Comparing Multiple Proportions

Research Objective

Research Question: Is there a relationship between which college someone is in and whether they use an apple or android phone?

Population: All BYU students.

Parameter of Interest:

- We have a lot! We want to know the proportion of students in each college/phone combination. For example, $\pi_{\text{Apple,Humanities}}$.

Sample: A convenience sample of 1575 BYU students who are in my class and completed the student survey.

Are there any issues with this study setup?

More Problem Definitions

Response Variable (y):

- Does the student have an Apple or Android phone. This is a **categorical variable** meaning it has to be one of a certain number of categories.

Explanatory Variable (x):

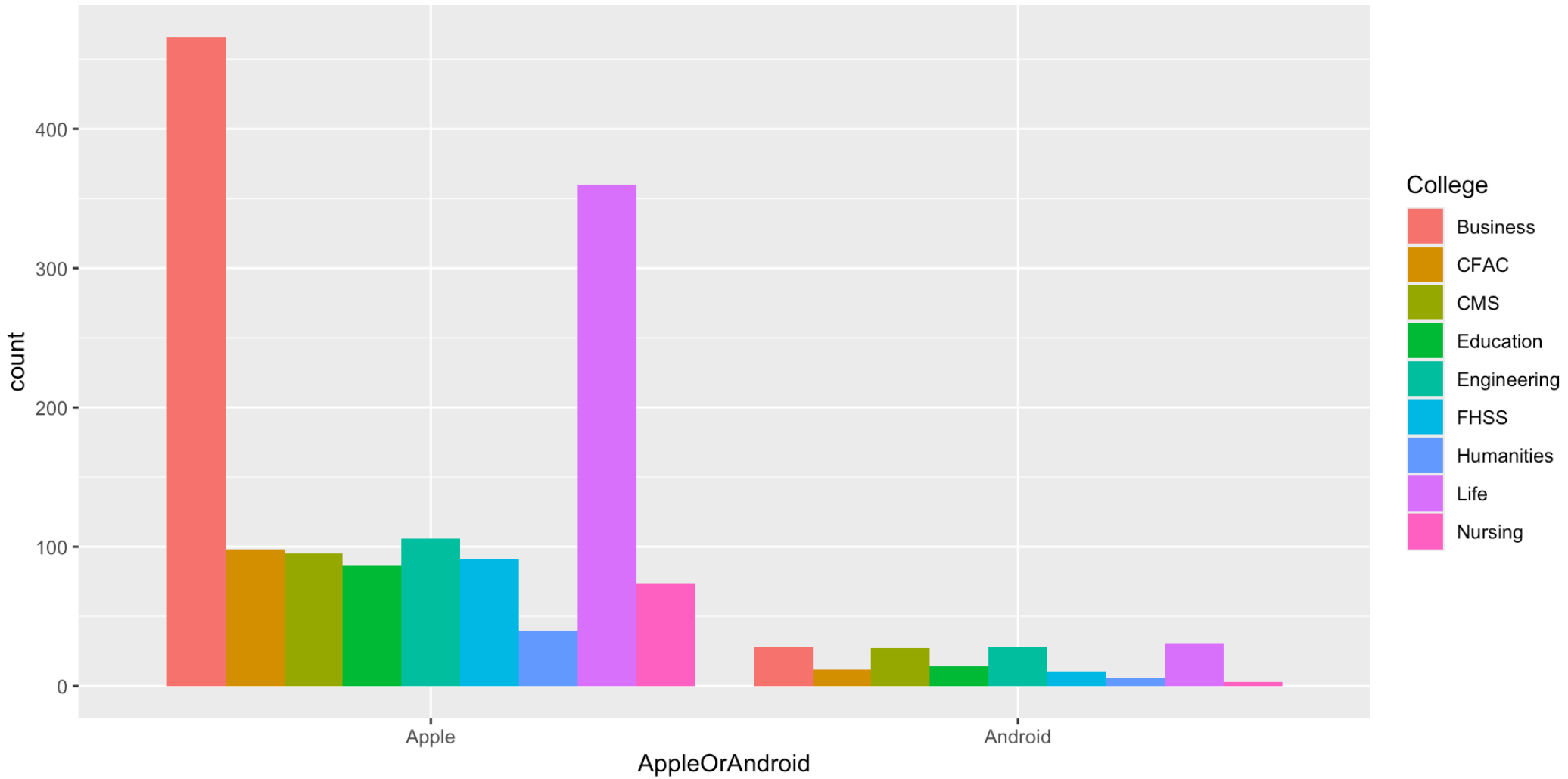
- The college.

Exploratory Data Analysis (EDA)

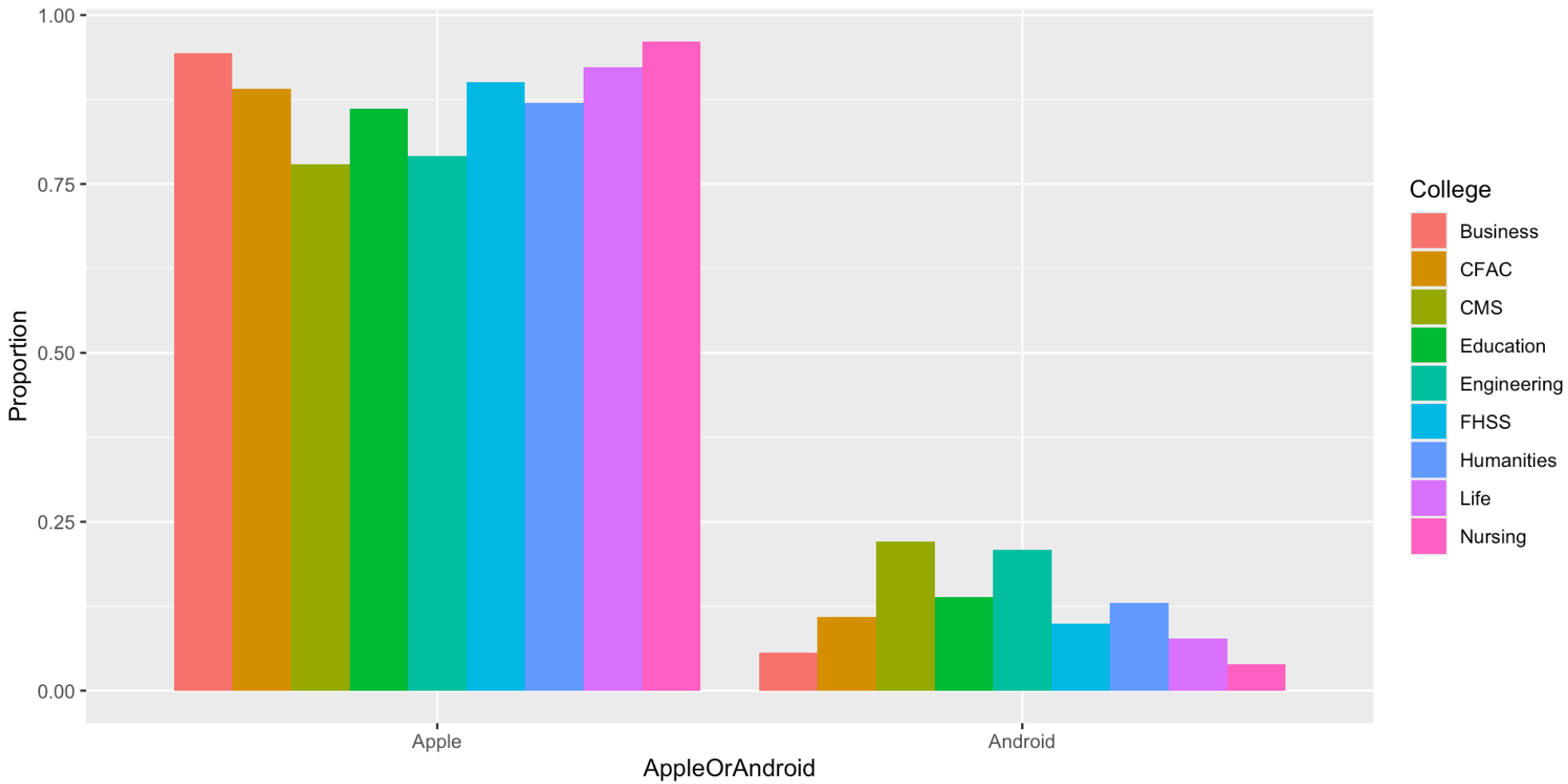
Main goal: Examine the RELATIONSHIP between College and Phone.

AppleOrAndroid	College
Apple	Nursing
Apple	Business
Apple	Business
Apple	Nursing
Apple	Business

EDA Tool #1 - Grouped Bar Charts



EDA Tool #2 - Grouped Bar Charts Proportions



EDA Tool #3 - Tables of Counts

	Apple	Android	Sum
Business	466	28	494
CFAC	98	12	110
CMS	95	27	122
Education	87	14	101
Engineering	106	28	134
FHSS	91	10	101
Humanities	40	6	46
Life	360	30	390
Nursing	74	3	77
Sum	1417	158	1575

EDA Tool #4 - Conditional and Marginal Distributions

Main Idea: Convert counts to proportions to account for differences in count sizes

Conditional Distribution of Row Variable given Column Variable:

- proportions sum to 1 down the rows
- divide cell counts by column totals

Conditional Distribution of Column Variable given Row Variable:

- proportions sum to 1 across the columns
- divide cell counts by row totals

EDA Tool #4 - Conditional and Marginal Distributions

Marginal Distribution of Column (or Row)

- proportions sum to 1 across total column (or row)
- divide column (or row) totals by table total

Relationship between variables is probably present if conditionals are different than marginal distributions.

Cond. Dists of Col. Given Row

	Apple	Android
Business	0.943	0.057
CFAC	0.891	0.109
CMS	0.779	0.221
Education	0.861	0.139
Engineering	0.791	0.209
FHSS	0.901	0.099
Humanities	0.870	0.130
Life	0.923	0.077
Nursing	0.961	0.039
Margin (Overall)	0.900	0.100

Cond. Dists of Row Given Col.

	Apple	Android	Margin (Overall)
Business	0.329	0.177	0.314
CFAC	0.069	0.076	0.070
CMS	0.067	0.171	0.077
Education	0.061	0.089	0.064
Engineering	0.075	0.177	0.085
FHSS	0.064	0.063	0.064
Humanities	0.028	0.038	0.029
Life	0.254	0.190	0.248
Nursing	0.052	0.019	0.049

Practice: Age vs. Distracted Driving

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	58	159	96	69	48	21	5	456
Not Distracted	2962	11278	8382	7328	7482	5282	4341	47055
Other Distracted	303	898	586	400	415	288	282	3172
Sum	3323	12335	9064	7797	7945	5591	4628	50683

Of those cell phone distracted drivers, what proportion are 15-19?

- 58/456

Is this a conditional or marginal proportion?

- Conditional

Practice: Age vs. Distracted Driving

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	58	159	96	69	48	21	5	456
Not Distracted	2962	11278	8382	7328	7482	5282	4341	47055
Other Distracted	303	898	586	400	415	288	282	3172
Sum	3323	12335	9064	7797	7945	5591	4628	50683

What is the conditional distribution of age for those who are cell phone distracted?

Practice: Age vs. Distracted Driving

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	58	159	96	69	48	21	5	456
Not Distracted	2962	11278	8382	7328	7482	5282	4341	47055
Other Distracted	303	898	586	400	415	288	282	3172
Sum	3323	12335	9064	7797	7945	5591	4628	50683

What is the conditional distribution of age for those who are cell phone distracted?

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	0.127	0.349	0.211	0.151	0.105	0.046	0.011	1

Practice: Age vs. Distracted Driving

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	58	159	96	69	48	21	5	456
Not Distracted	2962	11278	8382	7328	7482	5282	4341	47055
Other Distracted	303	898	586	400	415	288	282	3172
Sum	3323	12335	9064	7797	7945	5591	4628	50683

What is the conditional distribution of age for those who are cell phone distracted?

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	0.127	0.349	0.211	0.151	0.105	0.046	0.011	1
Not Distracted	0.063	0.240	0.178	0.156	0.159	0.112	0.092	1
Other Distracted	0.096	0.283	0.185	0.126	0.131	0.091	0.089	1
Margin (Overall)	0.066	0.243	0.179	0.154	0.157	0.110	0.091	1

Practice: Age vs. Distracted Driving

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	58	159	96	69	48	21	5	456
Not Distracted	2962	11278	8382	7328	7482	5282	4341	47055
Other Distracted	303	898	586	400	415	288	282	3172
Sum	3323	12335	9064	7797	7945	5591	4628	50683

What is the conditional distribution of distracted for those aged 20-29?

Practice: Age vs. Distracted Driving

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	58	159	96	69	48	21	5	456
Not Distracted	2962	11278	8382	7328	7482	5282	4341	47055
Other Distracted	303	898	586	400	415	288	282	3172
Sum	3323	12335	9064	7797	7945	5591	4628	50683

What is the conditional distribution of distracted for those aged 20-29?

	20-29
Cell Phone Distracted	0.013
Not Distracted	0.914
Other Distracted	0.073
Sum	1.000

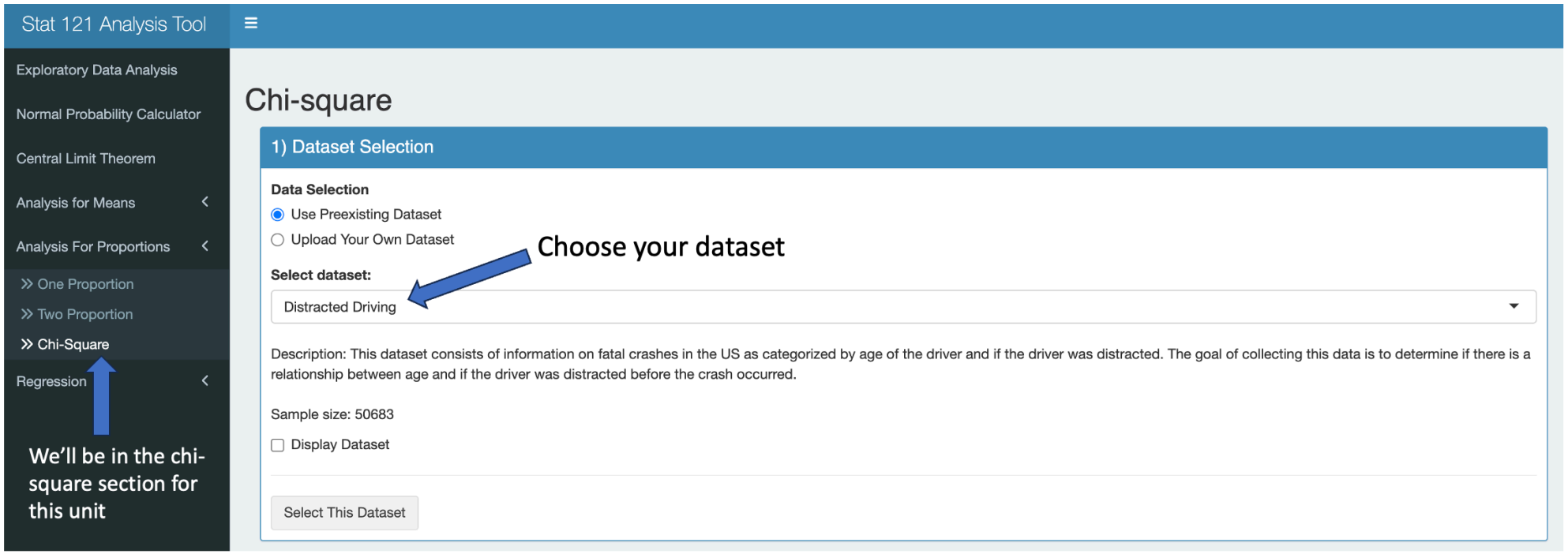
Practice: Age vs. Distracted Driving

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	58	159	96	69	48	21	5	456
Not Distracted	2962	11278	8382	7328	7482	5282	4341	47055
Other Distracted	303	898	586	400	415	288	282	3172
Sum	3323	12335	9064	7797	7945	5591	4628	50683

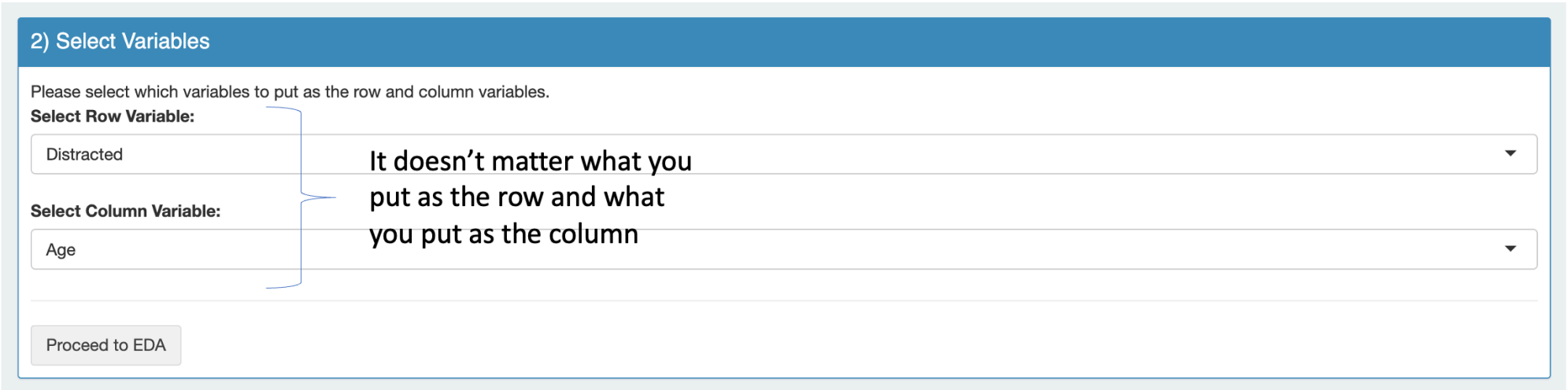
What is the conditional distribution of distracted for those aged 20-29?

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Margin (Overall)
Cell Phone Distracted	0.017	0.013	0.011	0.009	0.006	0.004	0.001	0.009
Not Distracted	0.891	0.914	0.925	0.940	0.942	0.945	0.938	0.928
Other Distracted	0.091	0.073	0.065	0.051	0.052	0.052	0.061	0.063

Using the Analysis Tool



Using the Analysis Tool

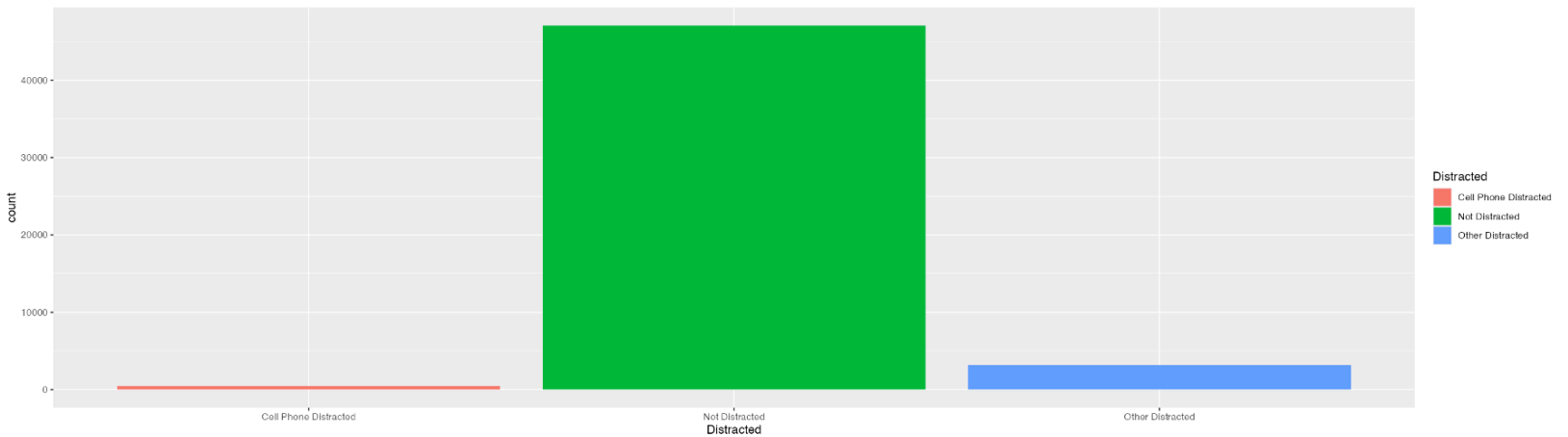


Using the Analysis Tool

3) Exploratory Data Analysis

What graphical summaries would you like to display?

Barplot of Row Variable ← Choose a graph you want to see



What numerical summaries would you like to calculate?

Two Way Table of Counts ← Choose numbers you want to calculate (a two-way table is the best place to start)

Show 10 entries Search:

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	58	159	96	69	48	21	5	456
Not Distracted	2962	11278	8382	7328	7482	5282	4341	47055
Other Distracted	303	898	586	400	415	288	282	3172
Sum	3323	12335	9064	7797	7945	5591	4628	50683

Showing 1 to 4 of 4 entries Previous 1 Next

Proceed to Statistical Inference

Using the Analysis Tool

Of those drivers aged 30-39, what proportion are not distracted?

What numerical summaries would you like to calculate?

Conditional Distribution of Row given Column

Choose “given column” because the question is asking about a specific column

Show 10 **entries** **Search:**

	15-19	20-29	30-39	40-49	50-59	60-69	70+
Cell Phone Distracted	0.0175	0.0129	0.0106	0.0088	0.006	0.0038	0.0011
Not Distracted	0.8914	0.9143	0.9248	0.9398	0.9417	0.9447	0.938
Other Distracted	0.0912	0.0728	0.0647	0.0513	0.0522	0.0515	0.0609
Sum	1.0001	1	1.0001	0.9999	0.9999	1	0.9999999999999999

Showing 1 to 4 of 4 entries

Previous 1 Next

Using the Analysis Tool

Of those “other distracted” drivers, what proportion are 60-69?

What numerical summaries would you like to calculate?

Conditional Distribution of Column given Row

Choose “given row” because the question is asking about a specific row

Show 10 **entries** **Search:**

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	0.1272	0.3487	0.2105	0.1513	0.1053	0.0461	0.011	1.0001
Not Distracted	0.0629	0.2397	0.1781	0.1557	0.159	0.1123	0.0923	1
Other Distracted	0.0955	0.2831	0.1847	0.1261	0.1308	0.0908	0.0889	0.9999

Showing 1 to 3 of 3 entries

Previous 1 Next

Statistical Model (Population)

The independence population model: The choice of apple vs. android product for a student is independent of the college of the student. In other words, the two variables are independent of each other.

Back to the Phone Example

	Apple	Android	Sum
Business	466	28	494
CFAC	98	12	110
CMS	95	27	122
Education	87	14	101
Engineering	106	28	134
FHSS	91	10	101
Humanities	40	6	46
Life	360	30	390
Nursing	74	3	77
Sum	1417	158	1575

Consequences of Independent Population Model

1. Because of independence…

$$\begin{aligned}\Pr(\text{Apple \& Business}) &= \Pr(\text{Apple})\Pr(\text{Business}) \\ &= (1417/1575) \times (494/1575) \\ &= 0.282\end{aligned}$$

2. IF variables are independent, expected number of people in each cell:

$$\begin{aligned}\text{Exp. No. of Apple/Business} &= n \times \Pr(\text{Apple})\Pr \times (\text{Business}) \\ &= 1575 \times 0.282 \\ &= 444.443\end{aligned}$$

Independence Model Practice

	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	58	159	96	69	48	21	5	456
Not Distracted	2962	11278	8382	7328	7482	5282	4341	47055
Other Distracted	303	898	586	400	415	288	282	3172
Sum	3323	12335	9064	7797	7945	5591	4628	50683

1. Under the independence model, what is the probability of being 15-19 and not distracted?

- $(3323/50683) \times (47055/50683) = 0.061$

Independence Model Practice

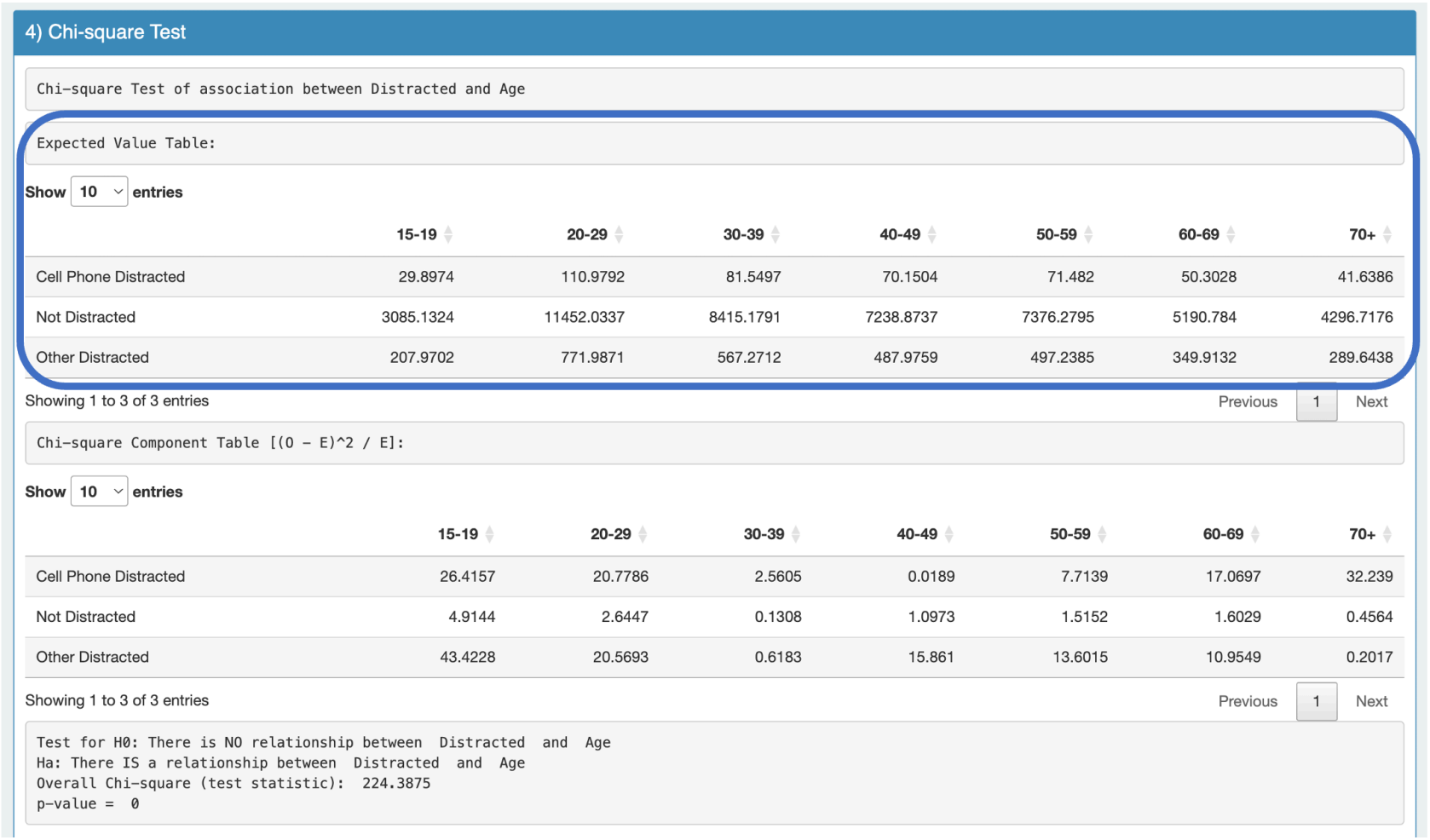
	15-19	20-29	30-39	40-49	50-59	60-69	70+	Sum
Cell Phone Distracted	58	159	96	69	48	21	5	456
Not Distracted	2962	11278	8382	7328	7482	5282	4341	47055
Other Distracted	303	898	586	400	415	288	282	3172
Sum	3323	12335	9064	7797	7945	5591	4628	50683

2. Under the independence model, what is the expected number of 15-19 year old drivers who are not distracted?

- $0.061 \times 50683 = 3085.132$

Using the Analysis Tool

Good news! The tool will calculate the expected counts for you. You just need to know where to look…



4) Chi-square Test

```
Chi-square Test of association between Distracted and Age
```

```
Expected Value Table:
```

Show 10 entries

	15-19	20-29	30-39	40-49	50-59	60-69	70+
Cell Phone Distracted	29.8974	110.9792	81.5497	70.1504	71.482	50.3028	41.6386
Not Distracted	3085.1324	11452.0337	8415.1791	7238.8737	7376.2795	5190.784	4296.7176
Other Distracted	207.9702	771.9871	567.2712	487.9759	497.2385	349.9132	289.6438

Showing 1 to 3 of 3 entries Previous 1 Next

```
Chi-square Component Table [(O - E)^2 / E]:
```

Show 10 entries

	15-19	20-29	30-39	40-49	50-59	60-69	70+
Cell Phone Distracted	26.4157	20.7786	2.5605	0.0189	7.7139	17.0697	32.239
Not Distracted	4.9144	2.6447	0.1308	1.0973	1.5152	1.6029	0.4564
Other Distracted	43.4228	20.5693	0.6183	15.861	13.6015	10.9549	0.2017

Showing 1 to 3 of 3 entries Previous 1 Next

```
Test for H0: There is NO relationship between  Distracted  and  Age
Ha: There IS a relationship between  Distracted  and  Age
Overall Chi-square (test statistic):  224.3875
p-value =  0
```

Hypothesis Testing

Recall the 3 steps of hypothesis testing:

- Formulate hypotheses
- See if data matches (or doesn’t) match the hypotheses
- Draw a conclusions about the parameter

Hypothesis Testing

Research Question: Is there a relationship between which college someone is in and whether they use an apple or android phone?

The hypotheses:

$$H_0 : \text{College and Phone are Independent}$$
$$H_a : \text{College and Phone are NOT Independent}$$

Hypothesis Testing

Step 2: See if the data matches the hypotheses.

How can we compare our observed data to hypotheses?

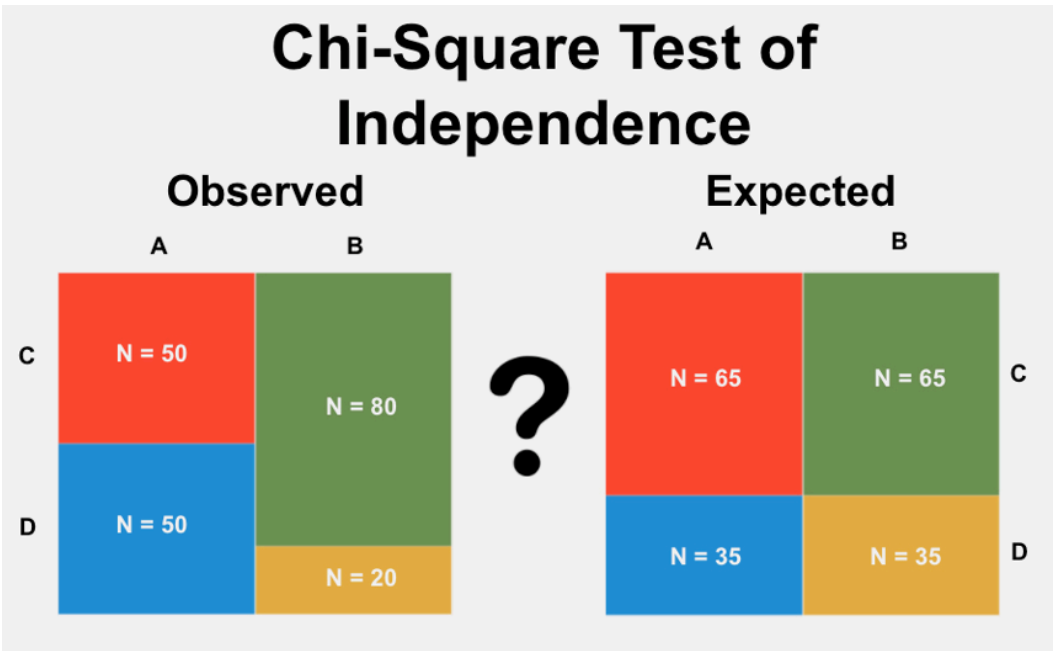
- Compare our data to what we expect to see IF the variables are independent.

Hypothesis Testing

Step 2: See if the data matches the hypotheses.

How can we compare our observed data to hypotheses?

- Compare our data to what we expect to see IF the variables are independent.



Hypothesis Testing

Step 2: See if the data matches the hypotheses.

How can we compare our observed data to hypotheses?

- Compare our data to what we expect to see IF the variables are independent.

The χ^2-statistic: (pronounced “kai-squared”)

$$\begin{aligned}\chi^2 &= \sum_{r=1}^{R}\sum_{c=1}^{C}\chi_{rc}^2 \\ &= \sum_{r=1}^{R}\sum_{c=1}^{C}\frac{(\mathrm{Obs}_{rc}-\mathrm{Exp}_{rc})^2}{\mathrm{Exp}_{rc}}\end{aligned}$$

Hypothesis Testing

Step 2: See if the data matches the hypotheses.

$$\chi^2 = \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(\text{Obs}_{rc} - \text{Exp}_{rc})^2}{\text{Exp}_{rc}}$$

Intuition

- If χ^2 is big, then the data favor H_a because what you observed is different than what you expected to observe IF H_0 was true.
- If any individual cell χ^2_{rc} is big, then that observed count is very different from what you expected it to be if H_0 were true.

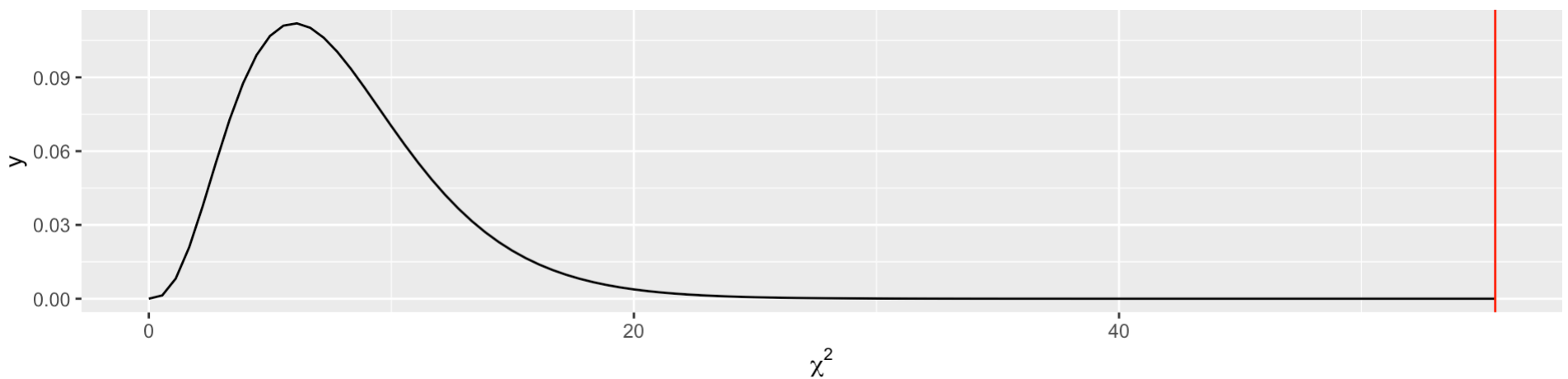
Hypothesis Testing

Step 2: See if the data matches the hypotheses.

Theorem. Sampling Distribution of Chi-2

If the independence model is appropriate AND all expected counts are > 5, then the χ^2 values that you should get when sampling follows an χ^2-distribution.

I am NOT going to get into details of what the χ^2 distribution is (it's technical) it looks like this.

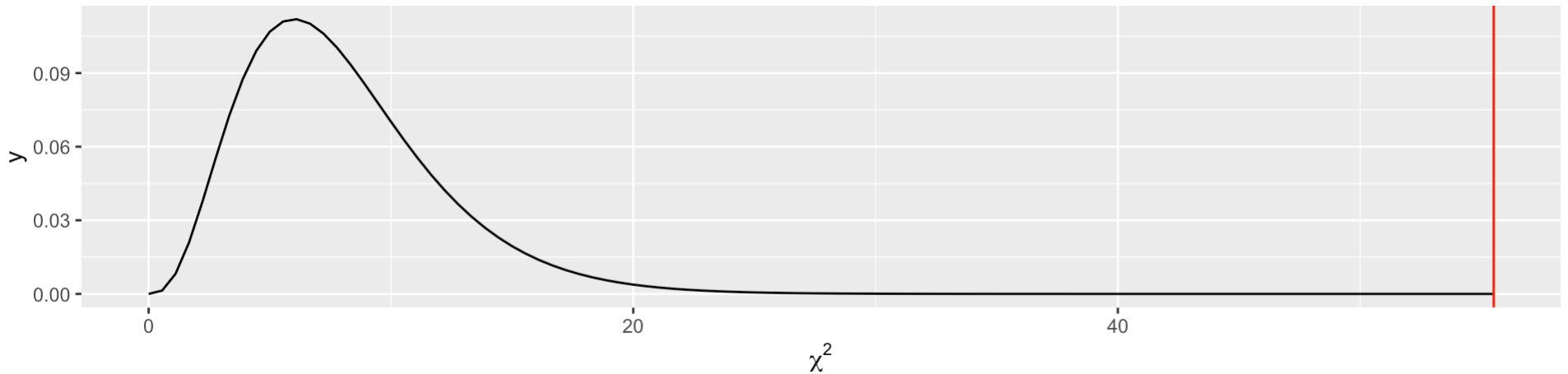


Hypothesis Testing

Step 2: See if the data matches the hypotheses (FIRST - check to make sure all expected counts > 5)

	Apple	Android
Business	444.443	49.557
CFAC	98.965	11.035
CMS	109.761	12.239
Education	90.868	10.132
Engineering	120.557	13.443
FHSS	90.868	10.132
Humanities	41.385	4.615
Life	350.876	39.124
Nursing	69.276	7.724

Hypothesis Testing



Step 2: See if the data matches the hypotheses.

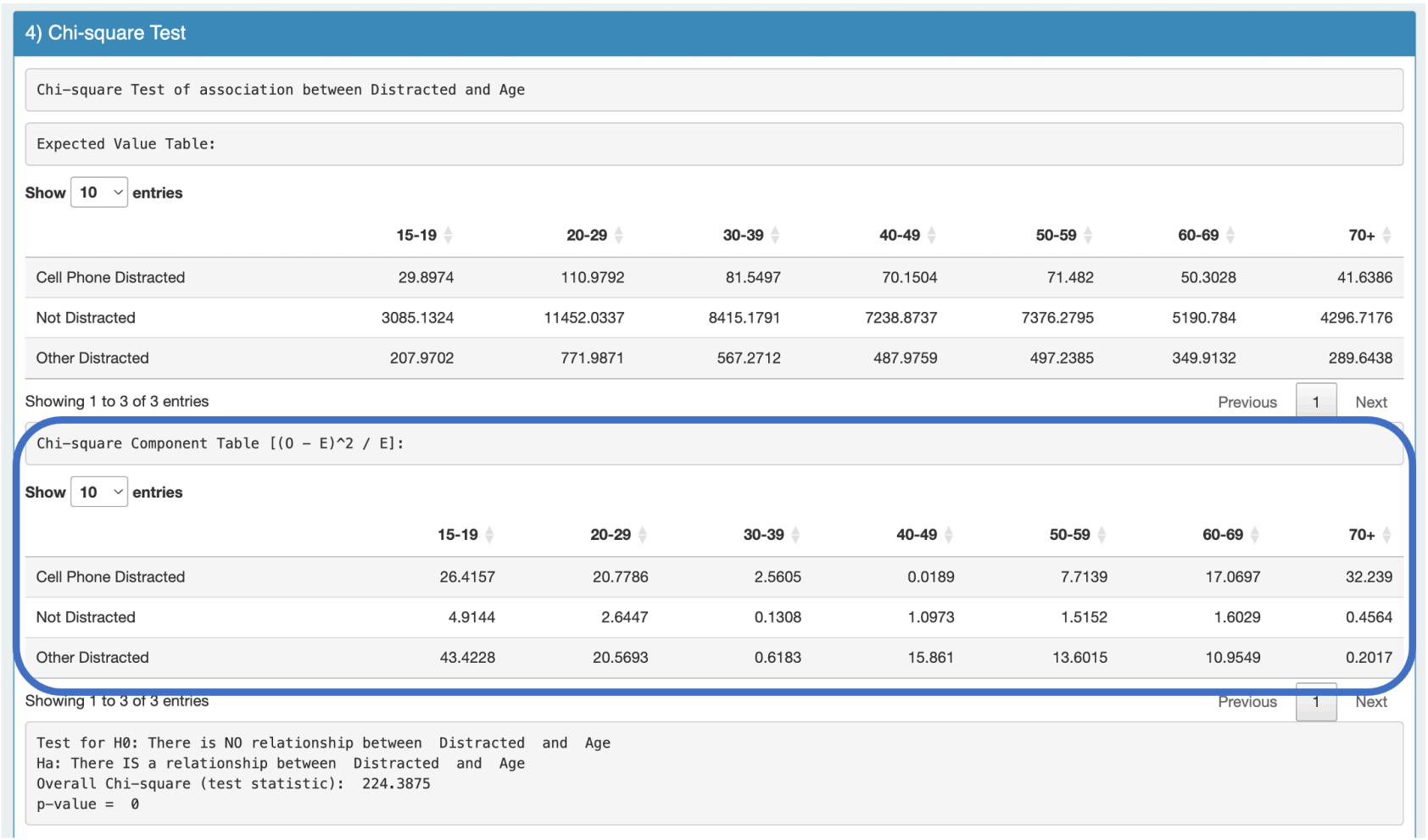
- $\chi^2 = 55.5102$
- p-value $= 0$

What is your conclusion at the $\alpha = 0.05$ level?

- The data are inconsistent with the null hypothesis so we conclude that the college and phone variables are NOT independent.

Using the Analysis Tool

The tool calculates the χ^2_{rc} values for you:



4) Chi-square Test

Chi-square Test of association between Distracted and Age

Expected Value Table:

Show 10 entries

	15-19	20-29	30-39	40-49	50-59	60-69	70+
Cell Phone Distracted	29.8974	110.9792	81.5497	70.1504	71.482	50.3028	41.6386
Not Distracted	3085.1324	11452.0337	8415.1791	7238.8737	7376.2795	5190.784	4296.7176
Other Distracted	207.9702	771.9871	567.2712	487.9759	497.2385	349.9132	289.6438

Showing 1 to 3 of 3 entries Previous 1 Next

Chi-square Component Table [(O - E)^2 / E]:

Show 10 entries

	15-19	20-29	30-39	40-49	50-59	60-69	70+
Cell Phone Distracted	26.4157	20.7786	2.5605	0.0189	7.7139	17.0697	32.239
Not Distracted	4.9144	2.6447	0.1308	1.0973	1.5152	1.6029	0.4564
Other Distracted	43.4228	20.5693	0.6183	15.861	13.6015	10.9549	0.2017

Showing 1 to 3 of 3 entries Previous 1 Next

Test for H0: There is NO relationship between Distracted and Age
Ha: There IS a relationship between Distracted and Age
Overall Chi-square (test statistic): 224.3875
p-value = 0

Using the Analysis Tool

4) Chi-square Test

```
Chi-square Test of association between Distracted and Age
```

```
Expected Value Table:
```

Show 10 entries

	15-19	20-29	30-39	40-49	50-59	60-69	70+
Cell Phone Distracted	29.8974	110.9792	81.5497	70.1504	71.482	50.3028	41.6386
Not Distracted	3085.1324	11452.0337	8415.1791	7238.8737	7376.2795	5190.784	4296.7176
Other Distracted	207.9702	771.9871	567.2712	487.9759	497.2385	349.9132	289.6438

Showing 1 to 3 of 3 entries Previous 1 Next

```
Chi-square Component Table [(O - E)^2 / E]:
```

Show 10 entries

	15-19	20-29	30-39	40-49	50-59	60-69	70+
Cell Phone Distracted	26.4157	20.7786	2.5605	0.0189	7.7139	17.0697	32.239
Not Distracted	4.9144	2.6447	0.1308	1.0973	1.5152	1.6029	0.4564
Other Distracted	43.4228	20.5693	0.6183	15.861	13.6015	10.9549	0.2017

Showing 1 to 3 of 3 entries Previous 1 Next

```
Test for H0: There is NO relationship between  Distracted  and  Age
Ha: There IS a relationship between  Distracted  and  Age
Overall Chi-square (test statistic):  224.3875
p-value =  0
```

Chi-square test statistic

p-value

Following up on χ^2 Test

IF you reject H_0, what can we say about where the relationship is? In other words, where are observed counts most different from expected counts?

- Check the individual cell χ^2 values.

Following up on χ^2 Test

Chi-square Vals

	Apple	Andr
Business	1.0	
CFAC	0.0	
CMS	2.0	1
Education	0.2	
Engineering	1.8	1
FHSS	0.0	
Humanities	0.0	
Life	0.2	
Nursing	0.3	

Obs. Counts

	Apple	Andr
Business	466	
CFAC	98	
CMS	95	
Education	87	
Engineering	106	
FHSS	91	
Humanities	40	
Life	360	
Nursing	74	

Exp. Counts

	Apple	Andr
Business	444.4	4
CFAC	99.0	1
CMS	109.8	1
Education	90.9	1
Engineering	120.6	1
FHSS	90.9	1
Humanities	41.4	
Life	350.9	3
Nursing	69.3	

Nuances of χ^2 Tests

1. What do we do if our expected counts aren't all > 5?

- Go get more data, combined small count categories or ask a statistician.

Key Terminology

- Conditional distributions
- Marginal Distributions
- Side-by-side bar charts
- Chi-square test
- Chi-square statistics