CS109/Stat121/AC209/E-109
Data Science
Bias and Sampling

Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu

This Week

- HW1 due tonight at 11:59 pm
- HW2 will be posted by tonight – start soon!
- Friday lab **10-11:30 am** in MD G115
  - *Pandas* with Rahul, Brandon, and Steffen

Some Forms of Bias

- selection bias
- publication bias (file drawer problem)
- censoring bias
- length bias
- sampling bias

Longevity Study

Profession	Average Longevity
chocolate maker	73.6
professors	66.6
clocksmiths	55.3
locksmiths	47.2
students	20.2

Sources: Lombard (1835), Wainer (1999), Stigler (2002)

Google macy's SIGN IN

Web Images Maps Shopping News More Search tools

About 24,200,000 results (0.22 seconds)

Ad related to **macy's**

Macys.com - Macy's® - Official Site
www.macys.com/ ★★★★☆ 48 seller reviews
Discover the Hottest Fashion Trends & Newest Brands at **Macy's**!
Macy's has 130,257 followers on Google+

Wedding Registry
Create a Registry or Buy a Wedding Gift from the Registry Gift Finder

20% Off Sale - Code: VIP
Secret 20% Off VIP Sale Ends Soon! Online & Instore - Use code: VIP

Macy's - Shop Fashion Clothing & Accessories - Official Site - Macys ...
www.macys.com/
Macy's - FREE Shipping at **Macys**.com. **Macy's** has the latest fashion brands on Women's and Men's Clothing, Accessories, Jewelry, Beauty, Shoes and Home ...
2.7 ★★½☆☆ 9 Google reviews · Write a review

Arden B. - Cambridgeside Galleria 100 Cambridgeside Pl, Cambridge, MA 02141
(617) 621-3800

Women's Clothing
Shop the Latest Designer Clothing for Women Online at Macys.com.

For The Home
Window Treatments - Electronics - Home Decor - Rugs - Mattresses

Macy's Wedding Registry
Macy's Wedding Registry- Create, modify or search a bridal ...

Store Locations & Hours
Events · Macy's Event Marketing · Shopping Services · Store ...

Shoes
Women's Shoes - Espadrilles and Wedges - Womens Sneakers - ...

Macy's in Cambridge, MA
Macy's in Cambridge is located at 100 Cambridge Side Place ...

More results from macys.com »

Macy's

Directions Write a review

Address: Arden B. - Cambridgeside Galleria, 100 Cambridgeside Pl, Cambridge, MA 02141

Phone: (617) 621-3800

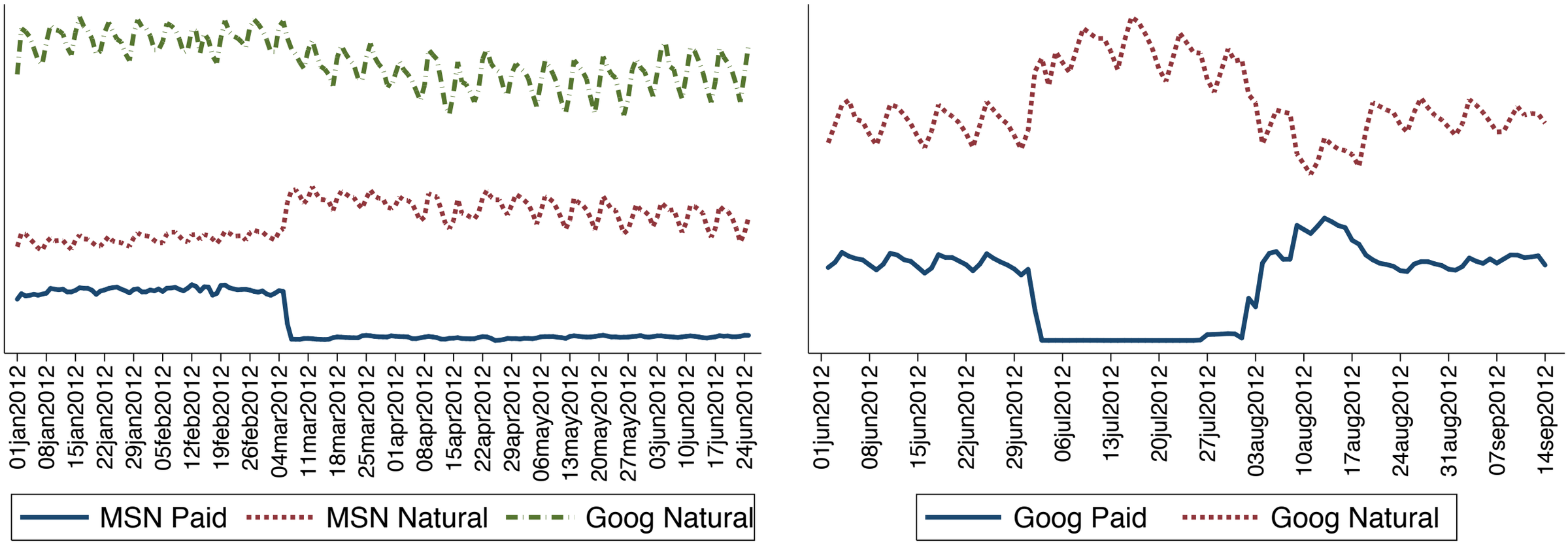
Prices: $$$$

Reviews

2.7 ★★½☆☆ 9 Google reviews

People also search for

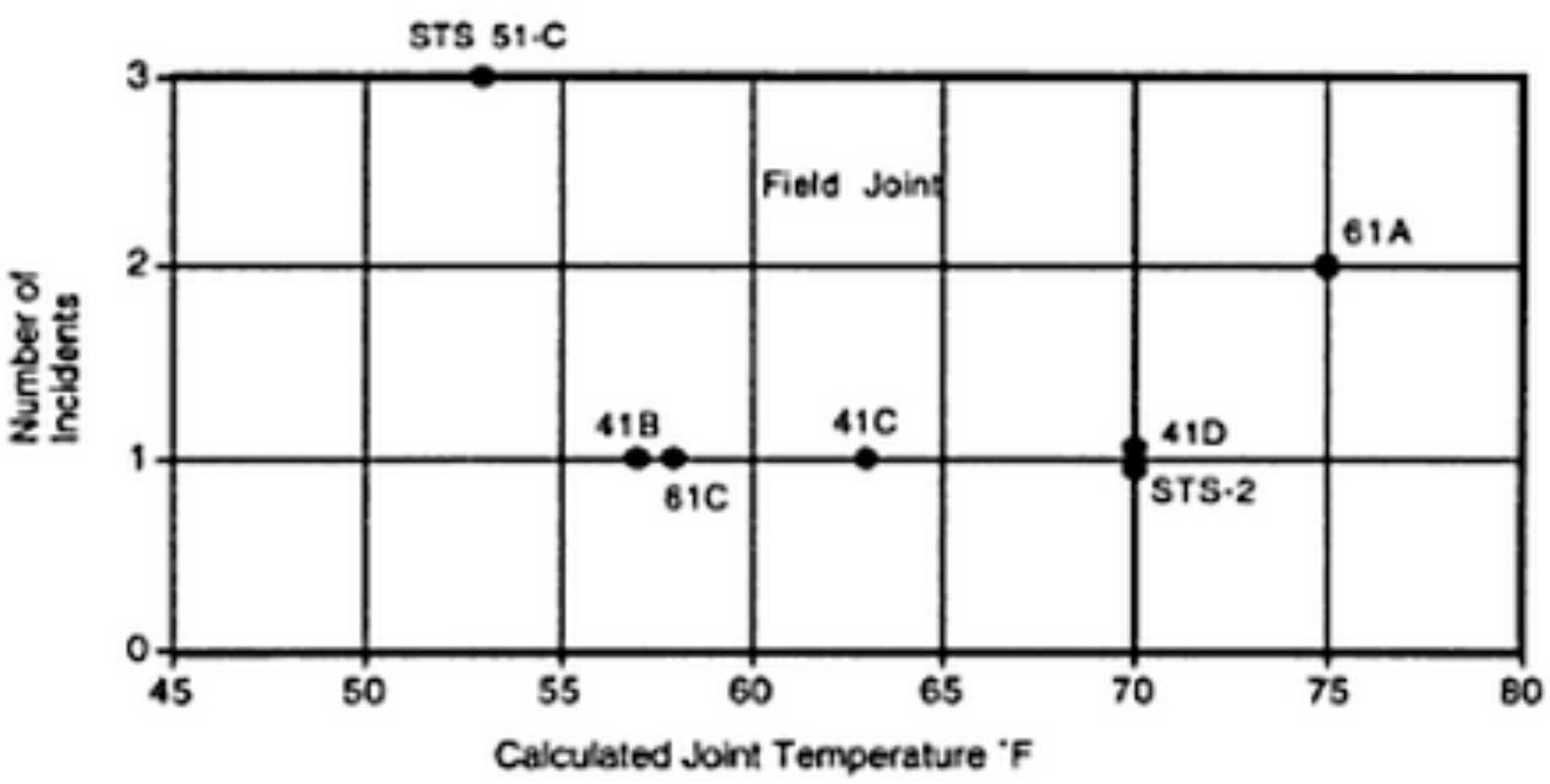
Figure 2: Brand Keyword Click Substitution

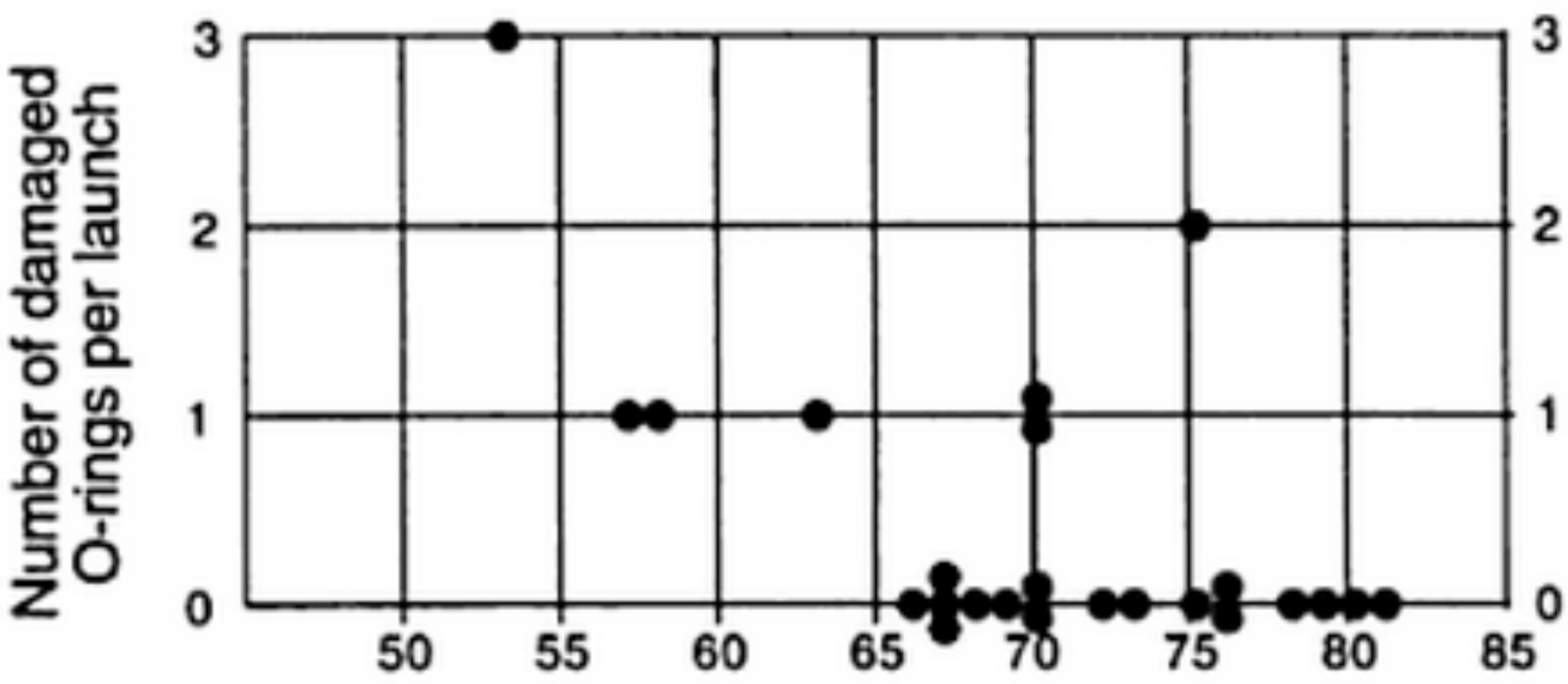


(a) MSN Test **(b) Google Test**

MSN and Google click traffic is shown for two events where paid search was suspended (Left) and suspended and resumed (Right).

result from Blake-Nosko-Tadelis (2013)
http://conference.nber.org/confer/2013/EoDs13/Tadelis.pdf





(Challenger Disaster)

Wainer (2000), *Visual Revelations*

Why sample from a population?

- often the only feasible option
- but it's useful to think about the question:
  What would you do if you had all the data?
- also often important for computational reasons

There are many sampling schemes...

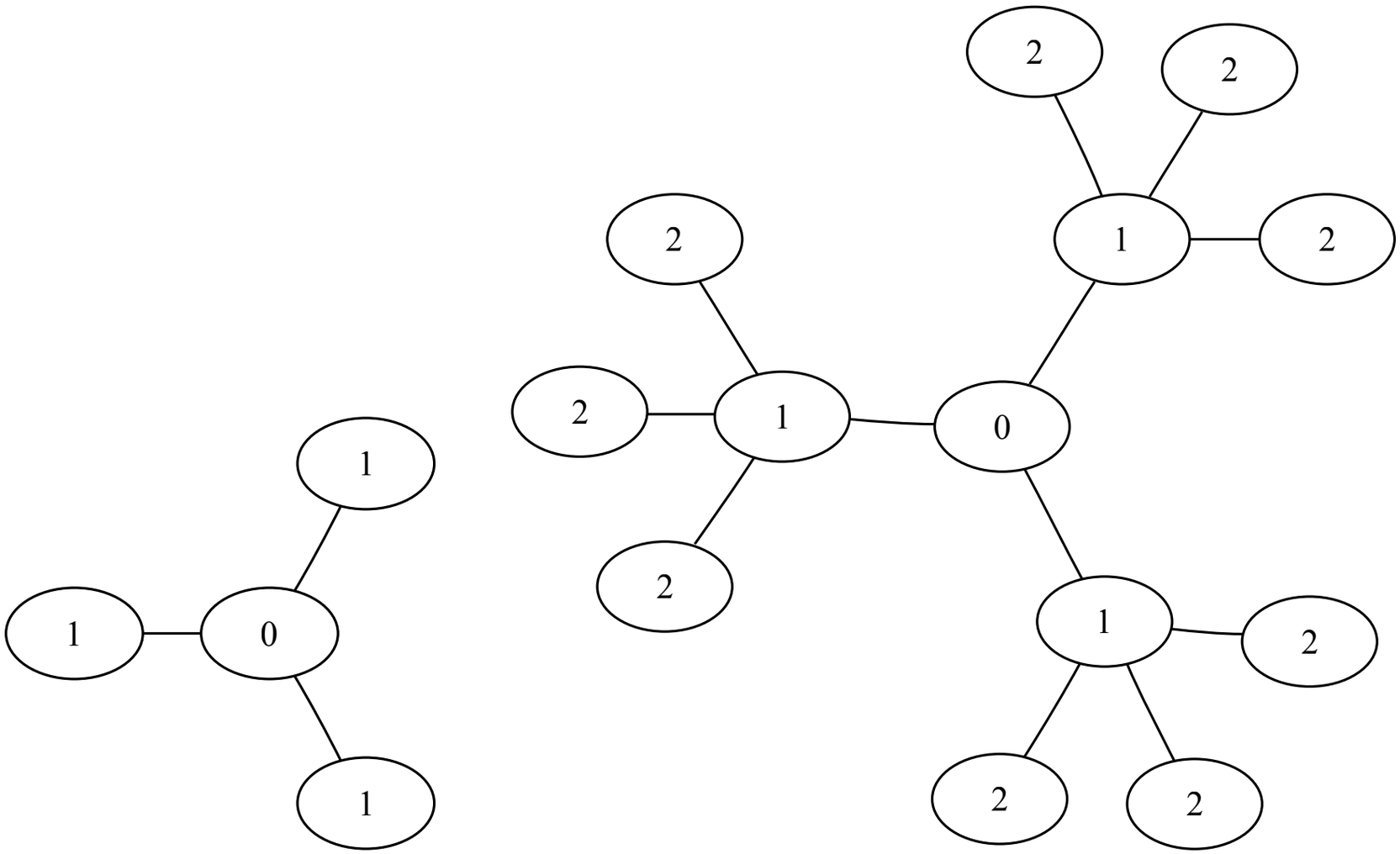
- simple random sampling
- stratified sampling
- cluster sampling
- snowball sampling

Absolute vs. relative

In simple random sampling, which matters more: the relative sample size, or the absolute sample size?

For example, how much bigger a sample should you collect in China vs. in the US, to get the same standard error?

Snowball Sampling (Link-Tracing)



(a) Stage 1 (b) Stage 2

Bias of an Estimator

The *bias* of an estimator is how far off it is on average:

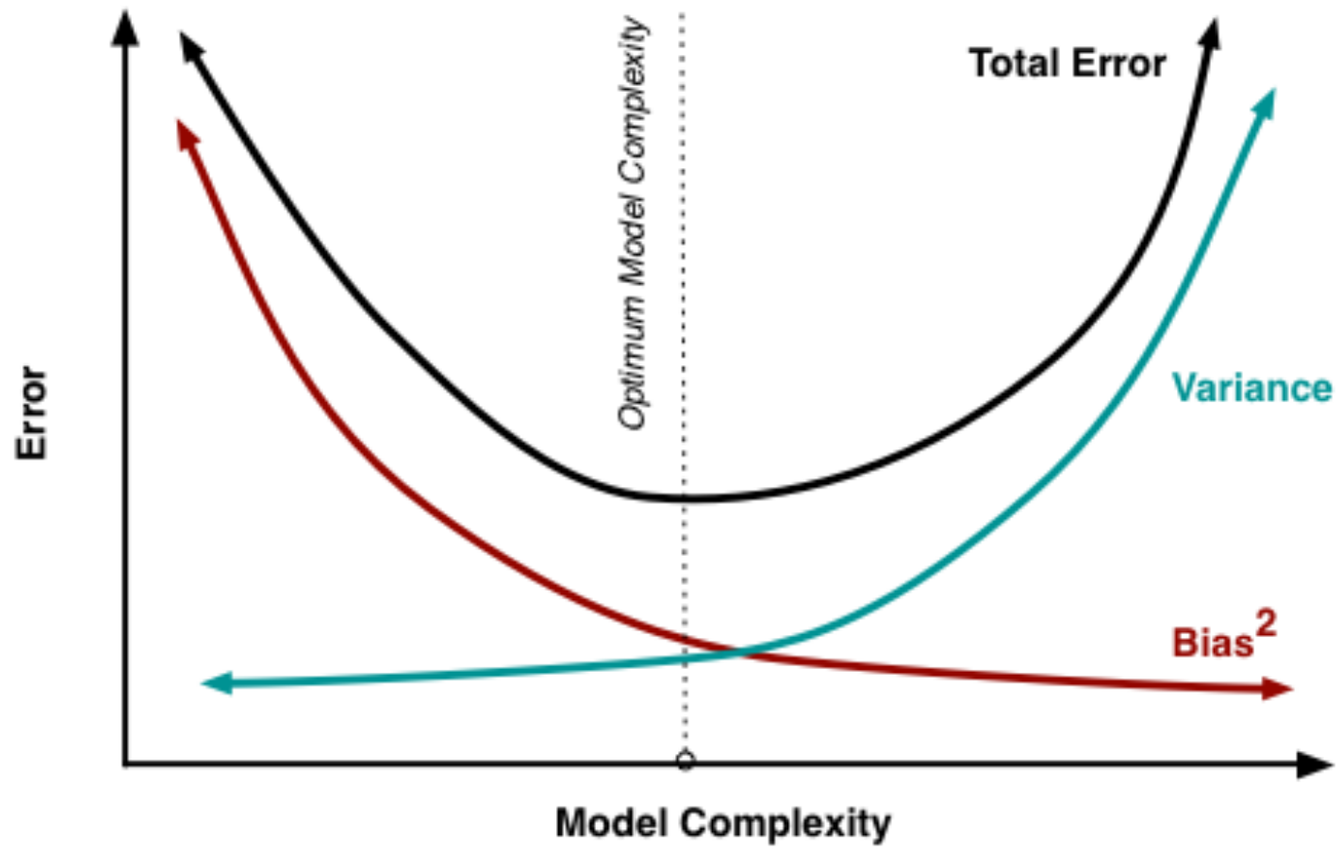
$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

So why not just subtract off the bias?

Bias-Variance Tradeoff

one form: $$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{bias}^2(\hat{\theta})$$

often a little bit of bias can make it possible to have much lower MSE



http://scott.fortmann-roe.com/docs/BiasVariance.html

Unbiased Estimation: Poisson Example

$$X \sim \text{Pois}(\lambda)$$

$$\text{Goal: estimate } e^{-2\lambda}$$

$(-1)^X$ is the best (and only) unbiased estimator of $e^{-2\lambda}$

sensible?

Basu's Elephant

Estimate the total weight of 50 elephants.

Horvitz-Thompson Estimator

Estimate the total of some variable for a finite population:

$$\hat{T}_y = \sum_{i \in S} \frac{y_i}{\pi_i}$$

where S is the sample and $\pi_i > 0$ is the probability of i being in the sample

Unbiased! But what about the variance?

Fisher Weighting

How should we combine independent, unbiased estimators for a parameter into one estimator?

$$\hat{\theta} = \sum_{i=1}^{k} w_i \hat{\theta}_i$$

The *weights* should sum to 1, but how should they be chosen?

$$w_i \propto \frac{1}{\mathrm{Var}(\hat{\theta}_i)}$$

(Inversely proportional to variance; why not SD?)