EDP308: STATISTICAL LITERACY

The University of Texas at Austin, Fall 2020

RAZ: Rebecca A. Zárate, MA

Overview

- Comparing Categorical Variables
- Chi-Squared Tests
  - Independence
  - Goodness of Fit
- Contingency Tables, Again
  - Marginal Probability
  - Conditional Probability
- Chi-Squared Test of Independence
  - Happiness and Income
  - Titanic Survivors
  - Test Grades and Studying
- Chi-Squared Test of Independence in R

Chi-Squared χ^2

Pronounced “Ki” (not like the tea “Chai”)

Comparing

What kind of things do we compare in an independent samples t-test?

What kind of variables do you need for such tests?

What if I want to compare two categorical variables, like, does being Male vs. Female affect which gym (Gregory, Rec Center, etc.) you work out in?

Chi-Squared and Categories

- The Chi-Squared test allows us to investigate associations between categorical variables like,
  - Sex, political affiliation, race, preferences, geographical area, the list goes on…
- There are two main types of Chi-Squared tests:
  - Chi-Squared Test of Independence
    - Tests whether two variables are independent
      - Ex. Is happiness independent of income?
  - Chi-Squared Goodness of Fit
    - Used to test a hypothesis for one variable
      - Ex. Is police brutality equal among different race-ethnicities?

Contingency Tables, Again

Contingency Tables

- Contingency tables show the frequency counts and probabilities for two different categorical variables
    - We used these for calculate probabilities in our last PPT
  - Ex. Below is a contingency table showing the counts for Happiness Level and Income Level
    - Do you think these two variables are independent?

	Unhappy	Neutral	Happy	Total
Lower Class	104	314	119	**537**
Middle Class	83	494	277	**854**
Upper Class	29	178	135	**342**
Total	**216**	**986**	**531**	**1733**

Marginal Probability

- Remember, a marginal probability is the probability of seeing a specific outcome, ex. the probability of being in the Middle Class

What proportion of people are unhappy?
What proportion are middle class?
What proportion are unhappy AND middle class?

	Unhappy	Neutral	Happy	Total
Lower Class	104	314	119	**537**
Middle Class	83	494	277	**854**
Upper Class	29	178	135	**342**
Total	**216**	**986**	**531**	**1733**

Marginal Probability

- To calculate marginal probability we take the total observed count of that variable and divide it by the total sample size

What proportion of people are unhappy?

Ex. $P(Unhappy) = \frac{216}{1733} \approx .125$

What proportion are middle class?

Ex. $P(Middle\ Class) = \frac{854}{1733} \approx .493$

	Unhappy	Neutral	Happy	Total
Lower Class	104	314	119	**537**
Middle Class	83	494	277	**854**
Upper Class	29	178	135	**342**
Total	**216**	**986**	**531**	**1733**

Marginal Distribution of Income

- Here is the marginal distribution for the different levels of Income
  - We're just finding the probability of being in a certain income bracket regardless of happiness

What is the marginal distribution of Income?

Which is most common?

Income	Total	P(Income Level)
Lower Class	537	$\frac{537}{1733} \approx \mathbf{.310}$
Middle Class	854	$\frac{854}{1733} \approx \mathbf{.493}$
Upper Class	342	$\frac{342}{1733} \approx \mathbf{.197}$

Marginal Distribution of Happiness

- Here is the marginal distribution for the different levels of Happiness
  - We're just finding the probability of being in a certain level of happiness regardless of your income

What is the marginal distribution of Happiness?

Which is most common?

Happiness	Unhappy	Neutral	Happy
Total	216	986	531
P(Happiness)	$\frac{216}{1733} \approx \mathbf{.125}$	$\frac{986}{1733} \approx \mathbf{.569}$	$\frac{531}{1733} \approx \mathbf{.306}$

Joint Probability

- Joint Probability is the probability of two things happening, ex. being unhappy and middle class

What proportion are unhappy AND middle class?

Ex. $P(Unhappy\ and\ Middle\ Class) = \frac{83}{1733} \approx .048$

	Unhappy	Neutral	Happy	Total
Lower Class	104	314	119	**537**
Middle Class	83	494	277	**854**
Upper Class	29	178	135	**342**
Total	**216**	**986**	**531**	**1733**

Conditional Probabilities

- Conditional probabilities are probabilities of a specific outcome from a categorical variable **given** a certain levels of another variable
  - "Given" means that you are already part of that group
    - What is the probability of being Happy GIVEN that you are in the Lower class?
    - What is the probability of being Upper class GIVEN that you are Unhappy?

Which do you think will have a higher probability? Happy given Lower class or Upper class given Unhappy?

	Unhappy	Neutral	Happy	Total
Lower Class	104	314	119	**537**
Middle Class	83	494	277	**854**
Upper Class	29	178	135	**342**
Total	**216**	**986**	**531**	**1733**

Conditional Probabilities

- Conditional probabilities are probabilities of a specific outcome from a categorical variable **given** a certain levels of another variable

What is the probability of being Happy GIVEN that you are in the Lower class?

Ex. $P(Happy|Lower\ Class) = \frac{119}{537} \approx .222$

What is the probability of being Upper class GIVEN that you are Happy?

Ex. $P(Upper\ Class|Unhappy) = \frac{29}{216} \approx .134$

	Unhappy	Neutral	Happy	Total
Lower Class	104	314	119	**537**
Middle Class	83	494	277	**854**
Upper Class	29	178	135	**342**
Total	**216**	**986**	**531**	**1733**

Conditional Probabilities

- To calculate conditional probabilities, we take the total number of observations in each cell for each level of happiness, then divide by the variable that is "given", here the row sum, (income level)

Find the conditional distribution of Happiness given Income:
i.e. Find P(Happiness | Income)

	Unhappy	Neutral	Happy	Total
Lower Class	$\frac{104}{537} \approx .194$	$\frac{314}{537} \approx .585$	$\frac{119}{537} \approx .222$	**537**
Middle Class	$\frac{83}{854} \approx .097$	$\frac{494}{854} \approx .578$	$\frac{277}{854} \approx .324$	**854**
Upper Class	$\frac{29}{342} \approx .085$	$\frac{178}{342} \approx .520$	$\frac{135}{342} \approx .395$	**342**

Chi-Squared Test of Independence

Independence of Two Variables

Do you think that happiness will be independent of the level of income you make?

Meaning, do you think it matters how much money you make when it comes to your happiness?

Or do you think that being in a certain income bracket is related to your level of happiness?

Let's test this hypothesis about Happiness and Income.

Chi-Squared Test of Independence

- Just like before… We are going to use hypothesis testing to test for independence.

What do you think our hypotheses will be?

Step 1: Independence and the Null Hypothesis

- As usual, the NULL hypothesis is going to assume there is nothing special going on. In this case, we assume independence. Why?
  - H_0: Happiness and Income are independent.
  - H_1: Happiness and Income are not independent.

Independence Helps Us Know What to Expect

- When we assume independence, we can calculate what we would EXPECT to see if the two variable are independent. How? Because…
  - If A and B are independent, meaning being in a particular social class has no impact on your happiness, then:
    - $P(Unhappy\ and\ Lower\ Class) = P(Unhappy) * P(Lower\ Class)$
    - $P(Unhappy\ and\ Middle\ Class) = P(Unhappy) * P(Middle\ Class)$
  - This means that if they are independent, the probability of both occurring (the joint probability) should be equal to the product of their individual marginal probabilities.

$$P(A\ and\ B) = P(A) * P(B)$$

- The research questions:
  - Is happiness independent of class?
    - Or do the two interact somehow…?
    - If you are of a certain class, are you more likely to be a certain level of happiness?
  - Are people who make more money happier?

Expected Frequencies

- If we assume that happiness and income are *independent*, then the EXPECTED number of people who would be lower class AND unhappy is…

$$\frac{216 * 537}{1733} \approx 66.93$$

Is this what we observed?

vs.

	Unhappy	Neutral	Happy	Total
Lower Class	104	314	119	**537**
Middle Class	83	494	277	**854**
Upper Class	29	178	135	**342**
Total	**216**	**986**	**531**	**1733**

Expected Frequencies

- Now we can find all the expected frequencies for each possible combination of happiness and income level
  - Does it look like Happiness are Income are going to be independent?

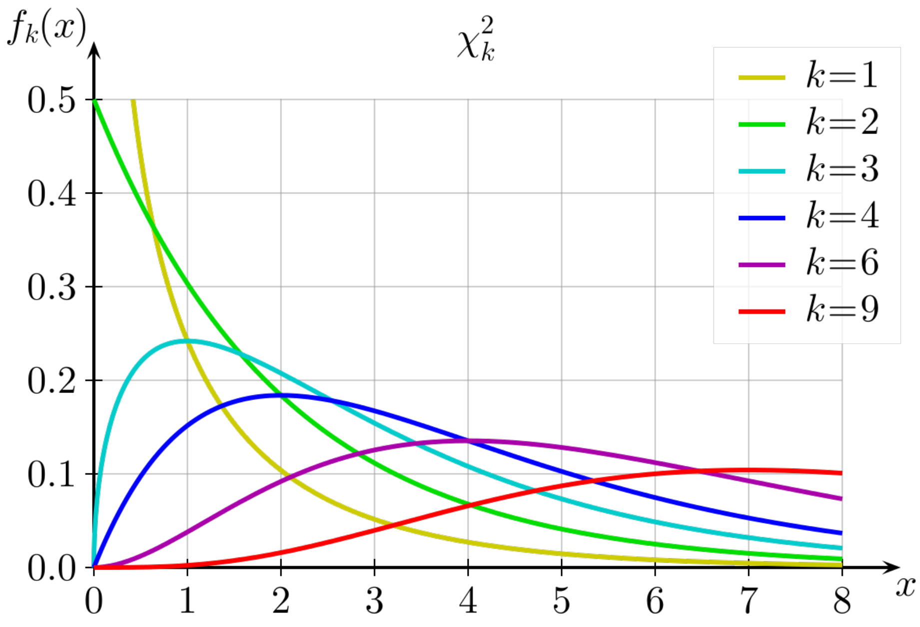
	Unhappy	Neutral	Happy	Total
Lower Class	104 (66.93)	314 (305.53)	119 (164.54)	**537**
Middle Class	83 (106.44)	494 (485.89)	277 (261.67)	**854**
Upper Class	29 (42.63)	178 (194.58)	135 (104.79)	**342**
Total	**216**	**986**	**531**	**1733**

Step 2 and 3: Significant and df

- The χ^2 statistic follows a χ^2 distribution
- Just like with F-statistics in ANOVA, Chi-squared tests are one-tailed tests (just like ANOVA) because the χ^2 statistic is always positive

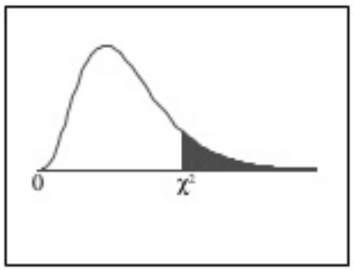
- To get the degrees of freedom we will subtract 1 from each of the two categorical variable' levels we are testing.
  - Happiness (3 levels) – 1 = 2
  - Income (3 levels) – 1 = 2
- Then we multiply the two values
  - $df = (3-1)(3-1) = 4$

$$df = 4$$
$$\alpha = .05$$



Step 4: Find the Critical Value

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

- In our example,
  - $\chi^2_{stat} = 54.04$,
  - $\alpha = .05$
  - $df = 4$
  - $\chi^2_{crit} = 9.488$

Chi-Squared Statistic

- Just as with other hypothesis tests, we know that even if the null were true, we won't get exactly the same values from our sample… But how close is close enough to assume independence?
  - This is what the Chi-Squared χ^2 test is testing…
  - Is what we see close enough to what we would expect if things were independent?
  - Or, is what we see further out from the realm of "reasonable null world" that we can reject the null?

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Step 5: Calculate the Chi-Squared Statistic

Calculate the χ^2 statistic for Happiness and Income

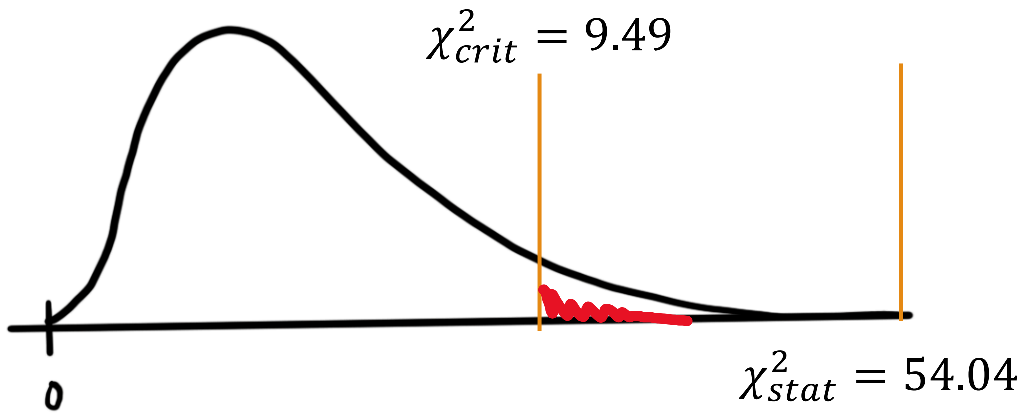
	Unhappy	Neutral	Happy	Total
Lower Class	104 (66.93)	314 (305.53)	119 (164.54)	**537**
Middle Class	83 (106.44)	494 (485.89)	277 (261.67)	**854**
Upper Class	29 (42.63)	178 (194.58)	135 (104.79)	**342**
Total	**216**	**986**	**531**	**1733**

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$\chi^2 = \frac{(104-66.93)^2}{66.93} + \frac{(314-305.53)^2}{305.53} + \frac{(119-164.54)^2}{164.54} + \frac{(83-106.44)^2}{106.44} + \frac{(494-485.89)^2}{485.89} + \frac{(277-261.67)^2}{261.67} + \frac{(29-42.63)^2}{42.63} + \frac{(178-194.58)^2}{194.58} + \frac{(135-104.79)^2}{104.79} \approx 54.04$$

Step 6: Draw Conclusions

- $\chi^2_{stat} = 54.04$
- $df = 4$
- $\alpha = .05$
- $\chi^2_{crit} = 9.49$



- Because χ^2_{stat} is past χ^2_{crit}, we reject H_0
- The observed frequencies are different enough from the expected frequencies that we can conclude Happiness and Class are NOT independent.

Try it. Titanic Data

- Think of what you know about the Titanic and it's survivors. Do you think First class (rich) passengers were more likely to survive compared to Third class passengers? Test if Class and Survival are independent at $\alpha = .05$.

	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

Step 1, 2, 3, and 4

Step 1:

$$H_0: Survial\ is\ independent\ of\ Class.$$
$$H_1: Survial\ is\ NOT\ independent\ of\ Class.$$

Step 2:

$$\alpha = .05$$

Step 3: Chi-Squared Test of Independence

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Step 4:

$$df = 3 \text{ and } \alpha = .05,\ \chi^2_{crit} = 7.81$$

Step 5: Compute Test Statistic

Step 5: The expected counts for each cell are shown in parenthesis below:

	First	Second	Third	Crew	Total
Alive	203 (104.98)	118 (92.06)	178 (228.06)	212 (285.89)	711
Dead	122 (220.01)	167 (192.94)	528 (477.94)	673 (599.11)	1490
Total	325	285	706	885	2201

Step 5: Compute Test Statistic

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

	First	Second	Third	Crew	Total
Alive	203 (104.98)	118 (92.06)	178 (228.06)	212 (285.89)	711
Dead	122 (220.01)	167 (192.94)	528 (477.94)	673 (599.11)	1490
Total	325	285	706	885	2201

$$\chi^2 = \frac{(203-104.98)^2}{104.98} + \frac{(118-92.06)^2}{92.06} + \frac{(178-228.06)^2}{228.06} + \frac{(212-285.89)^2}{285.89} + \frac{(122-220.01)^2}{220.01} + \frac{(167-192.94)^2}{192.94} + \frac{(528-477.94)^2}{477.94} + \frac{(673-599.11)^2}{599.11} \approx \mathbf{190.4}$$

Step 6: Draw Conclusions

Step 6:

Our $\chi^2_{stat} = 190.4$ is past $\chi^2_{crit} = 7.81$. Therefore, there is sufficient evidence to reject H_0.

"Our data show significant evidence to reject the null hypothesis that passenger class and survival status are independent. Instead, we conclude that on the Titanic, there is sufficient evidence to suggest that passenger class and survival status are dependent."

Test Grades and Studying

- Do you think Test Grades and whether or not you Studied are independent?

	Test Grade				Total
	A	B	C	D	
Studied	11	17	7	3	38
Didn't Study	1	4	12	15	32
Total	12	21	19	18	70

Test Grades and Studying

- Do you think Test Grades and whether or not you Studied are independent?
  - Definitely not…!

- $\chi^2_{stat} = 25.37$
- $df = 3$
- $\alpha = .05$
- $\chi^2_{crit} = 7.845$

	Test Grade				Total
	A	B	C	D	
Studied	11	17	7	3	38
Didn't Study	1	4	12	15	32
Total	12	21	19	18	70

Up Next…

- Our last topic will be another type of Chi-Squared test, but this time we are assessing how well something "fit" with what we'd expect…

Chi-Squared Goodness of Fit

Chi-Squared Test of Independence in R

Chi-Squared Test of Independence in R

```r
########################################
############# Chi-Squared ##############
######## Test of Independence ##########
########################################

# The data. Using data from our Contingency Tables PPT
# This is one way to unput the data if you are working with summary statistics.
happiness_income <- as.table(rbind(c(104, 314, 119), c(83, 494, 277), c(29, 178, 135)))
dimnames(happiness_income) <- list(Class = c("Lower", "Middle", "Upper"),
                    Happiness = c("Unhappy","Neutral", "Happy"))

chisq.test(happiness_income)

# Titanic Survival and Class.
titanic <- as.table(rbind(c(203, 118, 178, 212), c(122, 167, 528, 673)))
dimnames(titanic) <- list(Status = c("Survived", "Died"),
                    Class = c("First","Second", "Third", "Crew"))

chisq.test(titanic)

# The data. Using data from our Contingency Tables PPT
# Test Grade and Studying

study_grades <- as.table(rbind(c(11, 17, 7, 3), c(1, 4, 12, 15)))
dimnames(study_grades) <- list(Status = c("Studied", "Didn't Study"),
                    Letter_Grade = c("A","B", "C", "D"))

chisq.test(study_grades)
```

Using summary data.

Chi-Squared Test of Independence in R Output

- Here is the output from the chi-squared tests.
- All three examples were significant. We can conclude:
  - Happiness and Income are not independent.
  - Survival on Titanic and Class are not independent.
  - Test Grades and Studying are not independent.

```
> chisq.test(happiness_income)

        Pearson's Chi-squared test

data:  happiness_income
X-squared = 54.043, df = 4, p-value = 5.155e-11

> chisq.test(titanic)  # Prints test summary

        Pearson's Chi-squared test

data:  titanic
X-squared = 190.4, df = 3, p-value < 2.2e-16

> chisq.test(study_grades)

        Pearson's Chi-squared test

data:  study_grades
X-squared = 25.369, df = 3, p-value = 1.293e-05
```