronic) smoking (yes or no) were linked. 400 subjects were sampled.

- Decision?

	Smoking: Yes	Smoking: No
Alcohol: Yes	120	102
Alcohol: No	80	98

Example: Party Confidence

In a nationwide telephone poll of 1,000 adults representing Democrats, Republicans, and Independents, respondents were asked 2 questions: their party affiliation and if their confidence in the U.S. banking system had been shaken by the 2008 financial crisis.

- What are our hypotheses?

	Yes	No	No opinion
Democrat	175	220	55
Republican	150	165	35
Independent	75	105	20

Example: Party Confidence

In a nationwide telephone poll of 1,000 adults representing Democrats, Republicans, and Independents, respondents were asked 2 questions: their party affiliation and if their confidence in the U.S. banking system had been shaken by the 2008 financial crisis.

- Find the expected count table.

	Yes	No	No opinion
Democrat	175	220	55
Republican	150	165	35
Independent	75	105	20

Example: Party Confidence

In a nationwide telephone poll of 1,000 adults representing Democrats, Republicans, and Independents, respondents were asked 2 questions: their party affiliation and if their confidence in the U.S. banking system had been shaken by the 2008 financial crisis.

- Calculate the test statistic. Find the critical value.

	Yes	No	No opinion
Democrat	175	220	55
Republican	150	165	35
Independent	75	105	20

Example: Party Confidence

In a nationwide telephone poll of 1,000 adults representing Democrats, Republicans, and Independents, respondents were asked 2 questions: their party affiliation and if their confidence in the U.S. banking system had been shaken by the 2008 financial crisis.

- Decision?

	Yes	No	No opinion
Democrat	175	220	55
Republican	150	165	35
Independent	75	105	20