ST790 — Fall 2022
Imprecise-Probabilistic Foundations of Statistics

Ryan Martin
North Carolina State University
www4.stat.ncsu.edu/~rmartin

Week 14a

This lecture

- Simpson's paradox
  - general setup
  - what happens, why it's scary, etc.
  - foundational implications
  - connection to imprecise probability
- ....

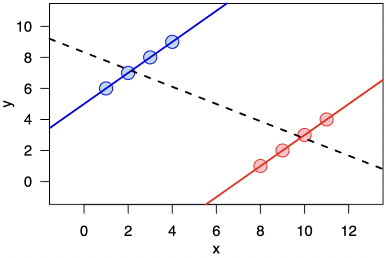
Introduction

- *Simpson's paradox* sometimes appears in "applied" courses
- Very roughly, Simpson's paradox[1] concerns cases where
  - conclusions go one way when a variable is included
  - but go completely the other way when it's excluded
- In other words, two "correct" statistical analyses of the same data can reach completely opposite conclusions
- Counter-intuitive, hence the name "paradox"
- There are connections between Simpson's paradox and
  - foundations of statistics
  - imprecise probability

[1] Idea is from Simpson (*JRSS-B* 1951), name is from Blyth (*JASA* 1972)

What is it?

- (X, Y) relationship?
- Counter-intuitive:
  - ignoring the colors, the trend is negative
  - but *both* color-specific trends are positive
- More formally:
  - two joint dist's,
    - $\rightarrow$ (X, Y)
    - $\rightarrow$ $\{(X, Y) \mid Z = z\}$
  - can be very different



https://en.wikipedia.org/wiki/Simpson's_paradox

Why does it matter?

- Two "correct" statistical analyses of the same data can reach completely opposite conclusions!
- Lots of real examples of this; one is below
- Can have major scientific and sociological consequences
- Relates to *How to lie with statistics*
  - intentionally
  - or unintentionally
- Have to be aware of the issue to avoid getting bitten by it

Example: UC Berkeley gender bias study[2][3]

- Famous admissions study from 1973
- Y = admitted (yes/no), X = gender (male/female)
- Data show signs of gender bias favoring male applicants

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

- i.e., $\mathrm{P}(Y = \text{yes} \mid X = \text{male}) \gg \mathrm{P}(Y = \text{yes} \mid X = \text{female})$
- Significant, can't be explained by sampling variation
- Fire administrators? More "inclusive" admission policies? ...

[2] `https://en.wikipedia.org/wiki/Simpson's_paradox`

[3] Another *racial bias* example in my ST503 Week 12b materials

Example, cont.

- Previous results ignore department-level data
- Z = department applied for (A, B, C,...)
- Different picture emerges after conditioning on Z

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

- We'd probably regret firing administrators or implementing admissions policies that penalized male applicants

Example, cont.

- Both of the analyses are "correct"
  - one makes conclusions about the marginal association
  - the other makes conclusions about conditional associations
- The question is which of these is most relevant
- Unfortunately, most people doing these analyses — and the ones retweeting the conclusions — don't know the difference
- Analysis with the conclusion that best aligns with one's opinion or agenda is the "correct" one
- Our statistics education (undergrad to PhD) falls way short on issues like this beyond "applied" and "theory"

Why is Simpson's paradox scary?

- The issue with Simpson's paradox, apparent in the plot and the gender bias illustration, is this:

  the marginal analysis doesn't show the whole story

- When we have access to the relevant variable Z, then it's arguably misleading to ignore/hide/marginalize it
- But every real-life problem is like this

$$(X, Y, \underbrace{Z_1^{(\mathrm{m})}, Z_2^{(\mathrm{m})}, \ldots}_{\text{measured}}, \underbrace{Z_1^{(\mathrm{u})}, Z_2^{(\mathrm{u})}, \ldots}_{\text{unmeasured}})$$

- *Question:* How do we know that the "whole story" we're after isn't contained in the *unmeasured* Z's?
- No method, algorithm, or theorem can protect us from this...

Math behind Simpson's paradox

- Focus on the discrete case, $\mathbb{X} = \mathbb{Y} = \{0, 1\}$
- We're looking for a situation in which
  - $\mathrm{P}(Y = 1 \mid X = 1, z) > \mathrm{P}(Y = 1 \mid X = 0, z)$ for all z
  - and $\mathrm{P}(Y = 1 \mid X = 1) < \mathrm{P}(Y = 1 \mid X = 0)$
- Total probability formula

$$\mathrm{P}(Y = 1 \mid X = 1) = \sum_z \mathrm{P}(Y = 1 \mid X = 1, z)\,\mathrm{P}(Z = z \mid X = 1)$$

$$\mathrm{P}(Y = 1 \mid X = 0) = \sum_z \mathrm{P}(Y = 1 \mid X = 0, z)\,\mathrm{P}(Z = z \mid X = 0)$$

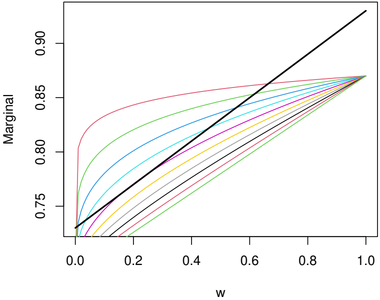
- No restrictions on $\mathrm{P}(Z = z \mid X = x)$
- So Simpson's reversal can happen when the distribution of Z, given $X = x$, differs greatly depending on x
- Clearly can't happen if Z is independent of X

Math behind Simpson, cont.

- Two sets of conditionals:
  - $p^1 = (0.93, 0.73)$
  - $p^0 = (0.87, 0.69)$
- Two marginals:

$$m^1(w) = wp_1^1 + (1-w)p_2^1$$
$$m^0(w) = w^r p_1^0 + (1-w^r)p_2^0$$

- Range of $r \in (0, 1]$
- When r is small:
  - $z \mid x$ depends on x
  - $m^0(w) > m^1(w)$



Foundational implications

- Recall the recent discussion of decision theory and the von Neumann–Morganstern result
- Roughly, vNM says that the following are equivalent:
  - agent makes decisions "rationally"
  - he has a probability–utility function pair and his decisions are based on maximizing expected utility
- Clear implications for Bayesian decision theory, etc.
- Leonard "Jimmie" Savage[4] aimed to do more, to demonstrate that *probability* is essential to this
- Surprisingly, Simpson's paradox turned out to be an obstacle

[4] *The Foundations of Statistics* book, 1954

Foundations, cont.

- Crucial to Savage's arguments is the *sure-thing principle*
- Roughly:
  - partition $\mathcal{B}$ of the universe $\mathbb{X}$
  - suppose $f^\star \succeq f$ when I know which $B \in \mathcal{B}$ occurs
  - STP: $f^\star \succeq f$ even when I don't know which $B \in \mathcal{B}$ occurs

Savage's example of the *sure-thing principle*.

A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he knew that the Democratic candidate were going to win, and decides that he would. Similarly, he considers whether he would buy if he knew that the Republican candidate were going to win, and again finds that he would. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event obtains...

Foundations, cont.

- Blyth's 1972 *JASA* paper spotted an issue with STP
  - basically, STP is based on an assumption that conditional preferences determine marginal preferences
  - Simpson's paradox is a counter-example
- Consider a modified version of Savage's story:[5]
  - shrink the scale, replace "president" with "mayor"
  - businessman doesn't support Democratic mayoral candidate
  - still makes the judgment that owning the property is better than not, regardless of who is mayor
  - *but now he worries that buying the property pre-election might help the Democratic candidate win*
- So: the businessman isn't necessarily irrational to wait until after the election to purchase the property, even though the purchase is justified regardless of who is mayor

[5] This is based on Judea Pearl's paper entitled "The sure-thing principle"

Connections to imprecise probability

- Simpson's paradox has the makings of a sure-loss situation
  - if my conditional probabilities are are ordered one way,
  - and my marginal probabilities are ordered the opposite
  - then there's a mistake or something weird is going on
  - so there must be a way to make me a fool
- Gong & Meng's[6] Theorem 5.1 states this formally
- Can imprecise probability help to avoid Simpson's paradox?
- In principle, *yes* — just introduce a vacuous model for the unmeasured variables
- But this only creates a new problem...
  - too much uncertainty, you handcuffed yourself
  - can't learn from what you observed

[6] `https://ruobingong.github.io/files/GongMeng2021_StatSci.pdf`

Next lecture

- Course wrap-up