Unit 5: Inference for categorical data
3. Chi-square testing

Sta 101 - Spring 2019

Duke University, Department of Statistical Science

Dr. Ellison

Slides posted at
https://www2.stat.duke.edu/courses/Spring19/sta101.001/

Outline

1. Housekeeping

2. Main ideas: Analyses that involve categorical variable(s) with >2 levels
   1. Which analysis to use: Categorical data: 2 levels → Z, >2 levels → χ^2 square
   2. χ^2 Distribution Properties: The χ^2 statistic is always positive and right skewed
   3. Conditions for χ^2 Analyses: At least 5 expected successes for χ^2 testing

3. Application exercises

4. Summary

Announcements

Coming up…

- Problem Set 5 due Wednesday 3/27
- Midterm 2 Review Wednesday 3/27
- Midterm 2 Review Thursday 3/28
- Performance Assessment 5 due Sunday 3/31
- Midterm 2 Monday 4/1

Outline

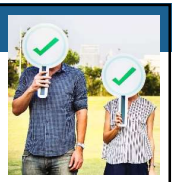
1. Housekeeping

2. Main ideas: Analyses that involve categorical variable(s) with >2 levels
   1. Which analysis to use: Categorical data: 2 levels → Z, >2 levels → χ^2 square
   2. χ^2 Distribution Properties: The χ^2 statistic is always positive and right skewed
   3. Conditions for χ^2 Analyses: At least 5 expected successes for χ^2 testing

3. Application exercises

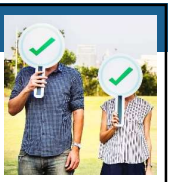
4. Summary

Last week… (5.1 and 5.2)



Categorical Variable(s) involved have only 2 levels

Last week… (5.1 and 5.2)



Categorical Variable(s) involved have only 2 levels

Analyses **HAVE one specified Population Parameter of Interest** and

- We can create a **confidence interval** for the population parameter.
- We can conduct a **hypothesis test** for the population parameter using:
  - Ho: $Pop.Param = null\ value$
  - Ha: $Pop.Param\ (\neq\ or\ <\ or\ >)\ null\ value$

Central Limit Theorem

Confidence Interval for Population Parameter

$$(point\ estimate) \pm (crit.value)SE$$

Types of Variable(s) Involved	Population Parameter	Point Estimate	Standard Error	Distribution: (1) To Get Critical Values From (CI) (2) That the Test Statistic Follows (HT)
Single **Categorical** Variable *(2 levels)*	p	$\hat{p}$	For Hypothesis Tests $\sqrt{\frac{p_0(1-p_0)}{n}}$ For Confidence Intervals $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	Z
-Categorical Response Variable **-Categorical** Explanatory Variable *(both have 2 levels)*	p1-p2	$\hat{p}_1 - \hat{p}_2$	For Hypothesis Tests (with Ho:p1-p2=0) $\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$ For Confidence Intervals $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	Z

Central Limit Theorem

Hypothesis Testing for Population Parameter

$Ho: pop.param = null\ value$
$Ha: pop.param (\neq\ or\ >\ or\ <) null\ value$

$$Test - Stat = \frac{(point\ estimate) - null\ value}{SE}$$

Types of Variable(s) Involved	Population Parameter	Point Estimate	Standard Error	Distribution: (1) To Get Critical Values From (CI) (2) That the Test Statistic Follows (HT)
Single **Categorical** Variable *(2 levels)*	p	$\hat{p}$	For Hypothesis Tests $\sqrt{\frac{p_0(1-p_0)}{n}}$ For Confidence Intervals $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	Z
-Categorical Response Variable **-Categorical** Explanatory Variable *(both have 2 levels)*	p1-p2	$\hat{p}_1 - \hat{p}_2$	For Hypothesis Tests (with Ho:p1-p2=0) $\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$ For Confidence Intervals $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	Z

Today… (5.3)

At least one of the Categorical Variable(s) involved has >2 levels

Today… (5.3)

At least one of the Categorical Variable(s) involved has >2 levels

Analyses that **DON'T HAVE one specified Population Parameter of Interest** and

- No confidence interval framework
- Hypothesis tests set up are different

Types of Variables	Analysis	Hypotheses
Single Categorical Variable (>2 levels)	Chi-Squared Goodness of Fit Test	Ho: The data follows the specified distribution. Ha: The data does not follow the specified distribution.
Categorical Response Variable Categorical Explanatory Variable (at least one has >2 levels)	Chi-Squared Independence Test	Ho: The two variables are independent/not associated. Ha: The two variables are dependent/associated.

Determining which test to use.

Clicker question

In a children's game that teaches kids about probability, the child can randomly select from a set of 2 green balls in the basket (labeled 1 and 2). We have historical data from the toy that indicates how often each of the balls were picked. We want to find out if each number is equally likely to be drawn by the child. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Total
Observed Data (from historical data)	294	296	590

Clicker question

In a children's game that teaches kids about probability, the child can randomly select from a set of 2 green balls in the basket (labeled 1 and 2). We have historical data from the toy that indicates how often each of the balls were picked. We want to find out if each number is equally likely to be drawn by the child. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Total
Observed Data (from historical data)	294	296	590

Clicker question

In a children's game that teaches kids about probability, the child can randomly select from a set of 2 green balls in the basket (labeled 1 and 2). We have historical data from the toy that indicates how often each of the balls were picked. We want to find out if each number is equally likely to be drawn by the child. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

One categorical variable with 2 levels *(ball 1 picked, ball 2 picked)*

Clicker question

In a children's game that teaches kids about probability, the child can randomly select from a set of 2 green balls in the basket (labeled 1 and 2). We have historical data from the toy that indicates how often each of the balls were picked. We want to find out if each number is equally likely to be drawn by the child. Which test is most appropriate?

Ho: $p = 1/2$
Ha: $p \neq 1/2$

p=probability of ball 1 being drawn

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Total
Observed Data (from historical data)	294	296	590

Clicker question

In a children's game that teaches kids about probability, the child can randomly select from a set of 2 green balls in the basket (labeled 1 and 2). We have historical data from the toy that indicates how often each of the balls were picked. We want to find out if each number is equally likely to be drawn by the child. Which test is most appropriate?

Ho: p = 1/2
Ha: p ≠ 1/2

p=probability of ball 1 being drawn
Inference on *p* automatically conducts inference on *(1-p)=probability of ball 2 being drawn*

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Total
Observed Data (from historical data)	294	296	590

Determining which test to use.

https://www.fool.com/retirement/2017/09/13/how-does-the-powerball-annuity-work.aspx

Clicker question

In the basic Powerball game players select 5 numbers from a set of 59 white balls. We have historical data from lottery outcomes such that we are able to calculate how many times each of the 59 white balls were picked. We want to find out if each number is equally likely to be drawn. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

https://www.fool.com/retirement/2017/09/13/how-does-the-powerball-annuity-work.aspx

Clicker question

In the basic Powerball game players select 5 numbers from a set of 59 white balls. We have historical data from lottery outcomes such that we are able to calculate how many times each of the 59 white balls were picked. We want to find out if each number is equally likely to be drawn. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 *test of goodness of fit*
(d) χ^2 test of independence

	Which ball was picked? *(Categorical 59 levels)*
Trial 1	Ball 44
Trial 2	Ball 28
Trial 3	Ball 7
...	...

One categorical variable with >2 levels *(ball 1 picked, ball 2 picked,... ball 59 picked)*

Clicker question

In the basic Powerball game players select 5 numbers from a set of 59 white balls. We have historical data from lottery outcomes such that we are able to calculate how many times each of the 59 white balls were picked. We want to find out if each number is equally likely to be drawn. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

- **One Categorical Variable of Interest (>2 levels):** Powerball outcomes 59 levels.
- Want to compare it to a "specified distribution"

Clicker question

In the basic Powerball game players select 5 numbers from a set of 59 white balls. We have historical data from lottery outcomes such that we are able to calculate how many times each of the 59 white balls were picked. We want to find out if each number is equally likely to be drawn. Which test is most appropriate?

χ^2 test of goodness of fit

Ho: The data follows the specified distribution
Ha: The data does not follow the specified distribution

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590

	Probability of Ball 1 Being Selected	Probability of Ball 2 Being Selected	Probability of Ball 3 Being Selected	...	Probability of Ball 59 Being Selected	Total
Specified Distribution: "Each number is equally likely to be drawn"	?	?	?		?	1

Clicker question

In the basic Powerball game players select 5 numbers from a set of 59 white balls. We have historical data from lottery outcomes such that we are able to calculate how many times each of the 59 white balls were picked. We want to find out if each number is equally likely to be drawn. Which test is most appropriate?

χ^2 test of goodness of fit

Ho: The data follows the specified distribution
Ha: The data does not follow the specified distribution

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590

	Probability of Ball 1 Being Selected	Probability of Ball 2 Being Selected	Probability of Ball 3 Being Selected	...	Probability of Ball 59 Being Selected	Total
Specified Distribution: "Each number is equally likely to be drawn"	1/59	1/59	1/59	...	1/59	1

Clicker question

If the data follows the specified distribution of interest (ie: each number is equally likely to be drawn), then what would be the expected number of times ball 3 would have been selected?

(a) 8
(b) 10
(c) 1/59
(d) 8/59

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590

	Probability of Ball 1 Being Selected	Probability of Ball 2 Being Selected	Probability of Ball 3 Being Selected	...	Probability of Ball 59 Being Selected	Total
Specified Distribution: "Each number is equally likely to be drawn"	1/59	1/59	1/59	...	1/59	1

Clicker question

If the data follows the specified distribution of interest (ie: each number is equally likely to be drawn), then what would be the expected number of times ball 3 would have been selected?

(a) 8
(b) 10= np_3 =590 (1/59)
(c) 1/59
(d) 8/59

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590

	Probability of Ball 1 Being Selected	Probability of Ball 2 Being Selected	Probability of Ball 3 Being Selected	...	Probability of Ball 59 Being Selected	Total
Specified Distribution: "Each number is equally likely to be drawn"	1/59	1/59	1/59	...	1/59	1

χ^2 test of goodness of fit

Ho: The data does resemble a specified distribution
Ha: The data does not resemble a specified distribution

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590
Data Expected from the Specified Distribution: "Each number is equally likely to be drawn"						590

	Probability of Ball 1 Being Selected	Probability of Ball 2 Being Selected	Probability of Ball 3 Being Selected	...	Probability of Ball 59 Being Selected	Total
Specified Distribution: "Each number is equally likely to be drawn"	1/59	1/59	1/59	...	1/59	1

χ^2 test of goodness of fit

Ho: The data does resemble a specified distribution
Ha: The data does not resemble a specified distribution

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590
Data Expected from the Specified Distribution: "Each number is equally likely to be drawn"	10 (ie: 590 *(1/59))	10 (ie: 590 *(1/59))	10 (ie: 590 *(1/59))	...	10 (ie: 590 *(1/59))	590

	Probability of Ball 1 Being Selected	Probability of Ball 2 Being Selected	Probability of Ball 3 Being Selected	...	Probability of Ball 59 Being Selected	Total
Specified Distribution: "Each number is equally likely to be drawn"	1/59	1/59	1/59	...	1/59	1

Determining which test to use.

Clicker question

A Gallup poll asked whether or not respondents identify as Tea Party Republican (yes / no) and whether or not they are motivated to vote in the upcoming midterm election (yes / no). We want to find out whether being a Tea Party Republican is associated with motivation to vote. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

Clicker question

A Gallup poll asked whether or not respondents identify as Tea Party Republican (yes / no) and whether or not they are motivated to vote in the upcoming midterm election (yes / no). We want to find out whether being a Tea Party Republican is associated with motivation to vote. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

- Categorical variable with 2 levels
- Categorical variable with 2 levels

Clicker question

A Gallup poll asked whether or not respondents identify as Tea Party Republican (yes / no) and whether or not they are motivated to vote in the upcoming midterm election (yes / no). We want to find out whether being a Tea Party Republican is **associated** with motivation to vote. Which test is most appropriate?

H_0: $p_{TPR} = p_{Other}$,
H_a: $p_{TPR} \neq p_{Other}$,

p_{TPR} = *probability/proportion of all TPR members motivated to vote*
p_{Other} = *probability/proportion of all non-TPR members motivated to vote*

Clicker question

A Gallup poll asked whether or not respondents identify as Tea Party Republican (yes / no) and whether or not they are motivated to vote in the upcoming midterm election (yes / no). We want to find out whether being a Tea Party Republican is **associated** with motivation to vote. Which test is most appropriate?

H_0: $p_{TPR} = p_{Other}$,
H_a: $p_{TPR} \neq p_{Other}$,

Equivalent to saying

H_0: $P(motivated\ to\ vote|TPR) = P(motivated\ to\ vote|other)$,
H_a: $P(motivated\ to\ vote|TPR) \neq P(motivated\ to\ vote|other)$

Clicker question

A Gallup poll asked whether or not respondents identify as Tea Party Republican (yes / no) and whether or not they are motivated to vote in the upcoming midterm election (yes / no). We want to find out whether being a Tea Party Republican is **associated** with motivation to vote. Which test is most appropriate?

$H_0: p_{TPR} = p_{Other}$,
$H_a: p_{TPR} \neq p_{Other}$,

Equivalent to saying

$H_0: P(\text{motivated to vote} | TPR) = P(\text{motivated to vote} | other)$,
$H_a: P(\text{motivated to vote} | TPR) \neq P(\text{motivated to vote} | other)$

Equivalent to saying

H_0: Tea Party affiliation is **not associated (or independent)** with being motivated to vote
H_a: Tea Party affiliation is **associated (or dependent)** with being motivated to vote

Determining which test to use.

Clicker question

Suppose the Gallup poll instead asked about

- party affiliation (Tea Party Republican, Other Republican, and Non-Republican), and
- motivation to vote (extremely unmotivated, very unmotivated, unmotivated, motivated, very motivated, extremely motivated)

We want to find out whether party affiliation is associated with motivation to vote. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence

Clicker question

Suppose the Gallup poll instead asked about

- party affiliation (Tea Party Republican, Other Republican, and Non-Republican), and
- motivation to vote (extremely unmotivated, very unmotivated, unmotivated, motivated, very motivated, extremely motivated)

We want to find out whether party affiliation is **associated** with motivation to vote. Which test is most appropriate?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 *test of independence*

- Categorical variable with >2 levels
- Categorical variable with >2 levels

H_0: *Party affiliation and motivation to vote are independent/not associated*
H_a: *Party affiliation and motivation to vote are dependent/associated*

Outline

1. Housekeeping

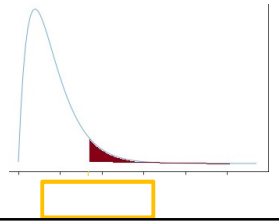
2. Main ideas: Analyses that involve categorical variable(s) with >2 levels
   1. Which analysis to use: Categorical data: 2 levels → Z, >2 levels → χ^2 square
   2. χ^2 Distribution Properties: The χ^2 statistic is always positive and right skewed
   3. Conditions for χ^2 Analyses: At least 5 expected successes for χ^2 testing

3. Application exercises

4. Summary

What is the test statistic for:

- **Chi-Squared Goodness of Fit Test?**
- **Chi-Squared Independence Test?**



What is the test statistic for:

- **Chi-Squared Goodness of Fit Test?**
- **Chi-Squared Independence Test?**

It's the same calculation!

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$ where k = total number of cells

The χ^2 statistic

χ^2 statistic: When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the *chi-square (χ^2) statistic*:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$ where k = total number of cells

***O**bserved # in cell i* ***E**xpected # in cell i*

	# of times Ball 1 is Selected	# of times Ball 2 is Selected	…	# of times Ball 59 is Selected	Total
Observed Data	9	11		12	590
(Expected Data)	(10)	(10)	…	(10)	(590)

Ex: Chi-Squared GOF, 59 cells

The χ^2 statistic

χ^2 statistic: When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the *chi-square (χ^2) statistic*:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$ where k = total number of cells

***O**bserved # in cell i*

***E**xpected # in cell i*

	dating	cohabiting	married	total
obese	81 (113)	103 (110)	147 (108)	331
not obese	359 (327)	326 (319)	277 (316)	962
total	440	429	424	1293

Ex: Chi-Squared Independence Test, 6 cells

The χ^2 statistic

χ^2 statistic: When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the *chi-square (χ^2) statistic*:

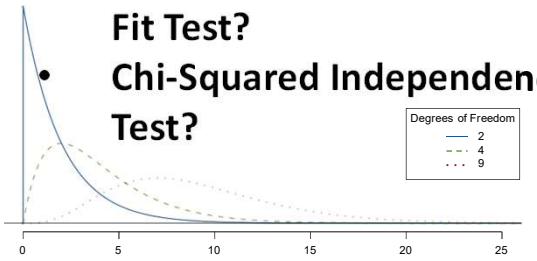
$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$ where k = total number of cells

Important points:

- Use **counts** (not **proportions**) in the calculation of the text statistic, even though we're truly interested in the proportions for inference
- Expected counts are calculated assuming the null hypothesis is true

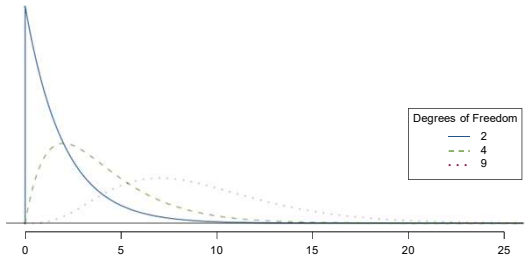
What are the degrees of freedom for the test statistic for a:

- Chi-Squared Goodness of Fit Test?
- Chi-Squared Independence Test?



The χ^2 distribution

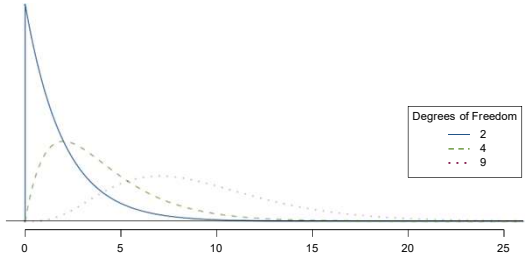
The χ^2 distribution has just one parameter, *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.



The χ^2 distribution

The χ^2 distribution has just one parameter, *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

- **For χ^2 GOF test**: $df = k - 1$
- **For χ^2 independence test:** $df = (R - 1) \times (C - 1)$



The χ^2 distribution

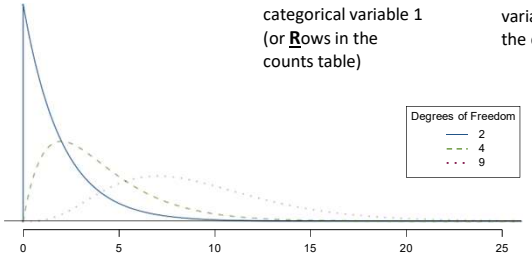
The χ^2 distribution has just one parameter, *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

- **For χ^2 GOF test**: $df = k - 1$
- **For χ^2 independence test**: $df = (R - 1) \times (C - 1)$

# of levels of the one categorical variable

# of levels of categorical variable 1 (or **R**ows in the counts table)

# of levels of categorical variable 2 (or **C**olumns in the counts table)



Finding areas under the chi-square curve

p-value = P($\chi^2 \geq \chi^2$- *statistic)*
=tail area under the chi-square distribution (as usual)

Finding areas under the chi-square curve

p-value = P($\chi^2 \geq \chi^2$- *statistic)*
=tail area under the chi-square distribution (as usual)

- Using the applet: *https://gallery.shinyapps.io/dist_calc/*

Finding areas under the chi-square curve

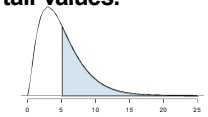
p-value = P($\chi^2 \geq \chi^2$-*statistic)*
=tail area under the chi-square distribution (as usual)

- Using the applet: *https://gallery.shinyapps.io/dist_calc/*
- Using R: `pchisq()`

Finding areas under the chi-square curve

p-value = P($\chi^2 \geq \chi^2$-*statistic)*
=tail area under the chi-square distribution (as usual)

- Using the applet: *https://gallery.shinyapps.io/dist_calc/*
- Using R: `pchisq()`
- Using the table: works a lot like the *t* table, but **only provides upper tail values.**



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
	…								

Outline

1. Housekeeping

2. Main ideas: Analyses that involve categorical variable(s) with >2 levels
   1. Which analysis to use: Categorical data: 2 levels → Z, >2 levels → χ^2 square
   2. χ^2 Distribution Properties: The χ^2 statistic is always positive and right skewed
   3. Conditions for χ^2 Analyses: At least 5 expected successes for χ^2 testing

3. Application exercises

4. Summary

What are the conditions for:

- **Chi-Squared Goodness of Fit Test?**
- **Chi-Squared Independence Test?**

Conditions for χ^2 IndependenceTest

1. *Independence:*
    1. For each population: random sampling/assignment
    2. n_1<10% of population 1, n_2<10% of population 2,
    3. Each case contributes to only one cell in the table.
2. *Sample size / distribution:*
    1. Each **expected cell** must have at least 5 *expected* cases.

Conditions for χ^2 testing

Conditions for χ^2 IndependenceTest

1. *Independence:*
    1. For each population: random sampling/assignment
    2. n_1<10% of population 1, n_2<10% of population 2,
    3. Each case contributes to only one cell in the table.
2. *Sample size / distribution:*
    1. Each **expected cell** must have at least 5 *expected* cases.

	Number of Times Ball 1 is Selected	Number of Times Ball 2 is Selected	Number of Times Ball 3 is Selected	...	Number of Times Ball 59 is Selected	Total
Observed Data (from historical data)	9	11	8	...	12	590
Data Expected from the Specified Distribution: "Each number is equally likely to be drawn"	10 (ie: 590 *(1/59))	10 (ie: 590 *(1/59))	10 (ie: 590 *(1/59))	...	10 (ie: 590 *(1/59))	590

Conditions for χ^2 testing

Conditions for χ^2 IndependenceTest

1. *Independence:*
    1. For each population: random sampling/assignment
    2. n<10% of population
    3. Each case contributes to only one cell in the table.
2. *Sample size / distribution:*
    1. Each cell must have at least 5 *expected* cases.

	dating	cohabiting	married	total
obese	81 (113)	103 (110)	147 (108)	331
not obese	359 (327)	326 (319)	277 (316)	962
total	440	429	424	1293

Determining which test to use.

Clicker question

Suppose a poll asked the following questions:

- How would you identify your socio-economic status: low, middle, high?
- What type of pet did you have growing up, select all that apply: cat, dog, fish, bird, rodent, none of the above?

What test is most appropriate for evaluating the relationship between these two variables?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence
(e) none of the above

Clicker question

Suppose a poll asked the following questions:

- How would you identify your socio-economic status: low, middle, high?
- What type of pet did you have growing up, **select all that apply**: cat, dog, fish, bird, rodent, none of the above?

What test is most appropriate for evaluating the relationship between these two variables?

(a) Z test for a single proportion
(b) Z test for comparing two proportions
(c) χ^2 test of goodness of fit
(d) χ^2 test of independence
(e) none of the above

Condition Violated: Each case MUST contribute to only one cell in the table.

Clicker question

Suppose a poll asked the following questions:

- How would you identify your socio-economic status: low, middle, high?
- What type of pet did you have growing up, **select all that apply**: cat, dog, fish, bird, rodent, none of the above?

What test is most appropriate for evaluating the relationship between these two variables?

	Had Dog	Had Cat	Had Fish	Had Bird	Had Rodent	Had None	Total
High Socioeconomic Status							
Middle Socioeconomic Status	1	1					
Low Socioeconomic Status							
Total							

Ex: I had dogs and cats growing up. My case contributed twice.-> CONDITIONS FOR χ^2 TEST OF INDEPENDENCE VIOLATED

Application exercise: 5.3 Chi-square tests

See course website for details.

Summary of main ideas

1. Categorical data: 2 levels → Z, >2 levels → χ^2 square
2. The χ^2 statistic is always positive and right skewed
3. At least 5 expected successes for χ^2 testing