Data and Culture

Prof. Andrew Goldstone (andrew.goldstone@rutgers.edu)
Prof. Meredith McGill (meredith.mcgill@rutgers.edu)

October 31, 2022. Counting, Representing, Seeing.

lab "solution set" (1)–(3)

- The rise and fall of "Marilyn" is *not* because of Marilyn Monroe
- Social Security registrations do not capture the whole US population
  - extreme sex ratio in the SSA data pre-1930: highly biased for cross-gender comparison (perhaps for other things too)
  - know thy data
- How important are top names?

continue

- Complete parts (4) and (5) of lab 6.
  - If everyone in the group is ready to present about the discussion questions, you can explore the bonus part (6).

lab “solution set” (4)

- names given to both genders almost never maintain an even split
- if a name becomes popular (above about 0.2%) for one gender, declines in the other
- more common for a boys’ name to become predominantly for girls than vice versa
  - but remember the SSA problem!

lab “solution set” (5)

- social-cultural questions to investigate in the data?
- what further data?

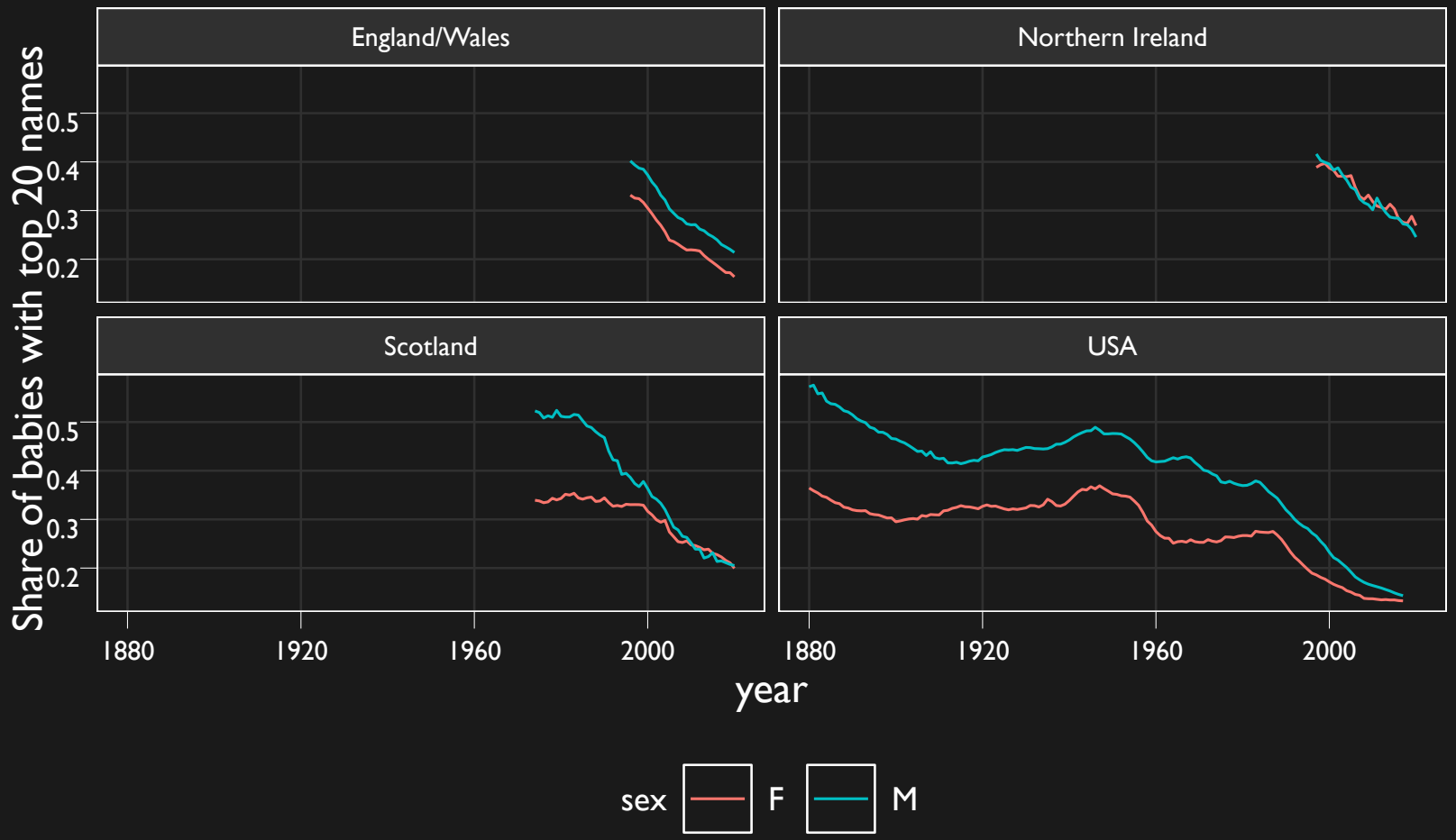
quick setup

```
library(dataculture)
library(babynames)
ukbabynames <- ukbabynames |>
    mutate(nation=fct_recode(nation,
        "England/Wales"="England & Wales"))
```

```
us20 <- babynames |> group_by(year, sex) |>
    mutate(prop=n / sum(n)) |>
    slice_max(prop, n=20) |>
    summarize(top20_prop=sum(prop)) |>
    mutate(nation="USA")

uk20 <- ukbabynames |> group_by(year, sex, nation) |>
    mutate(prop=n / sum(n)) |>
    slice_max(prop, n=20) |>
    summarize(top20_prop=sum(prop))
```

```
bind_rows(us20, uk20) |>
    ggplot(aes(x=year, y=top20_prop, color=sex)) +
        geom_line() + facet_wrap(vars(nation)) +
    labs(y="Share of babies with top 20 names")
```



representation: 3 problems in 1!

- as a fairness problem
  - Are women (racial minorities, the dominated class, etc.) justly represented among famous authors (musicians, scientists, politicians, etc.)?
- as an interpretive problem
  - In what ways does a given cultural phenomenon (novels published, baby names given, etc.) represent social facts?
- as a methodological problem
  - What choices of data (publication statistics, citation counts, plays on YouTube) will adequately represent the phenomena we want to study?

sourcing the data

> This study asks, How did the actual accomplishments of male and female writers specializing in various genres contribute differently to their fame or recognition? (Tuchman and Fortin, 73)

sourcing the data

This study asks, How did the actual accomplishments of male and female writers specializing in various genres contribute differently to their fame or recognition? (Tuchman and Fortin, 73)

- Look for a moment at the pages of the 1897 *DNB* on Mary Wollstonecraft Shelley linked from Canvas.
- How did these pages get made into data for Tuchman and Fortin for the computations in the tables? Find the specific steps described in the article.
  - *British Museum Catalogue* (*BMC*): now explore.bl.uk.

1. "We examined the *DNB* entries of all women and a random stratified sample of men born between 1750 and 1864 who…had published somewhere at least one 'not nonfiction work' " (81): who is in the universe?
2. "The birth dates…which we used… presume the transformation of the prestige of the novel in 1840…We chose to assume birth in 1814/1815 as the line between the two periods" (82): what kind of change is this looking for?
3. "We classified and counted by genre the entries credited to an individual in the *British Museum Catalogue*" (83): where do the categories come from?
4. "We used a measure of father's social status" (84): what kinds of explanatory factors are included?
5. "We developed a variable indicating whether an author's involvement in literature constituted professional participation" (84): what kinds of phenomena are being explained?

fame

"Fame is an index scored from 0 to 4." (83)

- What is fame? If you wanted to discuss the fame of a cultural producer today, what measurements would you use?

fame

"Fame is an index scored from 0 to 4." (83)

- What is fame? If you wanted to discuss the fame of a cultural producer today, what measurements would you use?
- How would you defend Tuchman and Fortin's choice? What interpretive issues might their choice create?

fame

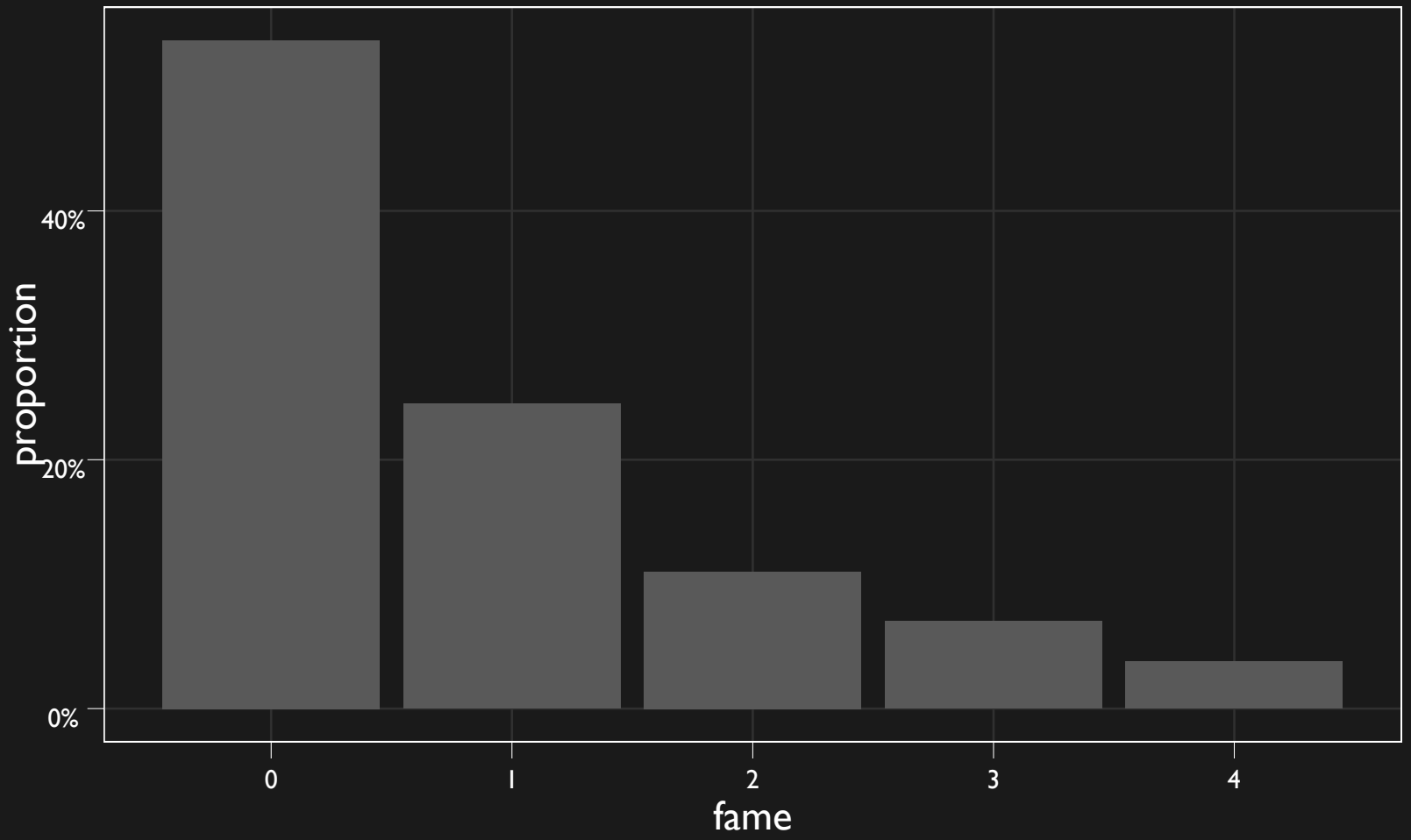
"Fame is an index scored from 0 to 4." (83)

- What is fame? If you wanted to discuss the fame of a cultural producer today, what measurements would you use?

- How would you defend Tuchman and Fortin's choice? What interpretive issues might their choice create?

"The peculiarities and limitations of the fame index result from the inherent limitations of the kinds of historical data in the *DNB*." (84n9)

“the sort of distribution one would expect”



Expos redux redux

- What are the key claims Tuchman and Fortin make on the basis of their quantitative evidence?

Expos redux redux

- What are the key claims Tuchman and Fortin make on the basis of their quantitative evidence?

Their [men's] ability to dominate fiction was largely accomplished through their strong numerical dominance of the more general field of nonfiction. (85)

With the rise of the English "man of letters," the mean fame of men increased while that of women decreased. (87)

Men and women achieve fame differently. (87)

The genre in which one writes contributes to whether one gains fame. (88)

Even as the number of novelists increased, especially among women, nonfiction dominated politically. (89)

What is a regression?

- a technique for figuring out how variation in one variable is related to (but not necessarily caused by) variations in other variables
- "regression of fame on specified variables": assume observed fame is some weighted sum of the specified variables plus some random noise; choose "best fit" weights b_i in

$$\begin{aligned}\text{Fame} = b_1(\text{Father's SES}) + b_2(\text{Fiction}) + b_3(\text{Poetry}) \\ + b_4(\text{Drama}) + b_5(\text{Nonfiction}) + b_0 + \epsilon\end{aligned}$$

- means: all else being equal, on average, a one unit increase in Fiction is associated with a b_2 increase in Fame
- does not mean: writing another fiction book makes you b_2 famouser

how to read a regression table

- What is the model?
- Does that model make any sense?
- What do the coefficients b mean?
- Is the model a good fit to the data?
- "Adjusted R^2": shrug
  - widely used to indicate goodness of fit
  - does not indicate goodness of fit
- ***: shrug
- $\beta_i = \hat{b}_i s_{x_i} / s_y$: shrug

upshot: Tuchman and Fortin's data sandwich

- data-making process required hard choices about simplification and reduction
  - and prior hypotheses about what kinds of factors matter (genre choice, class status, time)
- the data itself is much messier than prior conviction or forceful anecdote ("magnetizing of her brain by Shelley")
- the conclusion is focused on institutional factors, not on a broad atmosphere of sexism

representation nation

Reflection theory…simply states that cultural products such as literature in some way mirror the social order. (Griswold, 740)

Is the American novel unique, as it is often said to be? If so, do its peculiar properties reflect some American character or experience? (741)

Griswold's data

A random sample of all novels published in the United States between 1876 and 1910. The source of the sample was the *American National Catalogue*…A sample of 130 novels was divided into time periods. (746)

Griswold's data

A random sample of all novels published in the United States between 1876 and 1910. The source of the sample was the *American National Catalogue*…A sample of 130 novels was divided into time periods. (746)

Analysis focused on a number of variables pertaining to plot, author characteristics, and bibliographic information. (746)

what could be more American than © ?

So in addition to reflecting imperatives of the genre, the novels reflected differential market positions brought about by the state of American copyright laws. (760)

next

- read Manovich, “What Is Visualization?” (Canvas)
- go through *R for Data Science*, chapter 3
  - the example data is about cars (shrug)
  - mostly look at the pictures
  - copy and paste the code examples if you want
  - skip exercises unless you’re having fun