Relationships between categorical variables

EDUC 641: Unit 2

David D. Liebowitz

UNIVERSITY OF OREGON

Roadmap

Research is a partnership of questions and data		What types of data are collected?	
		Categorical data	**Continuous data**
What kinds of questions can be asked of those data?	**Descriptive questions**	• How many members of class have black hair? • What proportion of the class attends full-time?	• How tall are class members, on average • How many hours per week do class members report studying, on average?
	Relational questions	• Are male-identifying students more likely to study part-time? • Are PrevSci PhD students more likely to be female-identifying?	• Do people who say they study for more hours also think they'll finish their doctorate earlier? • Are computer-literate students less anxious about statistics?

Goals of the unit

- Describe relationships between quantitative data that are categorical
- Calculate an index of the strength of the relationship between two categorical variables, the chi-squared (χ^2) statistic
- Write R scripts to conduct these analyses
- Formulate and describe the purpose of a null hypothesis
- Conceptually describe the criteria to make a statistical inference from a sample to a population
- Interpret and report the results of a contingency-table analysis and a statistical inference from a chi-squared statistic

Reminder of motivating question

Were convicted murderers more likely to be sentenced to death in Georgia if they killed someone Black or if they killed someone white?

Materials

1. Death penalty data (in file called deathpenalty.csv)
2. Codebook describing the contents of said data
3. R script to conduct the data analytic tasks of the unit

What we've done

Up until now, we've been examining each variable by itself...

Relationships between variables

Two-way tables

Now we seek to create a *joint display* of the values of *RVICTIM* and *DEATHPEN*

```
table(df$deathpen, df$rvictim)
```

```
#> 
#>        Black White
#>   No    1483   863
#>   Yes     23   106
```

Could do this other ways...

```
xtabs(formula = ~ deathpen + rvictim, data = df)
```

```
#>         rvictim
#> deathpen Black White
#>      No   1483   863
#>      Yes    23   106
```

Would be helpful to have these in percentage terms. Proportion of convicted murderers sentenced to death in case of Black victim:

$$\frac{23}{1483 + 23} * 100 = 1.53\%$$

Two-way tables

Can ask R for this too:

```
round(prop.table(table(df$deathpen, df$rvictim), margin=2)*100, 2)
```

```
#>
#>        Black White
#>    No  98.47 89.06
#>    Yes  1.53 10.94
```

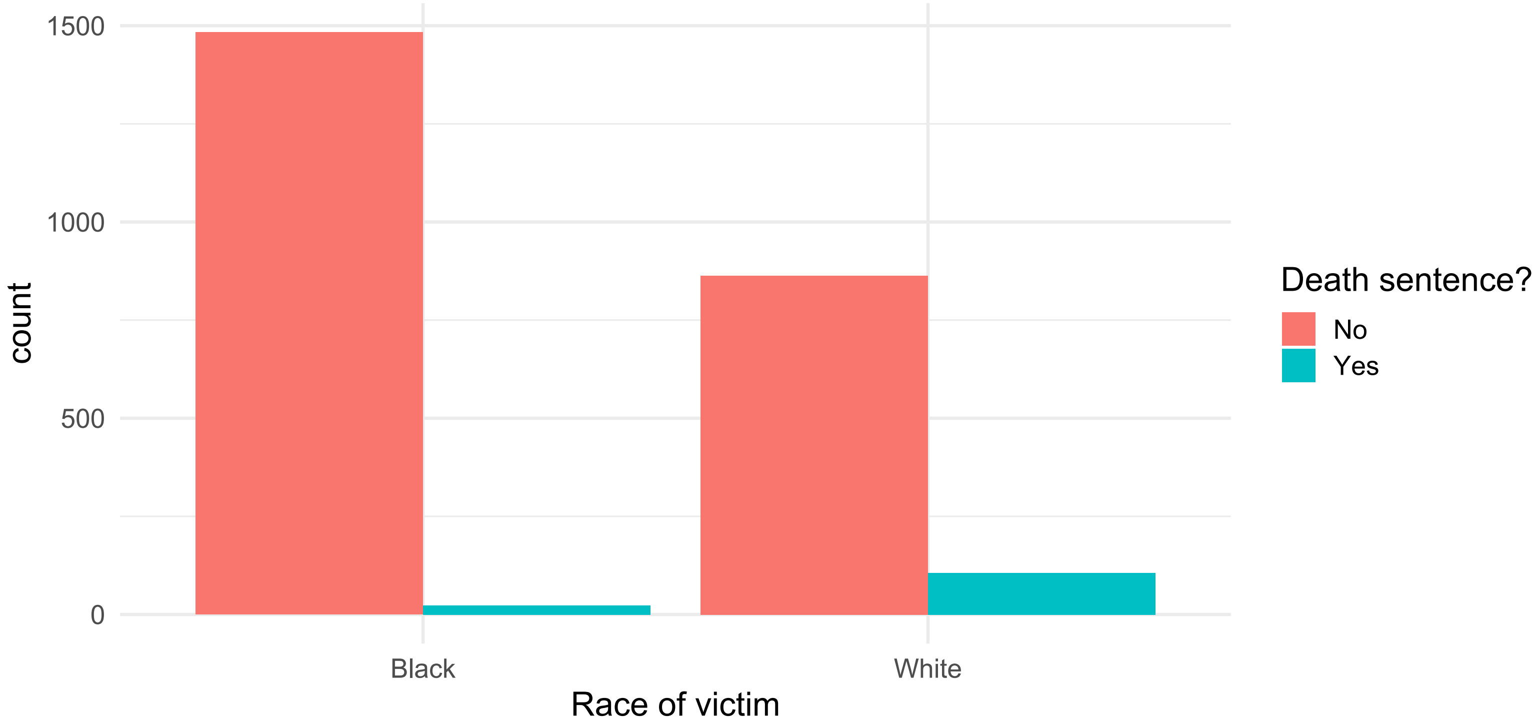
```
# the margin option 2 asks for the proportion of the columns
# if you want the proportion by rows, specify margin=1
```

Putting it into words

> In our sample of convicted murderers in Georgia, when a **Black person** was a victim...
>
> In our sample of convicted murderers in Georgia, when a **white person** was a victim...

Grouped charts

We can visualize these counts:



What is "related"?

To answer whether *DEATHPEN* and *RVICTIM* are related in our observed sample...

it might be helpful to imagine what the proportion of defendants sentenced to death would look like **if there were NO relationship**

Frequencies that we OBSERVE

```
#> 
#>        Black White  Sum
#>    No   1483   863 2346
#>    Yes    23   106  129
#>    Sum  1506   969 2475
```

Frequencies that we would EXPECT if there were NO relationship

deathpen	Black	White	Sum
No			2346
Yes			129
Sum	1506	969	2475

Observed vs. Expected

Frequencies that we OBSERVE

```
#>
#>         Black White  Sum
#>    No    1483   863 2346
#>    Yes     23   106  129
#>    Sum   1506   969 2475
```

Frequencies that we would EXPECT if there were NO relationship

deathpen	Black	White	Sum	Proport.
No			2346	0.948
Yes			129	0.052
Sum	1506	969	2475	1.000
Proportion	0.608	0.392	1.000	

Note: The proportions above are rounded, so if you use them to calculate the **EXPECTED** values, they will differ slightly from those on the next slide. If you calculate the proportions by hand (i.e., $1506/2474 = 0.60\overline{84}$), you will get the exact values, and then they will align with the rounded **EXPECTED** values on the next slide.

Observed vs. Expected

Frequencies that we OBSERVE

```
#> 
#>        Black White  Sum
#>   No    1483   863 2346
#>   Yes     23   106  129
#>   Sum   1506   969 2475
```

Frequencies that we would EXPECT if there were NO relationship

deathpen	Black	White	Sum
No	1428	918	2346
Yes	78	51	129
Sum	1506	969	2475

> What do you think? Is there a relationship between *DEATHPEN* and *RVICTIM*?

A desired index...?

Frequencies that we OBSERVE

```
#>
#>          Black White  Sum
#>     No    1483   863 2346
#>     Yes     23   106  129
#>     Sum   1506   969 2475
```

Frequencies that we would EXPECT if there were NO relationship

deathpen	Black	White	Sum
No	1428	918	2346
Yes	78	51	129
Sum	1506	969	2475

It would be nice to have an index of the **NET DISCREPANCY** between the **OBSERVED** and **EXPECTED** frequencies in the sample

The Chi-Squared χ^2 statistic

For a moment, assume that there is a powerful statistic that allows us to summarize the **NET DISCREPANCY** between the tables of **OBSERVED** and **EXPECTED** frequencies. Let's call this statistic the Pearson Chi-Squared (χ^2) statistic

Frequencies that we OBSERVE

```
#>
#>         Black White  Sum
#>     No   1483   863 2346
#>     Yes    23   106  129
#>     Sum  1506   969 2475
```

Frequencies that we would EXPECT if NO relationship

deathpen	Black	White	Sum
No	1428	918	2346
Yes	78	51	129
Sum	1506	969	2475

$$\chi^2 = \frac{(1483 - 1428)^2}{1428} + \frac{(863 - 918)^2}{918} + \frac{(23 - 78)^2}{78} + \frac{(106 - 51)^2}{51}$$

$$\chi^2 = 103.8$$

Yay! We got an answer, but what does it mean...?

Hypothesis testing and statistical inference

Big or small?

We can summarize the **NET DISCREPANCY** between the tables of **OBSERVED** and **EXPECTED** frequencies, using a statistic called the Pearson Chi-Squared (χ^2) statistic

Frequencies that we OBSERVE

```
#> 
#>        Black White  Sum
#>    No   1483   863 2346
#>    Yes    23   106  129
#>    Sum  1506   969 2475
```

Frequencies that we would EXPECT if NO relationship

deathpen	Black	White	Sum
No	1428	918	2346
Yes	78	51	129
Sum	1506	969	2475

$$\chi^2 = 103.8$$

Decision rule: If χ^2 is big, then declare that there is a relationship between *DEATHPEN* and *RVICTIM*; if χ^2 is zero (or close), then declare there is no relationship between *DEATHPEN* and *RVICTIM*... but what is **BIG**, what is **close to zero**, and is 103.8 **big** or **close to zero**? For that we will use this statistic to conduct a χ^2 goodness-of-fit test.

Statistical inference

Let's take a step back to capture the nature of the problem

- We've looked at some data on some convicted murderers in the state of Georgia
- We're not interested in only *these* murderers, but we're interested in a broader *population* of murderers from which our *sample* was drawn
  - In fact, even if we could observe outcomes for all murderers in the state of Georgia, our observation of them is imperfect due to *measurement error* and so we only ever observe samples, never populations (more on this later)
- Is there something about sampling from a population that could resolve our problem?
- Is there some way to generalize our conclusions about our *sample* relationship between *DEATHPEN* and *RVICTIM* to the *underlying population*?
  - This is called statistical inference and it is ***the*** critical contribution of quantitative methods to research

Sampling idiosyncrasy

When you generalize from a sample back to its underlying population, you must be careful that your empirical study has not been the victim of *sampling idiosyncrasy*

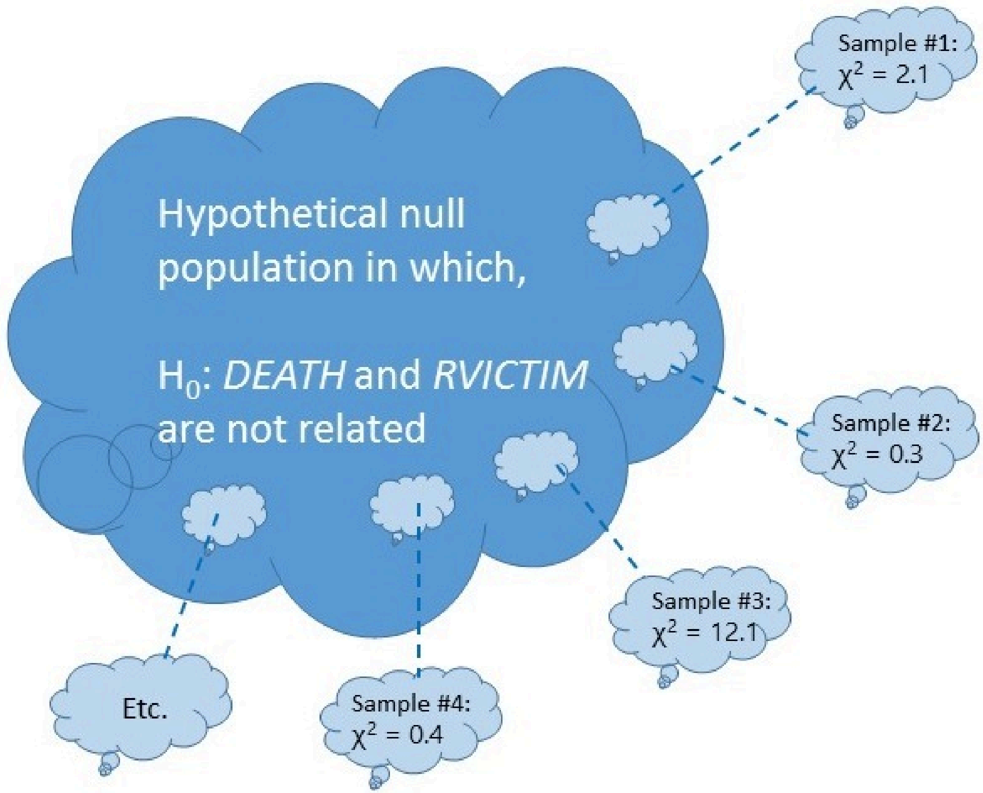
Is the following scenario plausible?

- There really is no relationship between *DEATHPEN* and *RVICTIM* **in the population**
- By accident, we have drawn an idiosyncratic sample from the population
- This sampling idiosyncrasy ended up giving us a χ^2 statistic as large as 103.8 by pure accident

How can we assess the plausibility of this scenario?

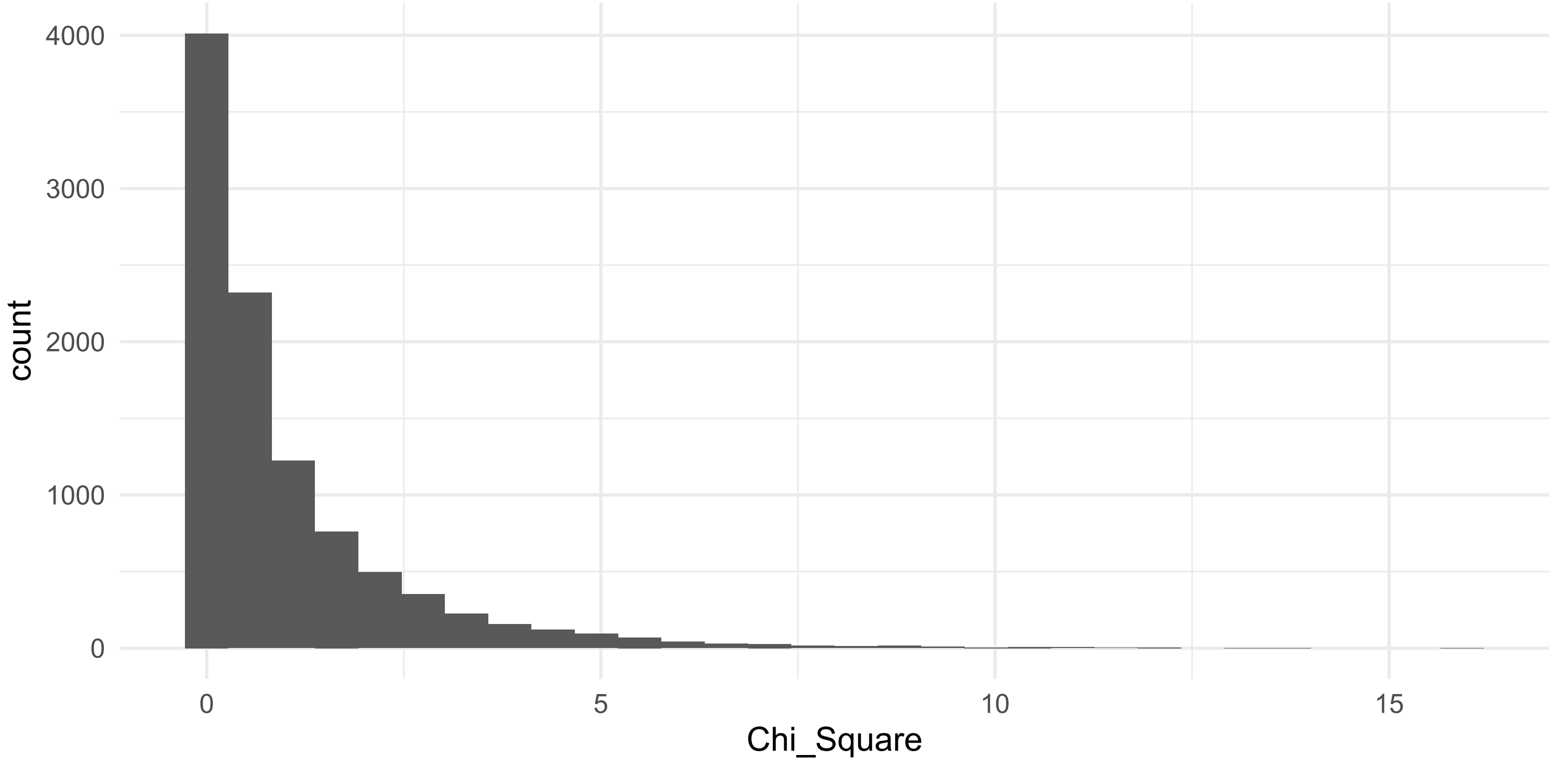
The Null Hypothesis (H_0)

We start by imagining a hypothetical world in which there is **no relationship** between *DEATHPEN* and *RVICTIM* in a true population of convicted murderers. Then, we imagine drawing a series of samples of convicted murders over and over again (say…10,000 times) from this hypothetical population. What values of the χ^2 statistic might we observe?



Testing the null

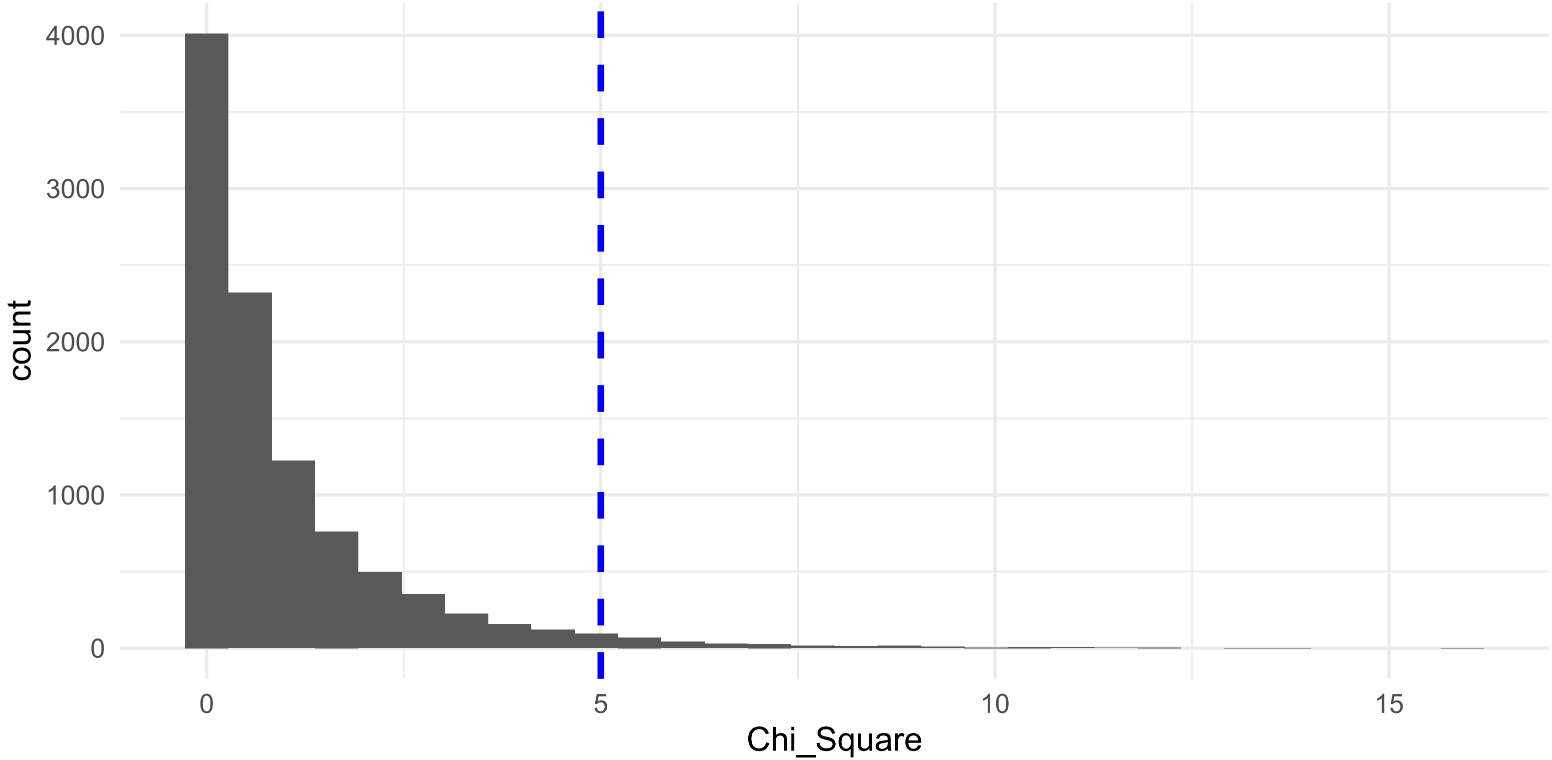
In this hypothetical example of repeated sampling from a null population, we could record all 10,000 values of the χ^2 statistic



The histogram summarizes the natural variation that could occur in a χ^2 statistic due to ***random sampling idiosyncrasy***, after drawing repeated samples from a hypothetical population in which there is no relationship between *DEATHPEN* and *RVICTIM*.

Testing the null

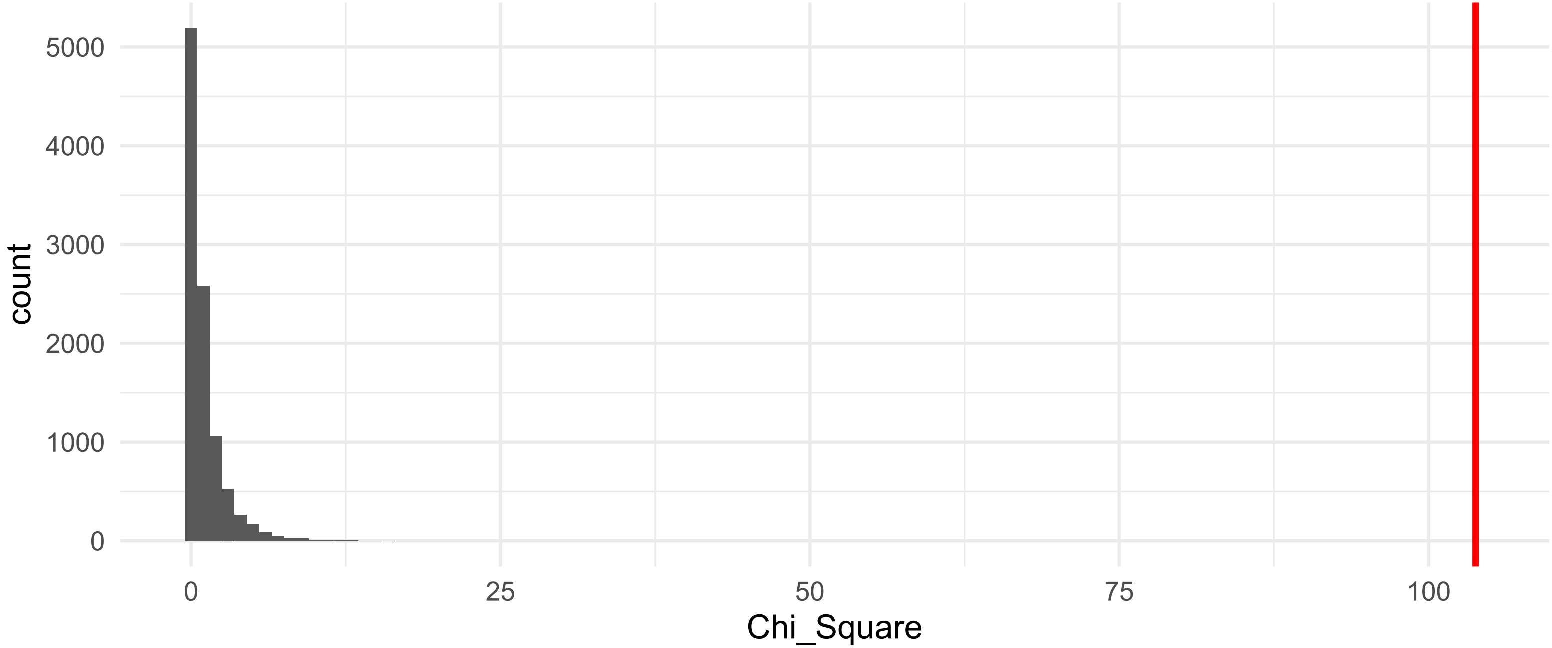
In this hypothetical example of repeated sampling from a null population, we could record all 10,000 values of the χ^2 statistic



If this were the histogram that could result from sampling idiosyncrasy, and this were the value of the chi-square statistic, what would you think?

Testing the null

In fact, this is the value of the χ^2 statistic we observed:

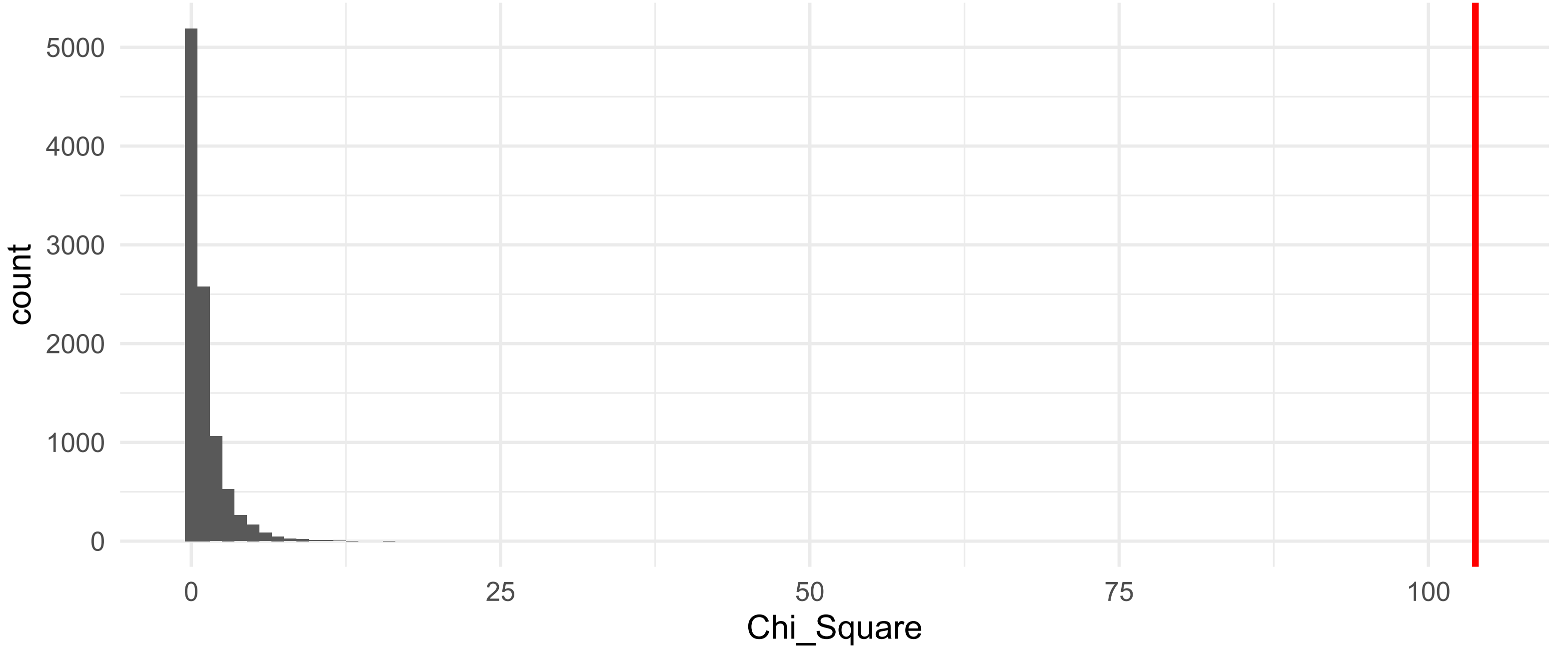


This is a histogram of possible chi-square values that could result from sampling idiosyncrasy, and the actual value of the chi-squared statistic in our sample. What do you think?

WOOHOO! In this thought exercise, you've just engaged in a rudimentary version of **Null-Hypothesis Significance Testing (NHST)**; the bedrock of most social science research.

Testing the null (*p*-values)

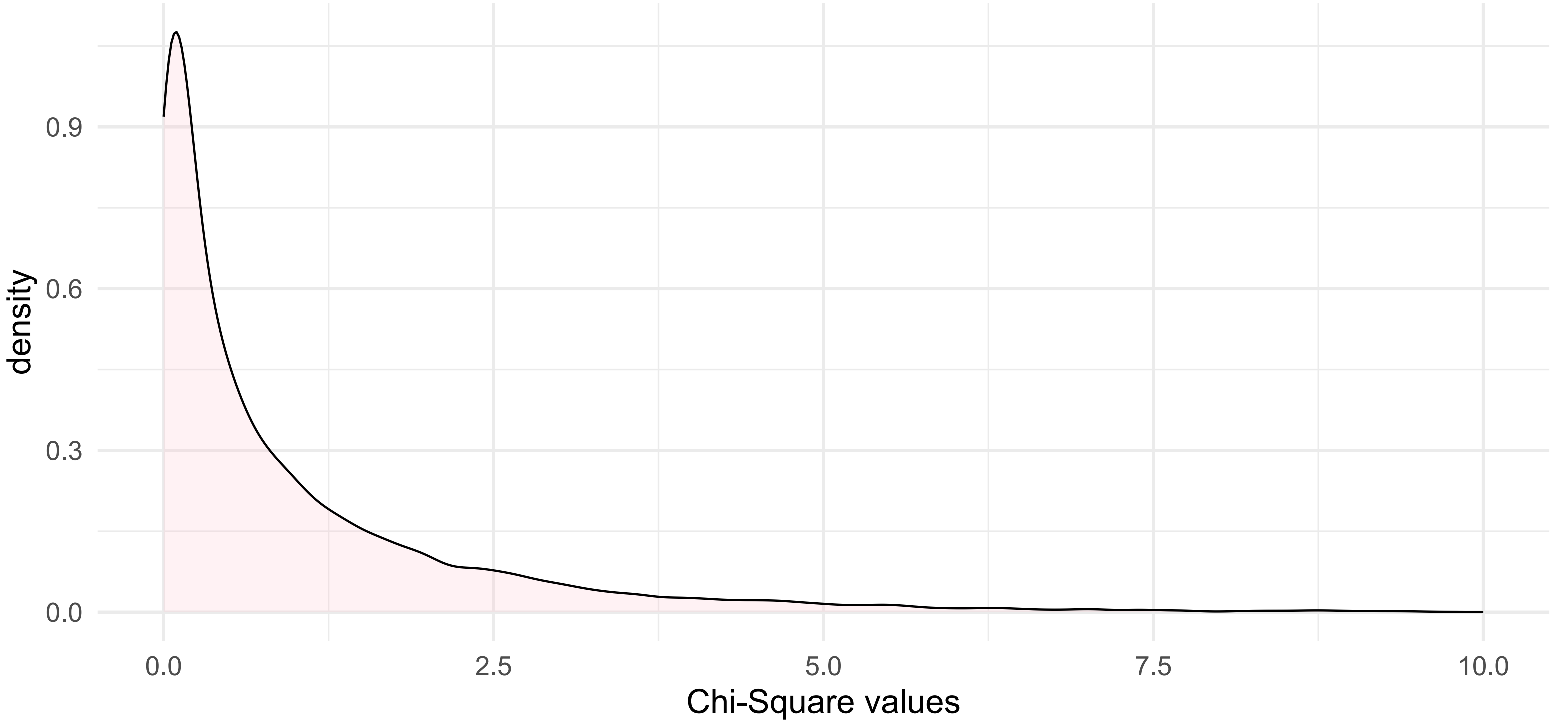
In fact, we don't need to examine the full histogram. Instead, we can say that in a hypothetical exercise of sampling repeatedly from a null population, less than 1 in a 1,000,000,000,000,000 (trillion) of all accidental values of the χ^2 statistic are larger than a value of 103.8



The statistic that captures the probability of observing a χ^2 statistic of a magnitude in a particular sample, in the presence of a null population, is called the ***p*-value**.

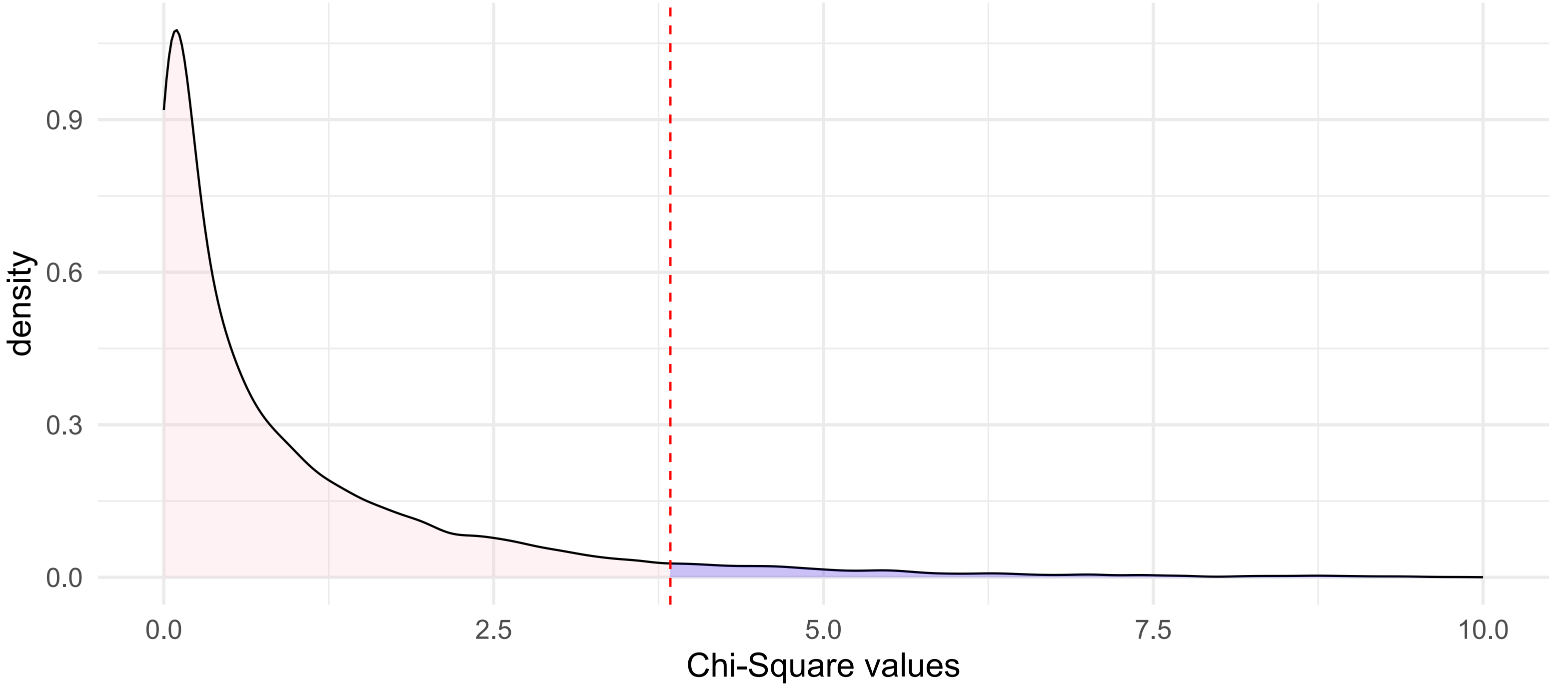
Testing the null (p-values)

At what p-value would you start to believe that the value of the χ^2 statistic in your own research was "big" (i.e., was unlikely to have occurred by accident)



Testing the null (p–values)

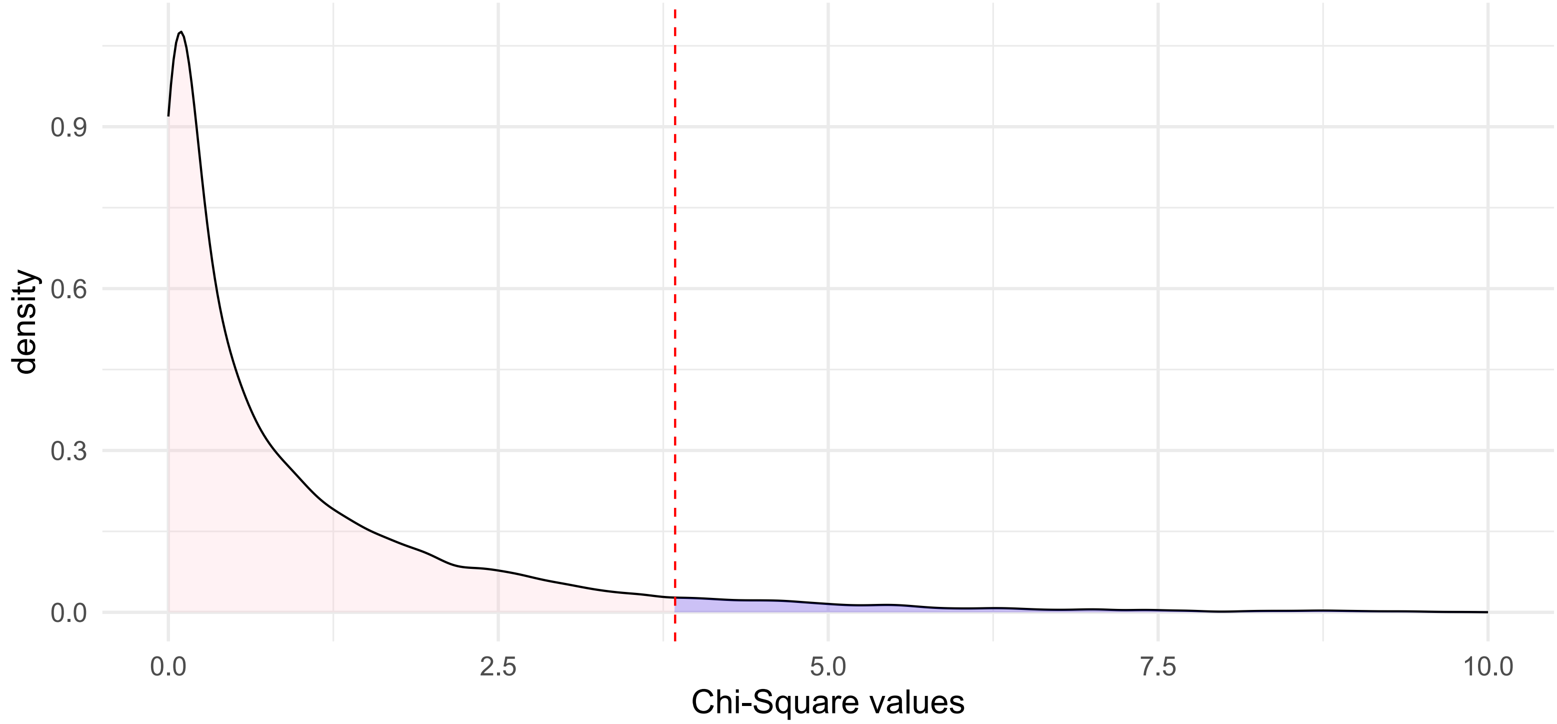
At what p-value would you start to believe that the value of the χ^2 statistic in your own research was "big" (i.e., was unlikely to have occurred by accident)



In social science research, it is customary to (arbitrarily) set that threshold at **5 percent (p<0.05)**. In other words, we say that if the difference between our observed data and our expected data would have happened in fewer than 1 out of 20 randomly drawn samples, that the difference reflects a true difference in the population.

Testing the null (p-values)

In social science research, it is customary to (arbitrarily) set an alpha-threshold and conduct a Null-Hypothesis Significance Test

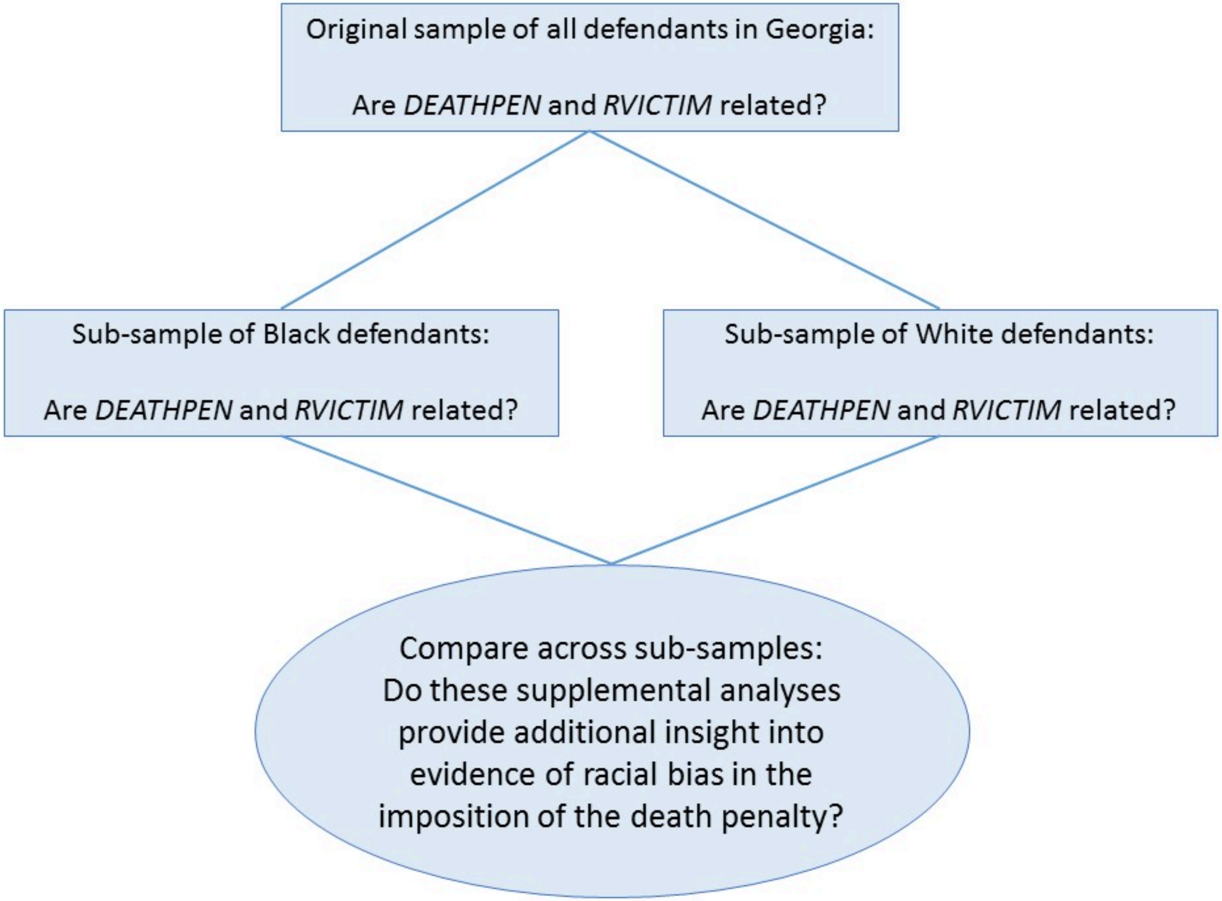


Is this the right thing to do? At the end of the course, we will revisit this concept.

Incorporating a third variable

Sub-sample comparisons

There are more complex ways of doing this, but one approach is to replicate the original contingency table analysis in interesting "slices" of the sample, defined by a third variable.



Cases with Black murderers

```
#>
#>           Black White
#>     No    1304    63
#>     Yes     18    50
```

Black murderers, Black victims

When a Black victim is killed by a Black murderer,

$$\frac{18}{18 + 1304} = 1.36\%$$

of the murderers are sentenced to death.

Black murderers, White victims

When a White victim is killed by a Black murderer,

$$\frac{50}{50 + 63} = 44.25\%$$

of the murderers are sentenced to death.

The percentage of Black murderers sentenced to death for killing a white victim is about 32.5 times the percentage of Black murderers sentenced to death for killing a Black victim, in Georgia.

I've subset my data to only cases with Black defendants. See the accompanying R script for how to do this.

A statistical test

Observed:

```
#>                df_b$rvictim
#> df_b$deathpen Black White
#>            No  1304    63
#>           Yes    18    50
```

Expected:

```
#>                df_b$rvictim
#> df_b$deathpen Black White
#>            No  1259   108
#>           Yes    63     5
```

χ^2 statistic:

```
#> X-squared
#>  414.7031
```

p-value

```
#> [1] 3.470593e-92
```

- H_0: *DEATHPEN* and *RVICTIM* are unrelated in the population of convicted Black murderers in GA
- χ^2 statistic: 414.7
- *p*-value: <0.0001
- Decision: Reject H_0
- Conclusion: There is a statistically significant relationship between the assignment of the death penalty and the race of the victim, on average, in the population of Black murderers in GA.

Cases with White murderers

```
#>
#>          Black White
#>     No     179   800
#>     Yes      5    56
```

White murderers, Black victims

When a Black victim is killed by a White murderer,

$$\frac{5}{5+179} = 2.71\%$$

of the murderers are sentenced to death.

White murderers, White victims

When a White victim is killed by a White murderer,

$$\frac{56}{56+800} = 6.89\%$$

of the murderers are sentenced to death.

The percentage of White murderers sentenced to death for killing a White victim is about 2.5 times the percentage of White murderers sentenced to death for killing a Black victim, in Georgia.

A statistical test

Observed:

```
#>                df_w$rvictim
#> df_w$deathpen Black White
#>            No   179   800
#>            Yes    5    56
```

Expected:

```
#>                df_w$rvictim
#> df_w$deathpen Black White
#>            No   173   806
#>            Yes   11    50
```

χ^2 statistic:

```
#> X-squared
#>  3.349547
```

p-value

```
#> [1] 0.06722351
```

- H_0: *DEATHPEN* and *RVICTIM* are unrelated in the population of convicted White murderers in GA
- χ^2 statistic: 3.35
- *p*-value: 0.067
- Decision: Fail to reject H_0
- Conclusion: There is not a statistically significant relationship between the assignment of the death penalty and the race of the victim, on average, in the population of white murderers in GA.

- Note that we **NEVER** accept the null-hypothesis. We only ever *fail to reject* it.

Putting it all together

Basic steps of classical statistical inference

1. State a research question, including a null hypothesis (H_0) which states there exists no relationship between our variables of interest, ***on average in the population***
2. Display and describe the observed data
3. Summarize the observed data in relationship to an expected value
4. Set a threshold at which we will no longer believe that the discrepancy between the observed and expected relationship is due to sampling idiosyncrasy
5. Estimate the p-value
6. Reject or fail to reject the null hypothesis
7. Interpret your findings drawing explicitly on plots, summary statistics and test statistics

Our intepretation

> In the population of convicted murderers in Georgia, the imposition of the death penalty and the race of the victim are, on average, related (χ^2 = 103.8, *p*< 0.001). The percentage of convicted murderers who were sentenced to death after killing a White victim was more than 8 times the percentage of convicted murderers who were sentenced to death after killing a Black victim. In Figure 1, we show…
>
> This phenomenon is largely driven by the imposition of the death penalty on Black defendants. Courts sentenced Black defendants to death for killing white victims at more than 32 times the frequency than when they were convicted of killing Black defendants (χ^2 = 414.7, *p*<0.001); whereas, we detect no statistical difference on average between white defendants convicted of murdering white -- compared to Black -- victims (χ^2 = 3.3, *p*=0.067). In Table 1, we show…

Estimating the χ^2 statistic in R

```
chi_df <- chisq.test(df$deathpen, df$rvictim)
chi_df
```

```
#>
#>     Pearson's Chi-squared test with Yates' continuity correction
#>
#> data:  df$deathpen and df$rvictim
#> X-squared = 103.82, df = 1, p-value < 2.2e-16
```

```
chi_df$expected
```

```
#>             df$rvictim
#> df$deathpen      Black     White
#>         No  1427.50545 918.49455
#>         Yes   78.49455  50.50545
```

```
round(chi_df$p.value, 5)
```

```
#> [1] 0
```

Synthesis and wrap-up

Goals of the unit

- Describe relationships between quantitative data that are categorical
- Calculate an index of the strength of the relationship between two categorical variables, the chi-squared (χ^2) statistic
- Write R scripts to conduct these analyses
- Formulate and describe the purpose of a null hypothesis
- Conceptually describe the criteria to make a statistical inference from a sample to a population
- Interpret and report the results of a contingency-table analysis and a statistical inference from a chi-squared statistic

To-Dos

Reading

- LSWR Chapter 11: hypothesis testing
- LSWR Chapter 12: categorical data analysis (chi-square test focus)
    - Please do not worry about fully understanding the discussions on sampling distributions, degrees of freedom, one- vs. two-sided tests, or variations of chi-squared calculations **(Sections 11.3, 11.4.3, 11.7, 11.8, 12.1.4-12.1.8, 12.3-12.9)**. We will (partially) cover these topics in future classes.
- Clayton (2020)
    - Last name A-L: Evans; Last Name M-Z: Clayton
    - Prep to summarize main ideas and key details

Optional follow-up

- Complete R Bootcamp Module 6 (matrices)
- Complete R Bootcamp Module 7 (lists)

Assignments

- Quiz on Units 1 & 2 on Oct. 17
- Assignment #2 Due October 25, 11:59pm