# DS-UA 112
# Introduction to Data Science

Lecture 4

# Agenda

- Review
- Lesson
- Demo

# Reminders

- Announcement
  - Section
  - Office Hours
- Homework
  - Upload to Gradescope
- Lecture
  - Lessons and Demos Links
  - Forums

# Review

## State-specific data on the relative frequency of given names in the population of U.S. births where the individual has a Social Security Number

***(Tabulated based on Social Security records as of March 3, 2019)***

For each of the 50 states and the District of Columbia we created a file called SC.txt, where SC is the state's postal code.

Each record in a file has the format: 2-digit state code, sex (M = male or F = female), 4-digit year of birth (starting with 1910), the 2-15 character name, and the number of occurrences of the name. Fields are delimited with a comma. Each file is sorted first on sex, then year of birth, and then on number of occurrences in descending order. When there is a tie on the number of occurrences names are listed in alphabetical order. This sorting makes it easy to determine a name's rank. The first record for each sex & year of birth has rank 1, the second record has rank 2, and so forth.

To safeguard privacy, we restrict our list of names to those with at least 5 occurrences. If a name has less than 5 occurrences for a year of birth in any state, the sum of the state counts for that year will be less than the national count.

# How to Switch the Order of Two Events?

$$
\begin{aligned}
\frac{P(n|y)P(y)}{P(n)} &= \frac{P(n \text{ and } y)}{P(y)} \frac{P(y)}{P(n)} \\
&= \frac{P(n \text{ and } y)}{P(n)} \\
&= P(y|n)
\end{aligned}
$$

# How to Switch the Order of Two Events?

$$P(y \mid n) = \frac{P(n \mid y)P(y)}{P(n)}$$

# How to Switch the Order of Two Events?

$$P(y|n) = \frac{P(n|y)P(y)}{P\left((n \text{ and } 1880) \text{ or } (n \text{ and } 1881)\ldots\right)}$$

$$= \frac{P(n|y)P(y)}{P\left(n \text{ and } 1880\right) + \ldots + P\left(n \text{ and } 1881\right)}$$

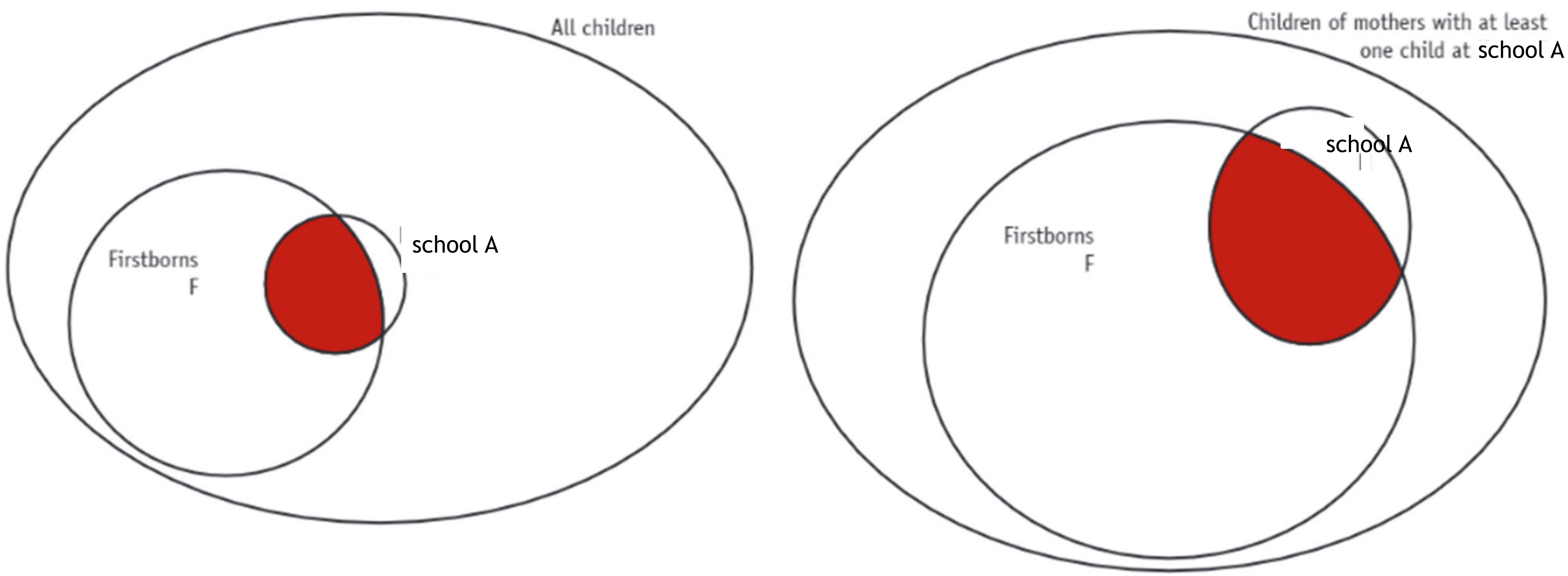
## How to Switch the Order of Two Events?

$$P(y \mid n) = \frac{P(n \mid y)P(y)}{\Sigma_y P(n \text{ and } y)}$$

# How to Switch the Order of Two Events?

$$P(y \mid n) = \frac{P(n \mid y)P(y)}{\Sigma_y P(n \mid y) \cdot P(y)}$$

# Lesson


All children
Firstborns
F
school A
Children of mothers with at least
one child at school A
school A
Firstborns
F

# Lesson

$$\frac{P\left(A|F\right)}{P\left(A|\text{not } F\right)} > 1$$

# Lesson

$$\frac{P\left(A|F\right)}{P\left(A|\text{not }F\right)} = \frac{P\left(F|A\right)P(A)}{P\left(F\right)} \cdot \left[\frac{P\left(\text{not }F|A\right)P(A)}{P\left(\text{not }F\right)}\right]^{-1}$$

# Lesson

$$\frac{P\left(A|F\right)}{P\left(A|\text{not }F\right)} = \frac{P\left(F|A\right)P(A)}{P\left(F\right)} \cdot \left[\frac{P\left(\text{not }F|A\right)P(A)}{P\left(\text{not }F\right)}\right]^{-1}$$
$$= \frac{P\left(F|A\right)}{P\left(\text{not }F|A\right)} \cdot \frac{P\left(\text{not }F\right)P(A)}{P\left(F\right)}$$

# Lesson

$$
\begin{aligned}
\frac{P(A|F)}{P(A|\text{not } F)} &= \frac{P(F|A)\,P(A)}{P(F)} \cdot \left[\frac{P(\text{not } F|A)\,P(A)}{P(\text{not } F)}\right]^{-1} \\
&= \frac{P(F|A)}{P(\text{not } F|A)} \cdot \frac{P(\text{not } F)\,P(A)}{P(F)} \\
&= \frac{P(F|A)}{1 - P(F|A)} \cdot \frac{1 - P(F)}{P(F)}
\end{aligned}
$$

# Lesson

$$
\begin{aligned}
\frac{P(A|F)}{P(A|\text{not } F)} &= \frac{P(F|A)\,P(A)}{P(F)} \cdot \left[\frac{P(\text{not } F|A)\,P(A)}{P(\text{not } F)}\right]^{-1} \\
&= \frac{P(F|A)}{P(\text{not } F|A)} \cdot \frac{P(\text{not } F)\,P(A)}{P(F)} \\
&= \frac{P(F|A)}{1 - P(F|A)} \cdot \frac{1 - P(F)}{P(F)} \\
&= \frac{P(F|A)}{1 - P(F|A)} \cdot \left(\frac{1}{P(F)} - 1\right)
\end{aligned}
$$

# Lesson

$$P(F) = \frac{N}{\lambda N}$$

# Lesson

$$
\begin{aligned}
\frac{P\left(A|F\right)}{P\left(A|\text{not } F\right)} &= \frac{P\left(F|A\right)P(A)}{P\left(F\right)} \cdot \left[\frac{P\left(\text{not } F|A\right)P(A)}{P\left(\text{not } F\right)}\right]^{-1} \\
&= \frac{P\left(F|A\right)}{P\left(\text{not } F|A\right)} \cdot \frac{P\left(\text{not } F\right)P(A)}{P\left(F\right)} \\
&= \frac{P\left(F|A\right)}{1-P\left(F|A\right)} \cdot \frac{1-P\left(F\right)}{P\left(F\right)} \\
&= \frac{P\left(F|A\right)}{1-P\left(F|A\right)} \cdot \left(\frac{1}{P\left(F\right)} - 1\right) \\
&= \frac{P\left(F|A\right)}{1-P\left(F|A\right)} \cdot \left(\lambda - 1\right)
\end{aligned}
$$

# Lesson

| | Landon (Rep) | Roosevelt (Dem) |
|---|---|---|
| Predicted | 57% | 43% |
| Actual | 38% | 62% |

# Lesson

| | Dewey (Rep) | Truman (Dem) |
|---|---|---|
| Predicted | 49.5% | 44.5% |
| Actual | 45.1% | 49.6% |

# Lesson

- Self-selected sample.
  - Sample is whoever chooses to answer.
- Convenience sample
  - Sample is whomever/whatever is convenient for investigator.
- Judgment sample
  - Sample is whomever/whatever investigator deliberately selects

# Lesson

- Probability sample
  - Sample is selected based on probabilistic procedure.
  - Assigns precise probability to the event that each particular sample is drawn from the population
  - This allows to quantify uncertainty/confidence about a prediction

# Lesson

- Probability sample
  - Simple Random Sample

# Lesson

- Probability sample
  - Simple Random Sample
  - Cluster Sample

# Lesson

- Probability sample
  - Simple Random Sample
  - Cluster Sample
  - Stratified Sample

# Demo

- Simulate votes for election
- Can large amounts of data correct for bias?