Chapter 11: χ^2 Tests

DSCC 462
Computational Introduction to Statistics

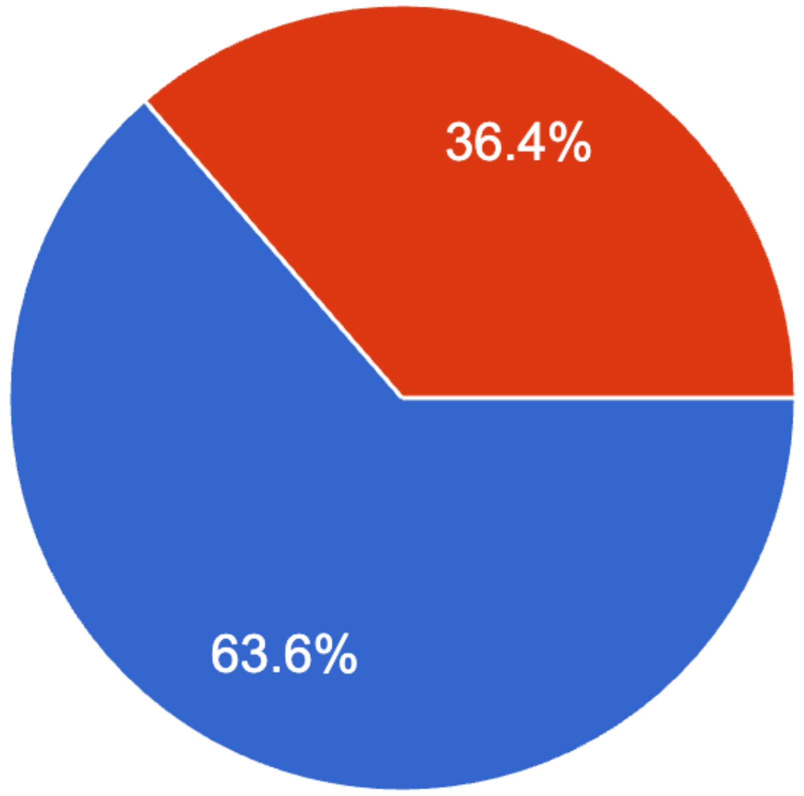
Anson Kahng
Fall 2022

Survey Feedback (n = 44 out of 54 so far)

The material is clear

The instructor presents concepts clearly.

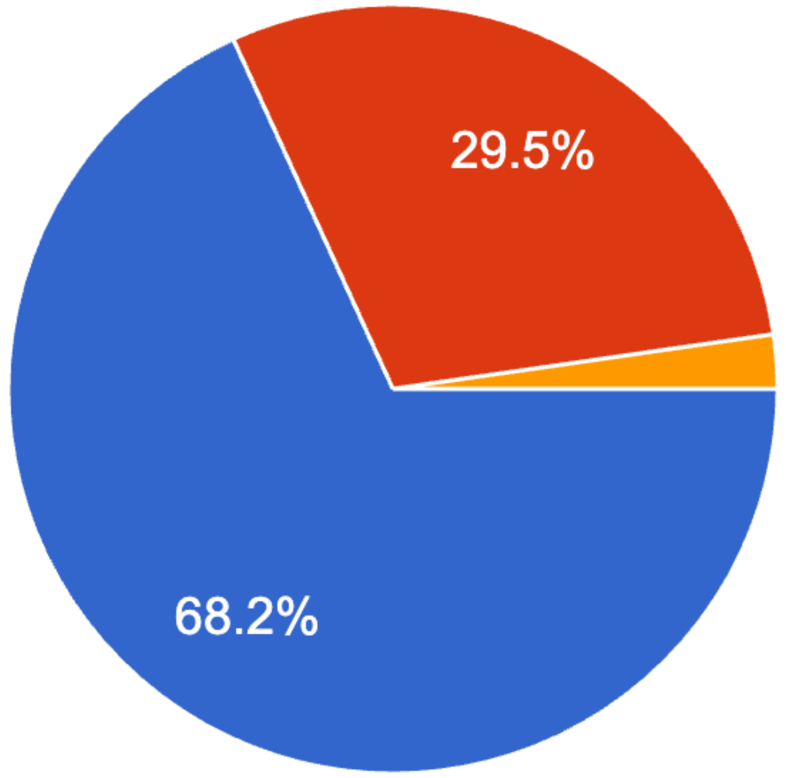
44 responses



The assignments are useful

The assignments help me apply concepts that we learn

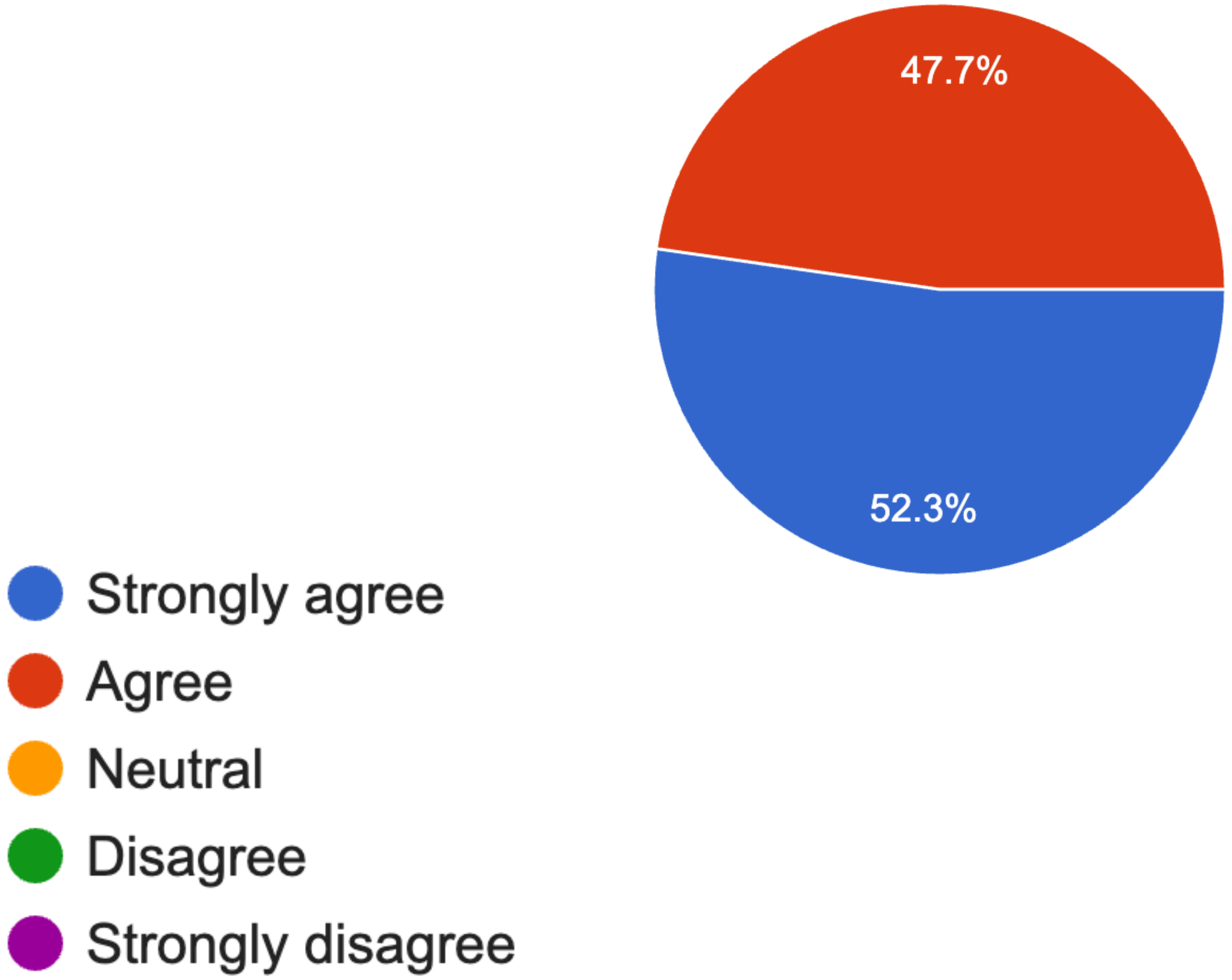
44 responses



You are learning!

I am learning a lot in this class.

44 responses



Survey Feedback (n = 44 out of 54 so far)

	Some of you mentioned that…	What we'll do going forward
START	• You like **outlines / summaries** of lecture • You find **examples** helpful • You want more **feedback** on homeworks • You want to see more **R code** on slides	• I will add **summary slides** to lecture • I will use **examples** to frame concepts • TAs will give more **detailed feedback** on assignments • I will add **R code** to slides
STOP	• Homework questions are **repetitive** • Homeworks are slightly **long**	• I can't promise homeworks will be shorter, but I'll make sure the questions **cover more concepts** 😊
CONTINUE	• Being **responsive** to students' needs • Seeking **feedback** from students • **Answering questions** during class • **Solving examples** in class	• I'll **continue doing these things**!

Key Reminders

- I have three office hours each week! Only a few of you have used them regularly
  - These are **by appointment** so please email me ahead of time
  - If you cannot make then, **email me so we can find another time**!
- The TAs are also here to support you
  - Learning from a group can help; be sure to prepare questions in advance
- Educational videos:
  - YouTube: zedstatistics, 3blue1brown, jbstatistics

Community Norms

- Let's keep side conversations to a minimum
- Let's stay awake

Lecture Plan for Today

- Goodness-of-Fit Test
  - True proportion = expected proportion?
  - Generalization of proportion hypothesis tests
- Chi-Squared (χ^2) Test of Independence
  - Are variables related or not?

Multi-Category Proportions

- Last lecture, we looked at inference for proportions
  - In this setting, a variable could take on one of two values
  - What if the variable had more categories?
- Example: Let's say we are trying to figure out what proportion of people have each season (winter, spring, summer, fall) as their favorite. How can we do inference in this setting?
  - Instead of one p, we have to infer values for p_1, p_2, p_3

Goodness-of-Fit

- Consider a categorical variable with multiple categories
  - E.g., eye color: brown, hazel, blue, other
- Perhaps we want to test whether the true proportion of people falling into each category is equal to some value
- Use the Goodness-of-Fit Test

Goodness-of-Fit

- Let p_i be the true proportion of the population that falls into category i
  - $i = 1,\ldots,k$, where k is the number of categories
  - Note that $\sum_{i=1}^{k} p_i = 1$
- Our hypotheses are as follows:
  - $H_0 : p_1 = p_{1_0}, p_2 = p_{2_0}, \ldots, p_k = p_{k_0}$
  - H_1 : at least one of these equalities does not hold

Goodness-of-Fit

- Test the hypothesis using a chi-squared (χ^2) test
- The test statistic is

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

- Here, O_i is the observed number of people who fall in category i
- E_i is the expected number of people who fall into category i under the null hypothesis

Goodness-of-Fit

- We calculate the test statistic using the following information
- Recall that the test statistic is $X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

Category	**1**	**2**	**...**	**k**
Observed	O_1	O_2	...	O_k
Expected	$E_1 = n*p_{1_0}$	$E_2 = n*p_{2_0}$	...	$E_k = n*p_{k_0}$

Goodness-of-Fit

- We are interested in the p-value of $\Pr(\chi^2 > X^2)$
- We can find this p-value by using a χ^2 distribution with $df = k - 1$
    - In R: $p =$ `1 - pchisq(X2, df)`
    - Always looking for *upper tail probability* (probability of seeing an outcome that is as extreme or more extreme than what we observed)
- If $p \leq \alpha$, we reject H_0
- If $p > \alpha$, we fail to reject H_0
- State your conclusion in the context of the problem
- Note: For this test to be valid, all the expected counts must be at least 5

Goodness-of-Fit: Example

- Consider eye color, broken down into four categories: brown, hazel, blue, other
- Based on previous knowledge, we believe that 40% of people have brown eyes, 10% have hazel eyes, 5% have blue eyes, and 45% have some other color eyes
- To see if this is correct, we go out and take a sample of 200 people and determine their eye color. We get 84 people with brown eyes, 17 people with hazel eyes, 16 people with blue eyes, and 83 people with other color eyes
- Test at the $\alpha = 0.05$ significance level

Goodness-of-Fit: Example

- $H_0 : p_1 = 0.4,\ p_2 = 0.1,\ p_3 = 0.05,\ p_4 = 0.45$ vs. H_1 : at least one of these proportions does not hold

- Calculate the test statistic:

Category	**Brown**	**Hazel**	**Blue**	**Other**
Observed				
Expected				

- $X^2 =$

- $\Pr(\chi^2 > X^2) =$

- Conclusion:

Goodness-of-Fit: Example

- $H_0: p_1 = 0.4,\ p_2 = 0.1,\ p_3 = 0.05,\ p_4 = 0.45$ vs. H_1 : at least one of these proportions does not hold

- Calculate the test statistic:

Category	**Brown**	**Hazel**	**Blue**	**Other**
Observed	84	17	16	83
Expected	80	20	10	90

- $X^2 = \dfrac{(84-80)^2}{80} + \dfrac{(17-20)^2}{20} + \dfrac{(16-10)^2}{10} + \dfrac{(83-90)^2}{90} = 4.79$

- $\Pr(\chi^2 > X^2)$ = 1 - pchisq(4.79, df=3) = 0.188

- Conclusion: Since $p = 0.188 > \alpha = 0.05$, we fail to reject H_0 – there is not sufficient evidence to conclude that the proportion of people with various eye colors has changed

Goodness-of-Fit: Example

- Can do directly in R as well

```
> chisq.test(c(84,17,16,83),p=c(0.4,0.1,0.05,0.45))

	Chi-squared test for given probabilities

data:  c(84, 17, 16, 83)
X-squared = 4.7944, df = 3, p-value = 0.1875
```

Contingency Table

- Now, let's consider the case of two categorical variables
  - We used a normal approximation to the binomial distribution and formed a two-proportion z-test for binary variables
- A generalized technique for testing proportions is through the χ^2 test of independence for contingency tables

Contingency Table

Variable 1	**Variable 2**		**Total**
	Yes	No	
Yes	O_{11}	O_{12}	r_1
No	O_{21}	O_{22}	r_2
Total	c_1	c_2	n

Testing Whether Variables are Independent

- Setting:
  - Consider two categorical variables: favorite season (winter, spring, summer, fall) and whether or not someone has pets (yes, no)
  - We are interested in whether a population's favorite season is independent of whether or not they are a pet owner
  - How can we test such a hypothesis?
- Idea: Extend the goodness-of-fit test to multiple dimensions

χ^2 Test of Independence

- We are testing the following hypotheses:
  - H_0 : the two variables are independent
  - H_1 : the two variables are associated (i.e., not independent)

χ^2 Test of Independence

- Similar to the goodness-of-fit test, the test of independence compares the observed frequencies in each category of the contingency table with the expected frequencies given that the null hypothesis is true
  - Let O be the observed frequencies
  - Let E be the expected frequencies under the null hypothesis
- Use the chi-square test to determine whether the deviations between the observed and expected frequencies are too large to be attributed to chance

Expected Contingency Table

- We compare what we observe with what we expect to see if the null hypothesis is true
- Calculate the expected counts as follows:

Variable 1	**Variable 2**		**Total**
	Yes	No	
Yes			r_1
No			r_2
Total	c_1	c_2	n

Expected Contingency Table

- We compare what we observe with what we expect to see if the null hypothesis is true

- Calculate the expected counts as follows:

Variable 1	**Variable 2**		**Total**
	Yes	No	
Yes	$E_{11} = \frac{r_1 \cdot c_1}{n}$	$E_{12} = \frac{r_1 \cdot c_2}{n}$	r_1
No	$E_{21} = \frac{r_2 \cdot c_1}{n}$	$E_{22} = \frac{r_2 \cdot c_2}{n}$	r_2
Total	c_1	c_2	n

χ^2 Test of Independence

- For a contingency table with r rows and c columns (for a total of rc cells), the chi-square test statistic is as follows:

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

- X^2 approximately follows a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom
- Find the p-value $p =$ `1-pchisq(`X^2`, df)`
- If $p \leq \alpha$, then reject H_0
- If $p > \alpha$, then fail to reject H_0
- In order for the χ^2 distribution to be appropriate, no cell should have an expected or observed frequency less than 5

χ^2 Test of Independence: Example

- To examine the effectiveness of flossing, we wish to know whether there is an association between the occurrence of gum disease and the use of floss
- Test the following hypotheses:
  - H_0 : Flossing and gum disease are independent
  - H_1 : Flossing and gum disease are associated
- Perform this test at the $\alpha = 0.05$ significance level

χ^2 Test of Independence: Example

- 350 dental patients were examined to determine whether flossing daily reduced their risk for gum disease

- Observed contingency table:

Daily Flossing	**Gum Disease**		**Total**
	Yes	No	
Yes	50	127	177
No	82	91	173
Total	132	218	350

χ^2 Test of Independence: Example

- Given the observed contingency table, what are expected counts?

Daily Flossing	Gum Disease		Total
	Yes	No	
Yes	50	127	177
No	82	91	173
Total	132	218	350

Daily Flossing	Gum Disease		Total
	Yes	No	
Yes			177
No			173
Total	132	218	350

- What is the X^2 statistic?

χ^2 Test of Independence: Example

- Given the observed contingency table, what are expected counts?

Daily Flossing	Gum Disease		Total
	Yes	No	
Yes	50	127	177
No	82	91	173
Total	132	218	350

Daily Flossing	Gum Disease		Total
	Yes	No	
Yes	177*132/350 =66.8	177*218/350 =110.2	177
No	173*132/350 =65.2	173*218/350 =107.8	173
Total	132	218	350

- What is the X^2 statistic?

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(50 - 66.8)^2}{66.8} + \frac{(127 - 110.2)^2}{110.2} + \frac{(82 - 65.2)^2}{65.2} + \frac{(91 - 107.8)^2}{107.8} = 13.659$$

χ^2 Test of Independence: Example

- How many degrees of freedom do we have? $(2-1)(2-1) = 1$ degree of freedom

- What is the p value?

χ^2 Test of Independence: Example

- How many degrees of freedom do we have?
  - $(2 - 1)(2 - 1) = 1$ degree of freedom
- What is the p-value?
  - $P(X^2 \geq 13.659) =$ `1-pchisq(13.659,1)` $= 0.00022$
- Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that there is an association between flossing and gum disease
- Could also use `chisq.test()` in R

χ^2 Test of Independence: R Code

```
> x <- matrix(c(50, 82, 127, 91), nrow=2, ncol=2)
> x
     [,1] [,2]
[1,]   50  127
[2,]   82   91
> chisq.test(x,correct=F)

	Pearson's Chi-squared test

data:  x
X-squared = 13.659, df = 1, p-value = 0.0002192
```

Larger Contingency Tables

- The 2×2 tables that we have talked about thus far are characterized by each variable having only two possible outcomes
  - This is the case comparable to the two-proportion z-test
- We can extend the χ^2 test to accommodate comparison of more than two proportions
  - $r \times c$ tables
  - r categories for the row variable
  - c categories for the column variable
- The inferential procedures are the same for $r \times c$ tables as for 2×2 tables

Larger Contingency Tables: Example

- Instead of yes/no flossing status, let's investigate different brands of floss
  - Oral-B, Colgate, and Reach
- Investigate gum disease prevalence for users of each floss type
- Conduct a study with 260 people who floss
  - Record the floss brand they use and whether they have gum disease
- Hypotheses:
  - H_0 : Floss brand and gum disease are independent
  - H_1 : There is an association between floss brand and gum disease
- Test at the $\alpha = 0.05$ significance level

Larger Contingency Tables: Example

Observed

Daily Flossing	**Gum Disease**		**Total**
	Yes	No	
Oral-B	14	70	84
Colgate	25	71	96
Reach	21	59	80
Total	60	200	260

Expected

Daily Flossing	**Gum Disease**		**Total**
	Yes	No	
Oral-B			84
Colgate			96
Reach			80
Total	60	200	260

Larger Contingency Tables: Example

Observed

Daily Flossing	**Gum Disease**		**Total**
	Yes	No	
Oral-B	14	70	84
Colgate	25	71	96
Reach	21	59	80
Total	60	200	260

Expected

Daily Flossing	**Gum Disease**		**Total**
	Yes	No	
Oral-B	19.39	64.61	84
Colgate	22.15	73.85	96
Reach	18.46	61.54	80
Total	60	200	260

Larger Contingency Tables: Example

- What is the test statistic?

- How many degrees of freedom?
- What is the p-value?

- Conclusion:

Larger Contingency Tables: Example

- What is the test statistic?

$$X^2 = \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
$$= \frac{(14 - 19.39)^2}{19.39} + \frac{(25 - 22.15)^2}{22.15} + \frac{(21 - 18.46)^2}{18.46}$$
$$+ \frac{(70 - 64.61)^2}{64.61} + \frac{(71 - 73.85)^2}{73.85} + \frac{(59 - 61.54)^2}{61.54}$$
$$= 2.8735.$$

- How many degrees of freedom?
  $(r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$

- What is the p-value?
  $P(X^2 > 2.8735) =$ `1-pchisq(2.28735,df=2)` $= 0.2377$

- Conclusion: Since the p-value $p = 0.2377 > \alpha = 0.05$, we fail to reject the null hypothesis; there is not sufficient evidence to conclude that gum disease is associated with brand of floss used

Summary

- Goodness-of-Fit Test
  - Use this to test if the true proportion = expected proportion
  - Generalization of proportion hypothesis tests
- Chi-Squared (χ^2) Test of Independence
  - Use this to check if variables are independent or not
- Both rely on the χ^2 distribution!