Soc500: Applied Social Statistics
Week 1: Introduction and Probability

Brandon Stewart[1]

Princeton

September 12, 2018

[1] These slides are heavily influenced by Matt Blackwell and Adam Glynn with contributions from Justin Grimmer and Matt Salganik. Illustrations by Shay O'Brien.

Where We've Been and Where We're Going...

- Last Week
  - methods camp
  - pre-grad school life
- This Week
  - Wednesday
    - welcome
    - basics of probability
- Next Week
  - random variables
  - joint distributions
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Questions?

Welcome and Introductions

- The tale of two classes: Soc400/Soc500 Applied Social Statistics
- I
  - ...am an Assistant Professor in Sociology.
  - ...am trained in political science and statistics
  - ...do research in methods and statistical text analysis
  - ...love doing collaborative research
  - ...talk very quickly
- Your Preceptors
  - sage guides of all things
  - Shay O'Brien (Soc500)
  - Alex Kindel (Soc400)
  - Ziyao Tian (Soc400)

Overview

- Goal: train you in statistical thinking
- First in a two course sequence $\rightsquigarrow$ replication project and longer arc
- Difficult course but with many resources to support you.
- When we are done you will be able to teach yourself many things
- Syllabus is a useful resource including philosophy of the class.

Specific Goals

- For the semester
  - critically read, interpret and replicate the quantitative content of many articles in the quantitative social sciences
  - conduct, interpret, and communicate results from analysis using multiple regression
  - explain the limitations of observational data for making causal claims
  - write clean, reusable, and reliable `R` code.
  - feel empowered working with data

Why R?

- It will give you super powers (but not at first)
- It is free and open source
- It is the *de facto* standard in many applied statistical fields

Why RMarkdown?

What you've done before

Image Credit: Baumer et al (2014)

Why RMarkdown?

RMarkdown

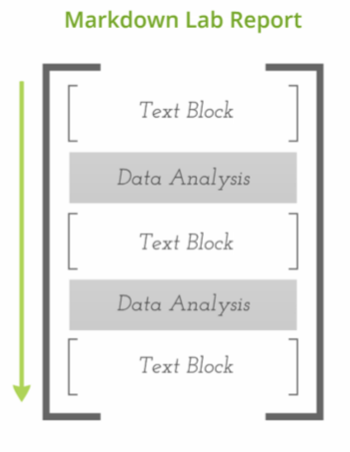


Image Credit: Baumer et al (2014)

Mathematical Prerequisites

- No formal pre-requisites
- Balancing rigor and intuition
  - no rigor for rigor's sake
  - we will tell you *why* you need the math, but also feel free to ask
  - course focus on how to reason about statistics, not just memorize guidelines
- We will teach you any math you need as we go along
- Crucially though- this class is not about innate statistical aptitude, it is about effort
- We all come from very different backgrounds. Please have patience with yourself and with others.

Ways to Learn

- Lecture
  learn broad topics
- Precept
  learn data analysis skills, get targeted help on assignments
- Readings
  support materials for lecture and precept

Reading

- Think of the lecture slides as primary reading.
- How to think about the reading?
- Key Books:
  - Angrist and Pischke (2008) *Mostly Harmless Econometrics*
  - Aronow and Miller. Forthcoming. *Foundations of Agnostic Statistics* (not yet available)
  - Blitzstein and Hwang. 2014. *Introduction to Probability* (available online through the library)
- Optional Books:
  - Fox (2016) *Applied Regression Analysis and Generalized Linear Models*
  - Imai (2017) *A First Course in Quantitative Social Science*
- Why so many books?

Ways to Learn

- Lecture
  learn broad topics
- Precept
  learn data analysis skills, get targeted help on assignments
- Readings
  support materials for lecture and precept
- Problem Sets
  reinforce understanding of material, practice

Problem Sets

- Schedule (available Wednesday, due 8 days later at precept)
- Grading and solutions
- Collaboration policy
- You may find these difficult. Start early and seek help!
- Most important part of the class

Ways to Learn

- Lecture
  learn broad topics
- Precept
  learn data analysis skills, get targeted help on assignments
- Readings
  support materials for lecture and precept
- Problem Sets
  reinforce understanding of material, practice
- Piazza
  ask questions of us and your classmates
- Office Hours
  ask even more questions.

Your Job: work hard and get help when you need it!

Daily Feedback

Can we go
over Bayes'
Rule again?

How to Get Help

1. Class and Precept
2. Daily Feedback
3. Readings and Slides
4. Piazza
5. Preceptor Office Hours
6. Instructor Office Hours
7. Final Exam Prep
8. External Consulting
9. Individual and Group Tutoring

Read the syllabus for more details.

Advice from Prior Generations

- Definitely take it! And be prepared to set aside a lot of time.
- Ask questions if you don't know what's going on!
- Study hard, work hard, review the slides.
- Investing a considerable amount of time in getting familiar with R and its various tools will pay off in the long run!
- Go over the lecture slides each week. This can be hard when you feel like you're treading water and just staying afloat, but I wish I had done this regularly.
- It's challenging but very doable and rewarding if you put the time in. There are plenty of resources to take advantage of for help.
- This course is very challenging but greatly contributed to my understanding of social statistics. If you're truly invested in the subject and willing to put in the work (more than you expect possibly), it will be one of the best courses you've taken.

Outline of Topics

Outline in reverse order:

- Causal Inference:
  assess the effect of a counterfactual intervention using observed associations.
- Regression:
  measure the association (expectation of a variable given a number of others).
- Inference:
  learn about things we don't know from the things we do know
- Probability:
  learn what data we would expect if we did know the truth.

Probability $\rightarrow$ Inference $\rightarrow$ Regression $\rightarrow$ Causal Inference

Attribution and Thanks

- My philosophy on teaching: don't reinvent the wheel- customize, refine, improve.
- Huge thanks to those who have provided slides particularly: Matt Blackwell, Adam Glynn, Justin Grimmer, Jens Hainmueller, Erin Hartman, Kevin Quinn
- Also thanks to those who have discussed with me at length including Dalton Conley, Chad Hazlett, Gary King, Kosuke Imai, Matt Salganik and Teppei Yamamoto.
- Previous generations of preceptors have also been incredible important: Clark Bernier, Elisha Cohen, Ian Lundberg and Simone Zhang.
- Shay O'Brien produced the hand-drawn illustrations used throughout.

This Class

Any questions about this class?

Let's get started

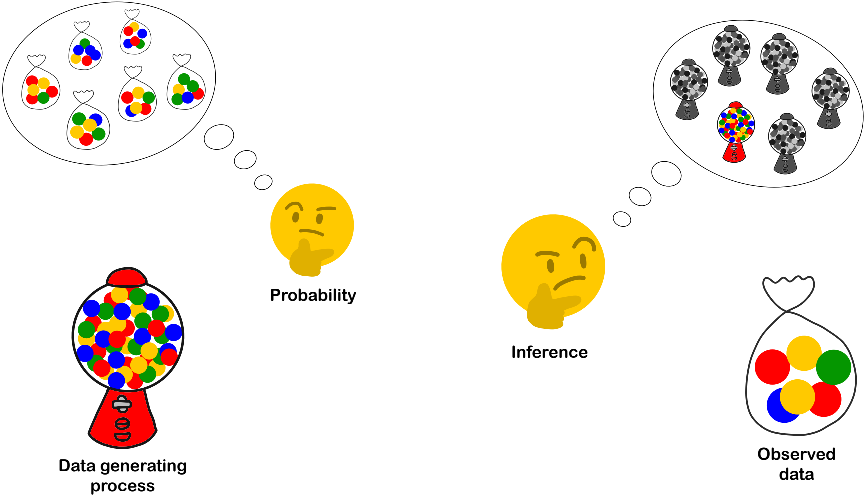
What is Statistics?

- Branch of mathematics studying collection and analysis of data
- The name statistic comes from the word state
- The arc of developments in statistics
  1) an applied scholar has a problem
  2) they solve the problem by inventing a specific method
  3) statisticians generalize and export the best of these methods
- Relatively recent field (started at the very end of the 19th century)
- Provides a way of making principled guesses based on stated assumptions.
- In practice, an essential part of research, policy making, political campaigns, selling people things...

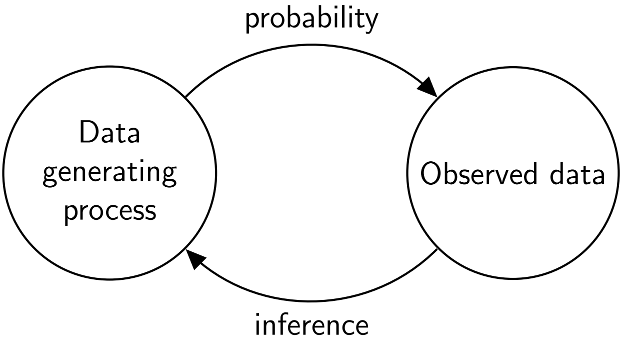
Why study probability?

It enables inference

In Picture Form



In Picture Form



Statistical Thought Experiments

- Start with probability
- Allows us to contemplate world under hypothetical scenarios
  - hypotheticals let us ask- is the observed relationship happening by chance or is it systematic?
  - it tells us what the world would look like under a certain assumption
- We will review probability today, but feel free to ask questions as needed throughout the semester.

Example: Fisher's Lady Tasting Tea

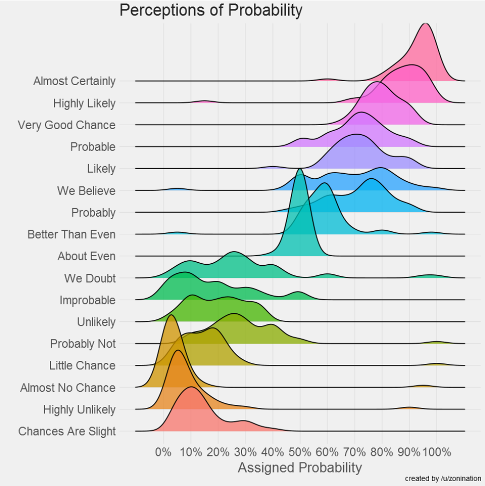
- The Story Setup
  (lady discerning about tea)
- The Experiment
  (perform a taste test)
- The Hypothetical
  (count possibilities)
- The Result
  (boom she was right)

Tea-Tasting Distribution

Success count	Permutations of selection	Number of permutations
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
Total		70

This became the Fisher Exact Test.

From 'Probably' to Probability



Can we make this more precise?

Why Probability?

- Helps us envision hypotheticals
- Describes uncertainty in how the data is generated
- Data Analysis: estimate probability that something will happen
- Thus: we need to know how probability gives rise to data

Intuitive Definition of Probability

While there are several interpretations of what probability is, most modern (post 1935 or so) researchers agree on an axiomatic definition of probability.

3 Axioms (Intuitive Version):

1. The probability of any particular event must be non-negative.
2. The probability of anything occurring among all possible events must be 1.
3. The probability of one of many mutually exclusive events happening is the sum of the individual probabilities.

All the rules of probability can be derived from these axioms.
(we will return to these in a minute)

Sample Spaces

To define probability we need to define the set of possible outcomes.

The sample space is the set of all possible outcomes, and is often written as **S** or Ω.

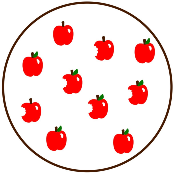
For example, if we flip a coin twice, there are four possible outcomes,

$$\mathbf{S} = \Big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\Big\}$$

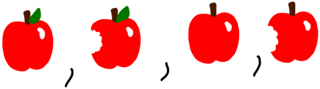
Thus the table in Lady Tasting Tea was defining the sample space. (Note we defined illogical guesses to be prob$= 0$)

A Running Visual Metaphor

Imagine that we sample an apple from a bag.
Looking in the bag we see:



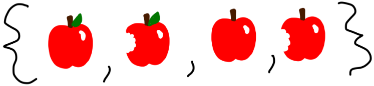
The sample space is:

$\Omega = \mathbf{S} = \{$  $\}$

Events

Events are subsets of the sample space.

For Example, if

$\Omega = \mathbf{S} =$ 

then

and

are both **events.**

Events Are a Kind of Set

Sets are collections of things, in this case collections of outcomes

One way to define an event is to describe the common property that all of the outcomes share. We write this as

$$\{\omega | \omega \text{ satisfies } P\},$$

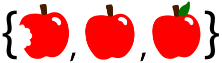
where P is the property that they all share.

If $A = \{\omega | \omega \text{ has a leaf}\}$:

 $\in A$,  $\in A$,  $\not\in A$,  $\not\in A$

Complement

A complement of event A is a set: A^c, is collection of all of the outcomes not in A. That is, it is "everything else" in the sample space.



and

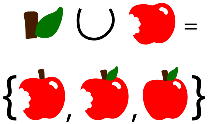
are **complements.**

$$A^c = \{\omega \in \Omega | \omega \notin A\}.$$

Important complement: $\Omega^c = \emptyset$, where $\emptyset$ is the empty set—it's just the event that nothing happens.

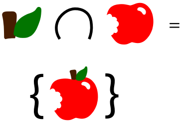
Unions and intersections (Operations on events)

The union of two events, A and B is the event that A or B occurs:



$$A \cup B = \{\omega | \omega \in A \text{ or } \omega \in B\}.$$

The intersection of two events, A and B is the event that both A and B occur:



$$A \cap B = \{\omega | \omega \in A \text{ and } \omega \in B\}.$$

Operations on Events

We say that two events A and B are disjoint or mutually exclusive if they don't share any elements or that $A \cap B = \emptyset$.

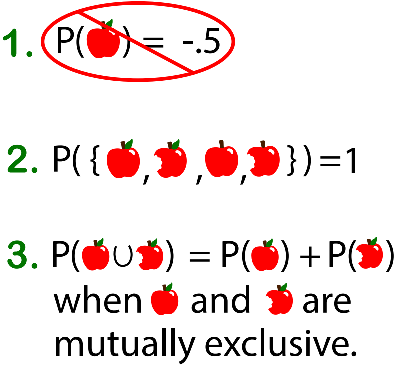
An event and its complement A and A^c are disjoint.

Sample spaces can have infinite events $A_1, A_2, \ldots$.

Probability Function

A probability function $P(\cdot)$ is a function defined over all subsets of a sample space $\mathbf{S}$ that satisfies the following three axioms:

1. $P(A) \geq 0$ for all A in the set of all events. nonnegativity
2. $P(\mathbf{S}) = 1$ normalization
3. if events $A_1, A_2, \ldots$ are mutually exclusive then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. additivity



All the rules of probability can be derived from these axioms.
(See Blitzstein & Hwang, Def 1.6.1.)

A Brief Word on Interpretation

Massive debate on interpretation:

- Subjective Interpretation
  - Example: The probability of drawing 5 red cards out of 10 drawn from a deck of cards is whatever you want it to be. But...
  - If you don't follow the axioms, a bookie can beat you
  - There is a correct way to update your beliefs with data.
- Frequency Interpretation
  - Probability is the relative frequency with which an event would occur if the process were repeated a large number of times under similar conditions.
  - Example: The probability of drawing 5 red cards out of 10 drawn from a deck of cards is the frequency with which this event occurs in repeated samples of 10 cards.

Three Big Ideas

Marginal, joint, and conditional probabilities

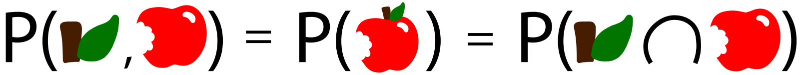
Bayes' rule

Independence

Marginal and Joint Probability

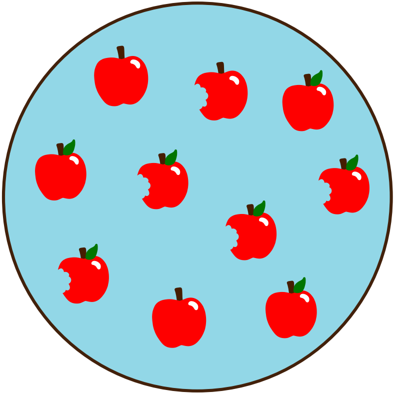
So far we have only considered situations where we are interested in the probability of a single event A occurring. We've denoted this $P(A)$. $P(A)$ is sometimes called a marginal probability.

Suppose we are now in a situation where we would like to express the probability that an event A and an event B occur. This quantity is written as $P(A \cap B)$, $P(B \cap A)$, $P(A, B)$, or $P(B, A)$ and is the joint probability of A and B.



P() = ?

P(, ) = ?



Conditional Probability

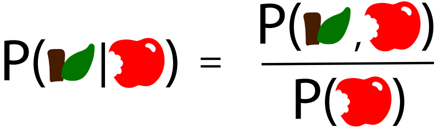
The "soul of statistics"
If $P(A) > 0$ then the probability of B conditional on A can be written as

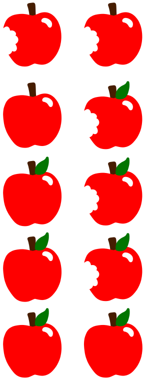
$$P(B|A) = \frac{P(A, B)}{P(A)}$$

This implies that

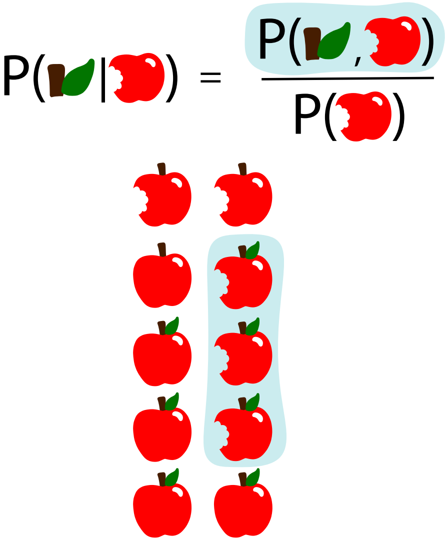
$$P(A, B) = P(A) \times P(B|A)$$

Conditional Probability: A Visual Example

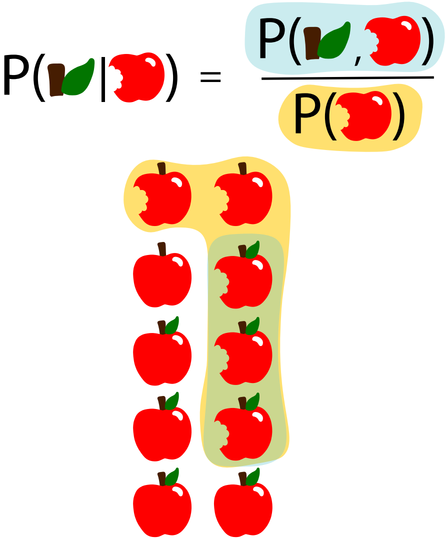




Conditional Probability: A Visual Example



Conditional Probability: A Visual Example



A Card Player's Example

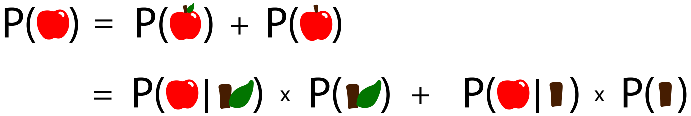
If we randomly draw two cards from a standard 52 card deck and define the events
$A = \{\text{King on Draw 1}\}$ and $B = \{\text{King on Draw 2}\}$, then

- $P(A) = 4/52$
- $P(B|A) = 3/51$
- $P(A, B) = P(A) \times P(B|A) = 4/52 \times 3/51 \approx .0045$

Law of Total Probability (LTP)

With 2 Events:

$$\begin{aligned} P(B) &= P(B, A) + P(B, A^c) \\ &= P(B|A) \times P(A) + P(B|A^c) \times P(A^c) \end{aligned}$$



Recall, if we randomly draw two cards from a standard 52 card deck and define the events $A = \{\text{King on Draw 1}\}$ and $B = \{\text{King on Draw 2}\}$, then

- $P(A) = 4/52$
- $P(B|A) = 3/51$
- $P(A, B) = P(A) \times P(B|A) = 4/52 \times 3/51$

Question: $P(B) =?$

Confirming Intuition with the LTP

$$\begin{aligned} P(B) &= P(B, A) + P(B, A^c) \\ &= P(B|A) \times P(A) + P(B|A^c) \times P(A^c) \end{aligned}$$

$$\begin{aligned} P(B) &= 3/51 \times 1/13 + 4/51 \times 12/13 \\ &= \frac{3 + 48}{51 \times 13} = \frac{1}{13} = \frac{4}{52} \end{aligned}$$

Example: Voter Mobilization

Suppose that we have put together a voter mobilization campaign and we want to know what the probability of voting is after the campaign: Pr[vote]. We know the following:

- $\Pr(\text{vote}|\text{mobilized}) = 0.75$
- $\Pr(\text{vote}|\text{not mobilized}) = 0.15$
- $\Pr(\text{mobilized}) = 0.6$ and so $\Pr(\text{ not mobilized}) = 0.4$

Note that mobilization partitions the data. Everyone is either mobilized or not. Thus, we can apply the LTP:

$$\begin{aligned}\Pr(\text{vote}) =& \Pr(\text{vote}|\text{mobilized}) \Pr(\text{mobilized}) + \\ & \Pr(\text{vote}|\text{not mobilized}) \Pr(\text{not mobilized}) \\ =& 0.75 \times 0.6 + 0.15 \times 0.4 \\ =& .51\end{aligned}$$

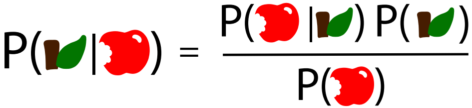
Bayes' Rule

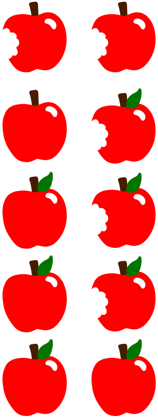
- Often we have information about $\Pr(B|A)$, but require $\Pr(A|B)$ instead.
- When this happens, always think: Bayes' rule
- Bayes' rule: if $\Pr(B) > 0$, then:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

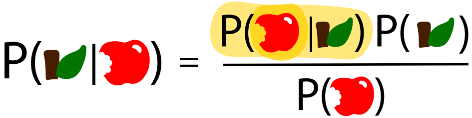
- Proof: combine the multiplication rule $\Pr(B|A)\Pr(A) = P(A \cap B)$, and the definition of conditional probability

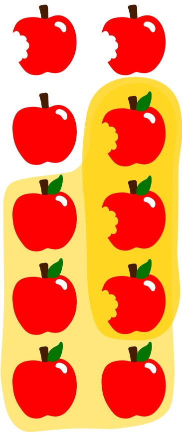
Bayes' Rule Mechanics



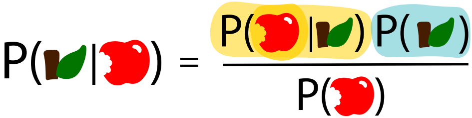


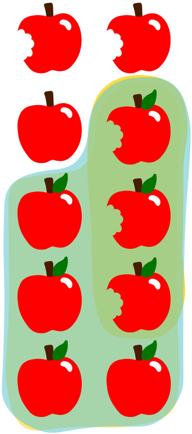
Bayes' Rule Mechanics



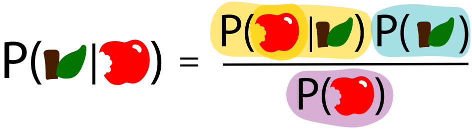


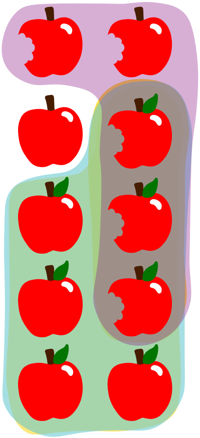
Bayes' Rule Mechanics





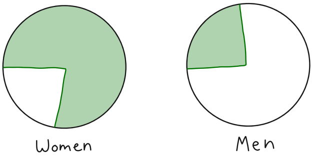
Bayes' Rule Mechanics





Bayes' Rule Example

U.S. Billionaires, 2014

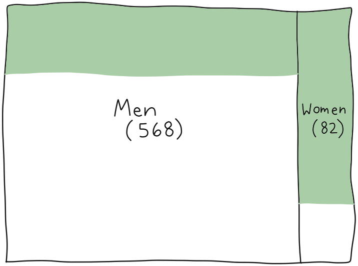


- 76.5% of female billionaires inherited their fortunes, compared to 24.5% of male billionaires
- So is P(woman | inherited billions) greater than P(man | inherited billions)?

$$P(W|I) = \frac{P(I|W)\,P(W)}{P(I)}$$

$$= \frac{.765\left(\frac{82}{568+82}\right)}{\frac{.765(82) + .245(568)}{568+82}}$$

$$= .31$$



*Data source = Billionaires characteristics database

Example: Race and Names

- Enos (2015): how do we identify a person's race from their name?
- First, note that the Census collects information on the distribution of names by race.
- For example, Washington is the most common last name among African-Americans in America:
  - $\Pr(\text{AfAm}) = 0.132$
  - $\Pr(\text{not AfAm}) = 1 - \Pr(\text{AfAm}) = .868$
  - $\Pr(\text{Washington}|\text{AfAm}) = 0.00378$
  - $\Pr(\text{Washington}|\text{not AfAm}) = 0.000061$
- We can now use Bayes' Rule

$$\Pr(\text{AfAm}|\text{Wash}) = \frac{\Pr(\text{Wash}|\text{AfAm})\Pr(\text{AfAm})}{\Pr(\text{Wash})}$$

Example: Race and Names

Note we don't have the probability of the name Washington.

Remember that we can calculate it from the LTP since the sets African-American and not African-American partition the sample space:

$$
\begin{aligned}
\Pr(\text{AfAm}|\text{Wash}) &= \frac{\Pr(\text{Wash}|\text{AfAm})\Pr(\text{AfAm})}{\Pr(\text{Wash})} \\
&= \frac{\Pr(\text{Wash}|\text{AfAm})\Pr(\text{AfAm})}{\Pr(\text{Wash}|\text{AfAm})\Pr(\text{AfAm}) + \Pr(\text{Wash}|\text{not AfAm})\Pr(\text{not AfAm})} \\
&= \frac{0.132 \times 0.00378}{0.132 \times 0.00378 + .868 \times 0.000061} \\
&\approx 0.9
\end{aligned}
$$

Independence

Intuitive Definition

Events A and B are independent if knowing whether A occurred provides no information about whether B occurred.

Formal Definition

$$P(A, B) = P(A)P(B) \implies A \perp\!\!\!\perp B$$

With all the usual > 0 restrictions, this implies

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$

Independence is a massively important concept in statistics.

Next Week

- Homework do on Thursday (8 days)! Why have homework assigned on the first day?
- Random Variables
- Reading for Random Variables
  - Blitzstein and Hwang Chapters 2, 3-3.2 (random variables), 4-4.2 (expectation), 4.4-4.6 (indicator rv, LOTUS, variance), 5.1-5.4 (continuous random variables), 7.0-7.3 (joint distributions)
  - Optional: Imai Chapter 6 (probability), Aronow and Miller Chapter 2
- A word from your preceptors

References

- Enos, Ryan D. "What the demolition of public housing teaches us about the impact of racial threat on political behavior." *American Journal of Political Science* (2015).
- J. Andrew Harris "Whats in a Name? A Method for Extracting Information about Ethnicity from Names" *Political Analysis* (2015).
- Salsburg, David. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century* (2002).