Ec140 - Causality and Selection Bias

Fernando Hoces la Guardia
06/29/2022

Today's Lecture

- Our First Causal Question in Real Life
    - Causality
    - Correlation v. Causation
    - Other things equal
- Selection Bias

Causal Inference to Inform Policy: Setting

Access to health care insurance is a huge political issue in the US. Subsidizing the provision and mandating the adoption of insurance was at the core of the, heavily debated, Affordable Health Care Act, also known as *Obamacare*.

Policy: Subsidize, and/or enforce, a health care insurance for the entire population.

Rationale: Increasing access to health care (through insurance), can improve the health outcomes of the population.

- Can you think of another rationale?

Let's look at some data to investigate this rationale.

National Health Interview Survey, 2009

- This is just a random sample of 100 observations from the real dataset. The complete data contains 80634 observations (individuals).

2009 National Health Interview Survey

	Insurance	Female?	Age	Health	Weight
1	1	1	24	3	1134
2	1	1	26	3	5514
3	1	1	15	4	3321
4	1	1	37	3	3466
5	0	1	43	2	6609
6	0	1	22	3	2360

Showing 1 to 6 of 100 entries

Previous 1 2 3 4 5 ... 17 Next

National Health Interview Survey, 2009

- This is just a random sample of 100 observations from the real dataset. The complete data contains 80634 observations (individuals).
- What tools from the course (so far) should we use to look at this data?

2009 National Health Interview Survey

	Insurance	Female?	Age	Health	Weight
1	1	1	24	3	1134
2	1	1	26	3	5514
3	1	1	15	4	3321
4	1	1	37	3	3466
5	0	1	43	2	6609
6	0	1	22	3	2360

Showing 1 to 6 of 100 entries

Previous 1 2 3 4 5 ... 17

Next

National Health Interview Survey, 2009 (MM, Ch1)

TABLE 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
B. Characteristics						
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

National Health Interview Survey, 2009: Notes

TABLE 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
B. Characteristics						
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

National Health Interview Survey, 2009: Notes

Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

Let's Read This Summary Statistics

- $\mathbb{E}(Y|X)$?
- σ ?

TABLE 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
B. Characteristics						
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47	3.49	3.93	−.43

National Health Interview Survey, 2009 (MM, Ch1)

- Can we interpret these differences **causally**?

TABLE 1.1
graphic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
I	No HI (2)	Difference (3)		Some HI (4)	No HI (5)	Difference (6)
		A. Health				
	3.70 [1.01]	.31 (.03)		4.02 [.92]	3.62 [1.01]	.39 (.04)
		B. Characteristics				
	.17	−.01 (.01)		.15	.17	−.02 (.01)

The Concept of Causality

Causality: what are we talking about?

- We say that X *causes* Y
  - if we were to intervene and *change* the value of X ***without changing anything else***...
  - then Y would also change ***as a result***.
- The key point here is the ***without changing anything else***, often referred as the **other things equal** assumption (or *ceteris paribus* if you want to sound fancy).
- ⚠️ It does **NOT** mean that X is the only factor that causes Y.

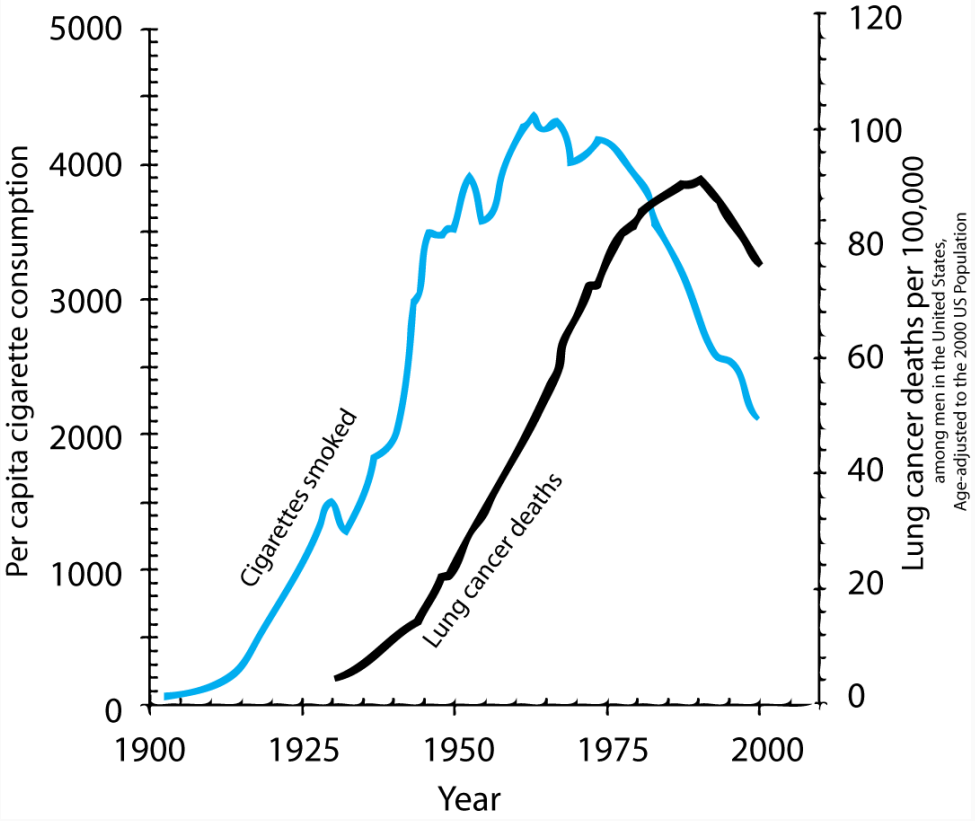
Correlation vs Causation

Correlation does not equal causation has become a ubiquitous mantra, but can you tell why it is true?

Some correlations obviously don't imply causation (e.g. spurious correlation website).

Correlation vs Causation: Smoking and Lung Cancer

But not all correlations are so easy to rule out



Does smoking cause lung cancer?

- Today, we know the answer is *YES*!
- But let's go back in the 1950's
  - We are at the start of a big increase in deaths from lung cancer...
  - ... which is happening after a fast growth in cigarette consumption
- It's very tempting to claim that smoking causes lung cancer based on this graph.

Correlation vs Causation: Smoking and Lung Cancer

At the time many people were still skeptical, including some famous statisticians:

Macro confounding factors:

Other macro factors which can cause cancers also changed between 1900 and 1950:

- Tarring of roads,
- Inhalation of motor exhausts (leaded gasoline fumes),
- General greater air pollution.

Self selection:

Smokers and non-smokers may be different in the first place:

- **Selection on observable characteristics**: age, education, income, etc.
- **Selection on unobservable characteristics**: genes (the hypothetical confounding genome theory of Fisher).

Back to Our Original Example: Health and Health Insurance

- Can we interpret these differences **causally**?
- Are all **other things equal** between insured and uninsured?

TABLE 1.1
graphic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
I	No HI (2)	Difference (3)		Some HI (4)	No HI (5)	Difference (6)
		A. Health				
	3.70 [1.01]	.31 (.03)		4.02 [.92]	3.62 [1.01]	.39 (.04)
		B. Characteristics				
	.17	−.01 (.01)		.15	.17	−.02 (.01)

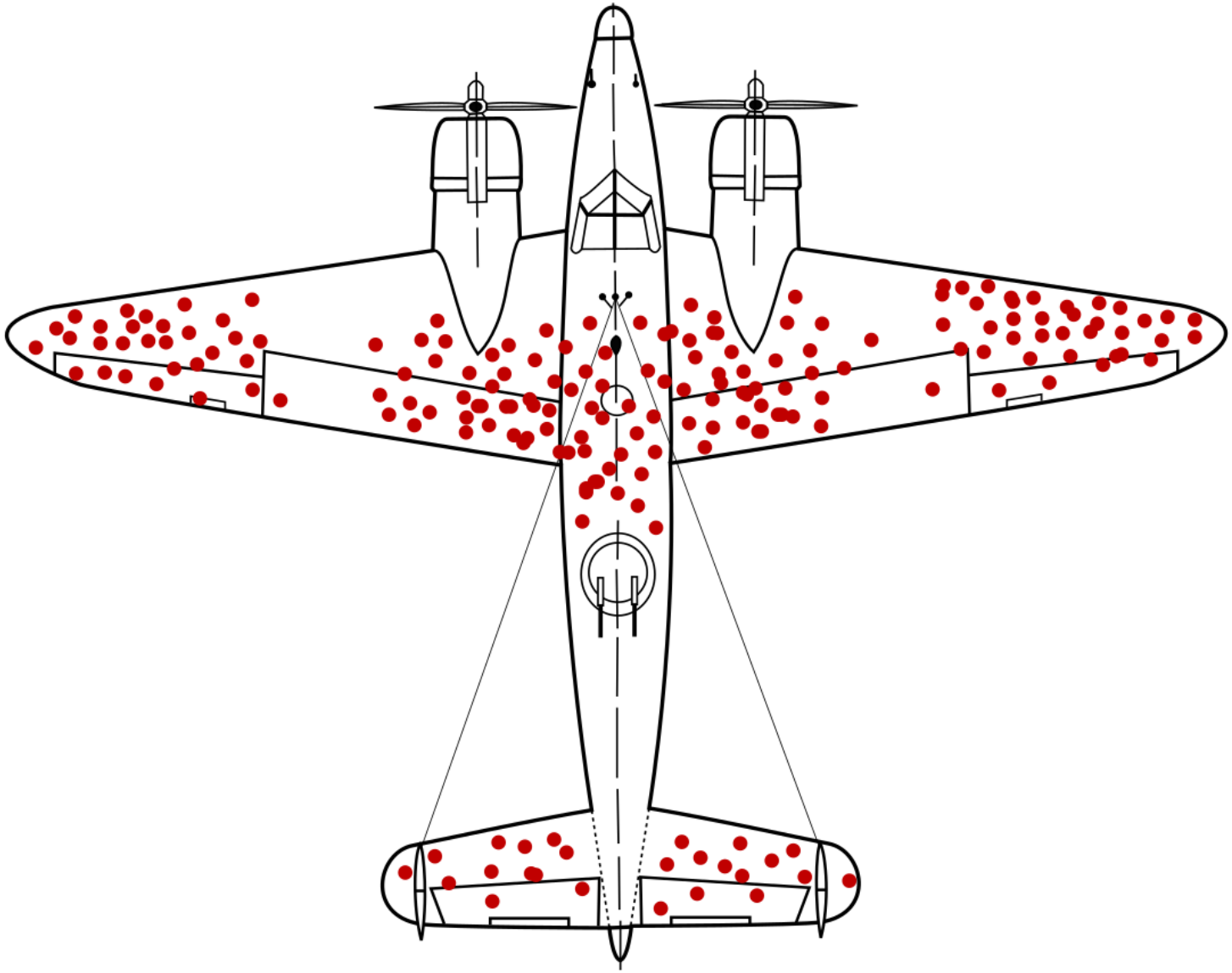
Selection Bias

Wikipedia Definition:

> Selection bias is the bias introduced by the selection of individuals, groups, or data for analysis in such a way that proper randomization is not achieved, thereby failing to ensure that the sample obtained is representative of the population intended to be analyzed.

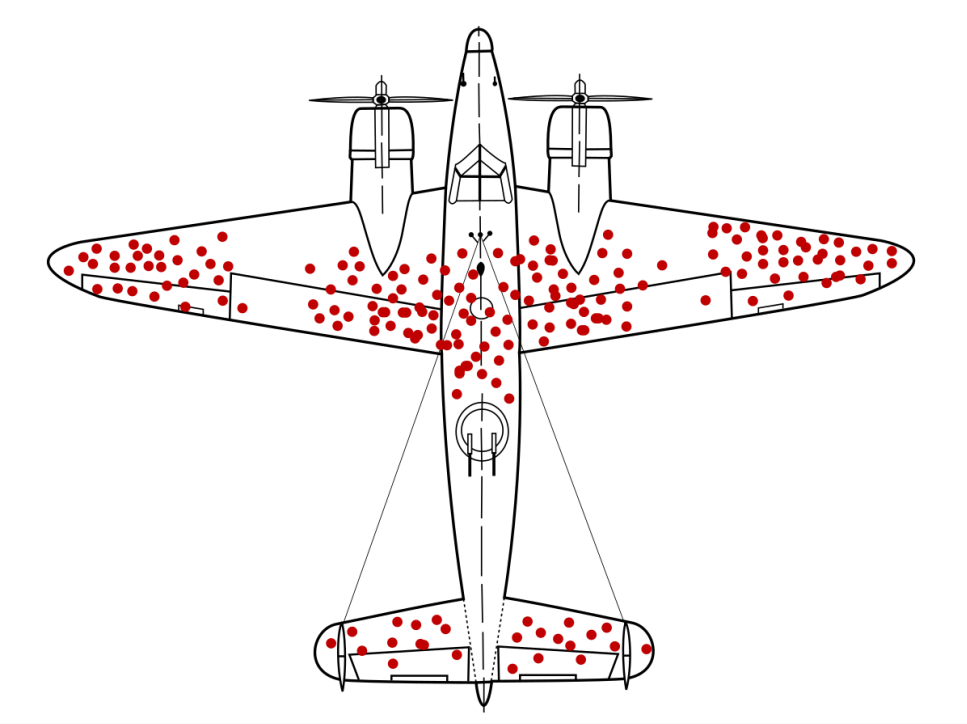
- Econometric textbooks, tend to define selection bias in term of a regression or (as MM) a randomized controlled trial.
- We will start from this more general definition to connect with the concept of **conditional expectation**.

SB Example 1: Airplanes in World War II



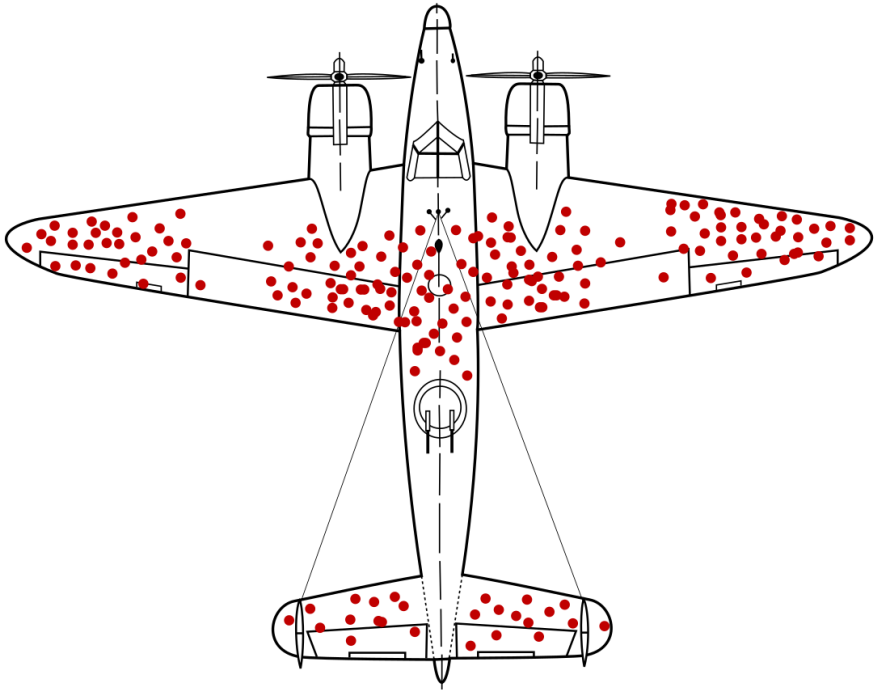
SB Example 1: Airplanes in World War II. Using Expectation 1/2

- How would you use conditional expectations to characterize this problem?
- Let's start by simplifying the problem by assuming that each plane only had two sections. Now define two random variables: binary variables (bernulli) to indicate if the plane received damage in locations one, and two.
  $(DL1 : \{\text{No damaged in lct 1, Damaged in lct1}\} \rightarrow \{0, 1\}$
  , same for $DL2)$.
- We also need to define random variable for that we are conditioning on. In this case, let's use a binary variable for return
  $(R : \{\text{Plane didn't return, Plane returned}\} \rightarrow \{0, 1\})$



SB Example 1: Airplanes in World War II. Using Expectation 2/2

- One way of characterizing the problem would be that the engineers thought they where observing $\mathbb{E}(DL1)$ and $\mathbb{E}(DL2)$ and concluding $\mathbb{E}(DL1) > \mathbb{E}(DL2)$.
- But in they were actually observing $\mathbb{E}(DL1|R = 1)$ and $\mathbb{E}(DL2|R = 1)$ and most likely $\mathbb{E}(DL1|R = 0) < \mathbb{E}(DL2|R = 0)$
- If you don't like the math notation, you can provide the same answer, but in narrative form.
- This is called survivorship bias, and is a type of selection bias.



SB Example 2: Health Insurance 1/2

- We can do something similar for our health insurance example.
- The "hidden" information could be many things. For example: maybe uninsured people are less have different standards of what constitutes good health, and for the same true health status, uninsured tend to report much higher scores than insured (thanks Andy!).

TABLE 1.1
graphic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
I	No HI (2)	Difference (3)		Some HI (4)	No HI (5)	Difference (6)
	A. Health					
	3.70 [1.01]	.31 (.03)		4.02 [.92]	3.62 [1.01]	.39 (.04)
	B. Characteristics					
	.17	−.01 (.01)		.15	.17	−.02 (.01)
	41.26	2.71		42.24	39.62	2.62

SB Example 2: Health Insurance 2/2

- Define a binary random variable that represents if an individual tends to over report good health or not $(ORep : \{\text{no over report, over reports}\} \to \{0, 1\})$. In this case the previous comparison translates into:
- $\mathbb{E}(H|HI = 1, Orep = 1)$ for column (4), and $\mathbb{E}(H|HI = 0, Orep = 0)$ for column (5).
- This is a violation of *other things equal* assumption.

TABLE 1.1
graphic characteristics of insured and uninsured couples in the NHIS

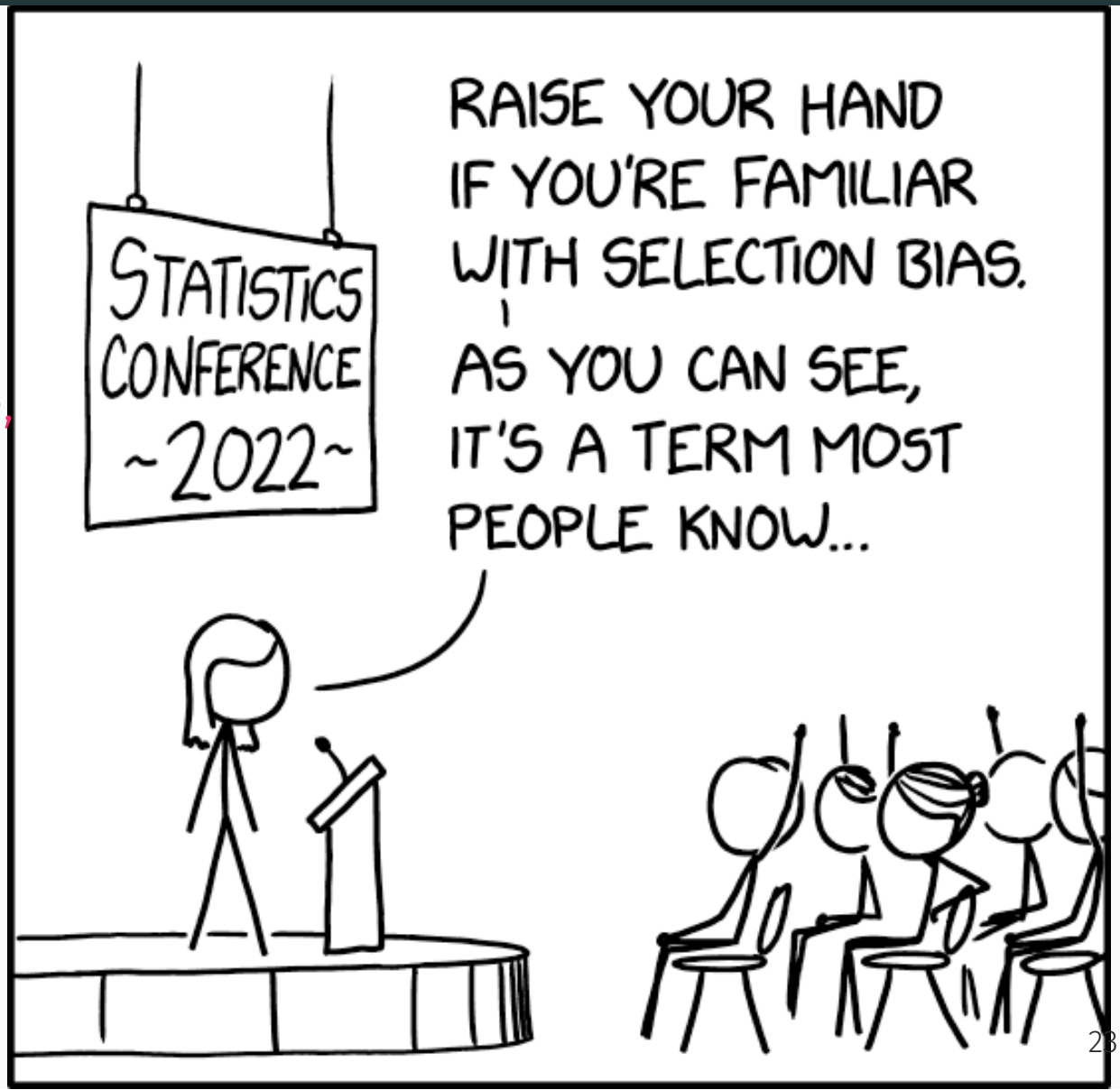
Husbands			Wives		
I	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
	A. Health				
	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
	B. Characteristics				
	.17	−.01 (.01)	.15	.17	−.02 (.01)
	41.26	2.71	42.24	39.62	2.62

SB Example 3: Country Characterization by Foreign Visitors

- Characterization of Americans according to foreigners visiting Berkeley.
- Characterization of Chinese according to foreigner visiting a specific city.

More Examples

- Convention of Statisticians. XQCD
- Heike Crabs
- Appearance and Intelligence of Movie Stars (From Causal Inference, The Mixtape)
- Think of at least two examples yourself!
- (Hernan Cascicari on Surveys [in Spanish, and strong language warning])



Acknowledgments

- Kyle Raze's Undergraduate Econometrics 1
- SoPo
- XQCD
- MM
- Matt Hollian
- Causal Mixtape (Also Hanny Fry)
- The plane pic
- MM bookdown and MM blog post