# Summarizing and displaying categorical data

EDUC 641: Unit 1

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?<br>• What proportion of the class attends full-time? | • How tall are class members, on average<br>• How many hours per week do class members report studying, on average? |
| | **Relational questions** | • Are male-identifying students more likely to study part-time?<br>• 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?<br>• Are computer-literate students less anxious about statistics? |

*This unit borrows heavily (with permission) from John B. Willett's course S-010Y*

# Class goals

1. Understand and implement principles of tabular (rectangular) data in R
2. Describe and summarize quantitative data that are categorical
3. Create visualizations of quantitative data that are categorical
4. Write R scripts to conduct these analyses

# The story behind the data

- Warren McCleskey sentenced to death for murdering police officer during armed robbery in Georgia (1978)
- McCleskey appeals to the Supreme Court, arguing his death sentence is due to racial bias in sentencing...his appeal is rejected
- Senator Edward Kennedy sponsors the Racial Justice Act to enforce "the Constitution's promise of equality under the law"
- McCleskey executed in Georgia (1991)

- Baldus, Pulaski & Woodworth (1983). A comparative review of death sentences: An empirical study of the Georgia experience. *J. Crim Law.*

**Our 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?**

*Note: Racial bias may exist in the criminal justice system irrespective of our answer to this question. Exploring such an issue would knit together more complete evidence from multiple research traditions.*

# Goals of the unit

We'll use a simplified version of the quantitative data accumulated by Baldus et al. (1983) to:

- Understand and implement principles of tabular data in R
- Describe and summarize quantitative data that are categorical
- Create visualizations of quantitative data that are categorical
- Write R scripts to conduct these analyses

# Materials

1. Death penalty data (in file called deathpenalty.csv)
2. Codebook describing the contents of said data (deathpenalty_codebook.pdf)
3. R script to conduct the data analytic tasks of the unit (EDUC641_3_code.R)

# Let's first access the data

Baldus, Pulaski and Woodworth (1983) compiled criminal justice and crude demographic information on all convicted murderers in the state of Georgia from in the mid-1970s. We have access to a simplified version of these data.

```
d <- read.csv('C:/Users/daviddl/Documents/EDUC/EDUC641_22F/data/deathpen
# but there are a lot of reasons to avoid using a hard-coded filepath
# so instead, we'll use our first R package

# Side note, I can write anything I want after a # sign and it
# doesn't cause any problems for your program. These are called
# comments and are critical to remind you (and tell others) what
# you are doing!
```

# Let's first access the data

Baldus, Pulaski and Woodworth (1983) compiled criminal justice and crude demographic information on all convicted murderers in the state of Georgia from in the mid-1970s. We have access to a simplified version of these data.

```
# The first thing we'll do is to install the package
# install.packages("here")  <- when you do this, remove the #
# don't forget the quotes around the package name!
# Then, we'll load the package:

library(here)
```

While you're at it, do the same for `tidyverse`. The `tidyverse` is a collection of helpful packages for manipulating and visualizing data. In particular, it includes the extremely useful `dplyr` and `ggplot2` !

# Let's first access the data

Baldus, Pulaski and Woodworth (1983) compiled criminal justice and crude demographic information on all convicted murderers in the state of Georgia from in the mid-1970s. We have access to a simplified version of these data.

```
# Ok, now we're ready to get started:

df <- read.csv(here("data/deathpenalty.csv"))
# it is common in R to name our datasets short names
# so as to reduce typing (df for dataframe)
```

but where did our data go?

# Understanding data structure

```
df
```

```
#>         id deathpen rdefend rvictim
#> 1        1        0       2       2
#> 2        2        0       2       1
#> 3        3        1       2       2
#> 4        4        0       2       2
#> 5        5        0       1       1
#> 6        6        0       2       1
#> 7        7        0       1       1
#> 8        8        0       2       2
#> 9        9        0       2       2
#> 10      10        0       2       1
#> 11      11        0       1       1
#> 12      12        0       1       1
#> 13      13        0       2       2
#> 14      14        0       1       1
#> 15      15        0       2       2
#> 16      16        0       2       2
#> 17      17        0       1       1
#> 18      18        0       1       2
#> 19      19        0       1       1
#> 20      20        0       2       2
```

# Understanding data structure

```
names(df)
```

```
#> [1] "id"       "deathpen" "rdefend"  "rvictim"
```

```
head(df)
```

```
#>   id deathpen rdefend rvictim
#> 1  1        0       2       2
#> 2  2        0       2       1
#> 3  3        1       2       2
#> 4  4        0       2       2
#> 5  5        0       1       1
#> 6  6        0       2       1
```

Some vocabulary:

- Observation: each row contains information on one person
- Variable: each column (R stores these as "vectors")
- Values: entries in the column

# Putting words to numbers

```
str(df)
```

```
#> 'data.frame':    2475 obs. of  4 variables:
#>  $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
#>  $ deathpen: int  0 0 1 0 0 0 0 0 0 0 ...
#>  $ rdefend : int  2 2 2 2 1 2 1 2 2 2 ...
#>  $ rvictim : int  2 1 2 2 1 1 1 2 2 1 ...
```

We know from our codebook what this long list of numbers means, so let's make our dataset a little more readable...

```
df$deathpen <- factor(df$deathpen,
                      levels = c(0,1), labels = c("No", "Yes"))
head(df)
```

```
#>   id deathpen rdefend rvictim
#> 1  1       No   White   White
#> 2  2       No   White   Black
#> 3  3      Yes   White   White
#> 4  4       No   White   White
#> 5  5       No   Black   Black
#> 6  6       No   White   Black
```

# Questions go with data

## What sorts of questions could you ask already of this data?

> Based on the contents of this data, what questions can you think of that ask you to describe single variables?
>
> Based on the contents of this data, what questions can you think of that ask you to inquire about relationships between two or more variables?

# Summarizing data: Tables

One powerful thing to do with data is to just count it up. How many defendants were sentenced to death?

```
table(df$deathpen)
```

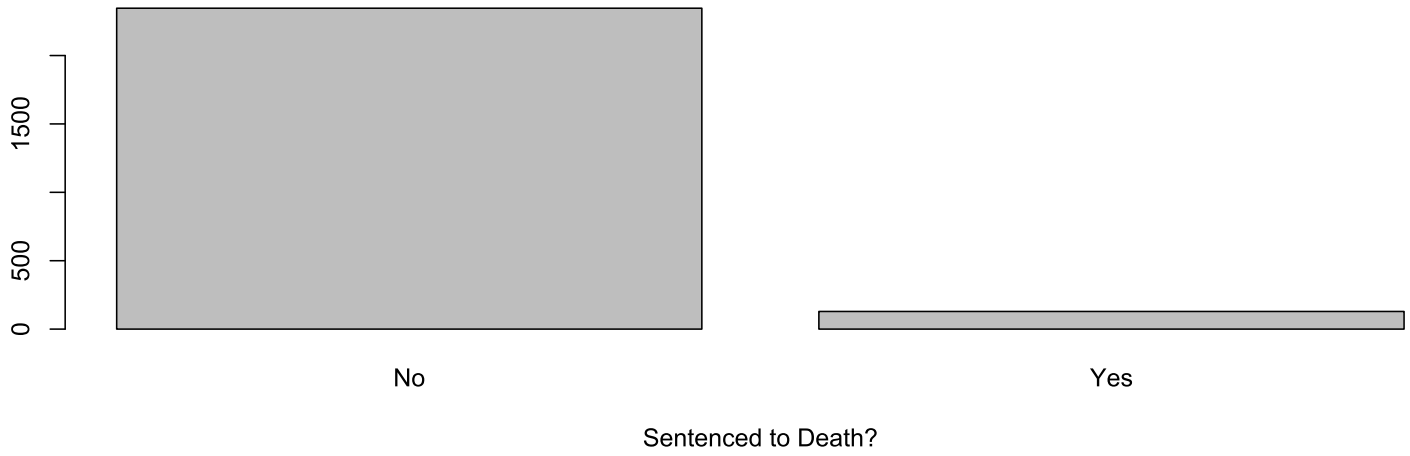
```
#>
#>   No  Yes
#> 2346  129
```

> From the *table* of the values of the variable *DEATHPEN*, we conclude ...

# Summarizing data: Charts

Even more powerful is to visualize these counts!

```
counts <- table(df$deathpen)
barplot(counts, xlab = "Sentenced to Death?")
```




> Sometimes, *proportions* are better ... can you calculate by hand the proportion of defendants who are sentenced to death?

# Summarizing data: Proportions

We can also ask R to do this for us

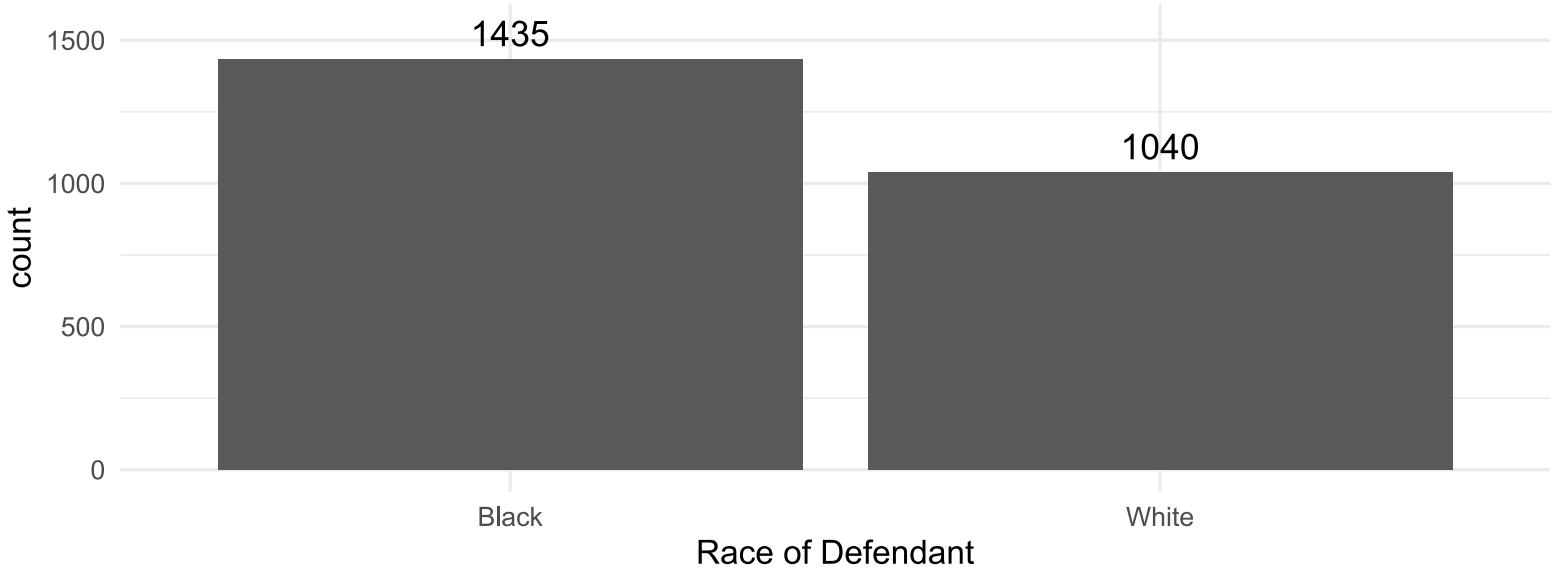
```
prop <- prop.table(table(df$rvictim))
barplot(prop, xlab = "Race of Victim",
        ylab = "Proportion of victims")
```


> After inspecting this *chart of the proportions* of the variable *RVICTIM* in our data, we conclude ...

# Another way to visualize

Let's start to get familiar with the beauty of `ggplot`

```
rd <- ggplot(df, aes(rdefend)) + geom_bar() +
        xlab("Race of Defendant")
```




> After inspecting this *chart* of the values of the variable *RDEFEND* in our data, we conclude ...

# BUT...

We still have not answered our 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?**

We will do so next class and start our journey into the wild world of **statistical inference**!

(*how's that for a cliffhanger?*)

# Synthesis and wrap-up

# Class goals

1. Understand and implement principles of tabular (rectangular) data in R
2. Describe and summarize quantitative data that are categorical
3. Create visualizations of quantitative data that are categorical
4. Write R scripts to conduct these analyses

# To-Dos

## Reading:

- LSWR Chapter 6: an overview of graphing in R; great to return to again and again
- Evans (2020)

## Optional follow-up

- Complete R Bootcamp Module 3 (installing and loading a package)
- Complete R Bootcamp Module 9 (importing data)
- Complete R Bootcamp Module 11 (creating plots)
- Complete R Bootcamp Module 12 (creating a project)

## Assignment:

- Assignment #1 Due October 14, 11:59pm

# Quiz