9.S918: Statistical Inference in Brain and Cognitive Sciences

Week 1 Day 2: Introduction to causal inference

Roger Levy
Dept. of Brain & Cognitive Sciences
Massachusetts Institute of Technology

April 4, 2024

A tiny bit of statistics

- On Tuesday we reviewed basics of **probability**: the logical calculus of uncertainty–a branch of mathematics
- The primary focus of this class is **statistics**: the mathematics, science, craft, and art of drawing inferences from data
- The two fields are fundamentally different
- But, probability is used extensively throughout statistics

Perhaps the simplest probability **distribution**

- Consider a binary random variable Y with two possible outcomes: 0 and 1
- Y is a **Bernoulli random variable** with **parameter** $P(\text{heads}) = \pi$, where $0 \leq \pi \leq 1$
- Figuring out from observed data what the weighting is likely to be is **parameter estimation**
- In general, we will use $\mathbf{y}$ to refer to observed-outcome **data** and θ to refer to the model parameters to be estimated

Statistical estimators

- **Estimator:** a procedure for guessing a quantity of interest within a population from a sample from that population
- For example, the **relative frequency estimator:** if we observe r instances of heads in n coin flips,

"this is an estimator"

$$\boxed{\widehat{}}\!\!\!\!\pi = \frac{r}{n}$$

- Data are stochastic, so estimators give random variables!
- **Bias** of an estimator is $E[\widehat{\theta}] - \theta$

Here we used **linearity of the expectation**

$$E[\widehat{\pi}] = E[\frac{r}{n}] = \frac{1}{n}E[r] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \frac{1}{n}n\pi = \pi$$

...so $\widehat{\pi}$ is **unbiased**

- **Variance** of an estimator is ordinary variance

$$\text{Var}(X) \equiv E[(X - E[X])^2] \qquad \text{Var}(\widehat{\pi}) = \frac{\pi(1-\pi)}{n}$$

(see reading materials)

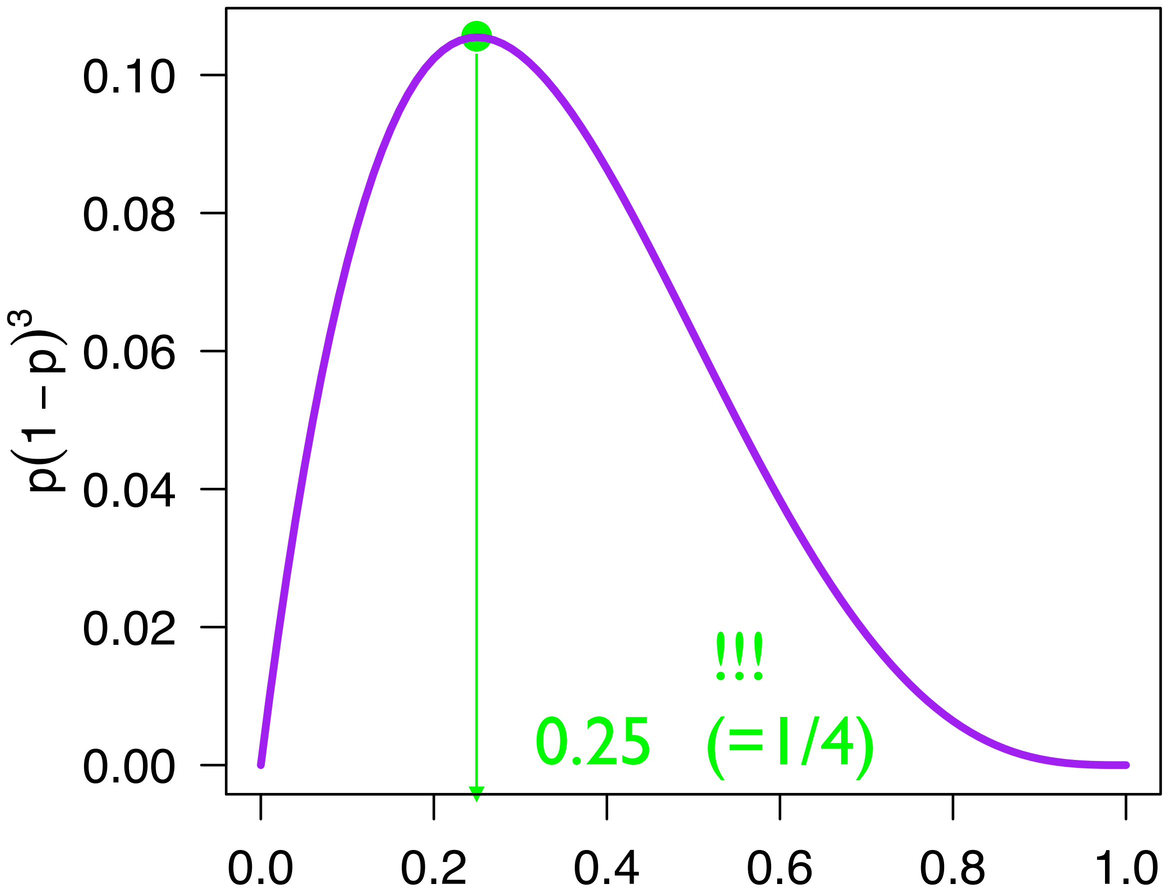
- Good estimators have favorable **bias–variance** tradeoff

Maximum likelihood estimation

$$\mathrm{Lik}(\boldsymbol{\theta};\boldsymbol{y}) \equiv P(\boldsymbol{y}|\boldsymbol{\theta}) \qquad \hat{\boldsymbol{\theta}}_{MLE} \stackrel{\mathrm{def}}{=} \arg\max_{\boldsymbol{\theta}} \mathrm{Lik}(\boldsymbol{\theta};\boldsymbol{y})$$

i	y_i
1	T
2	T
3	H
4	T

- p refers to the value of P(coin toss$_i$ = Heads)
- Likelihood for the following dataset



This is choosing the *maximum likelihood estimate* (**MLE**)

The **MLE** also turns out to be the *relative frequency estimate* (RFE)

(repeat slide from lecture 3)

Introductory causal inference

- You have probably had previous exposure to both probability and statistics
- You are less likely to have had exposure to **causal inference**
- Causal inference uses probability and statistics, but it is something separate from the traditional construal of those two fields
- You can think of causal inference as being a framework extending more traditional statistics by:
  - Adding new probability-based mathematical constructs; and,
  - Developing a set of practice for statistical inference based on those constructs
- Two causal inference frameworks:
  - The **potential outcomes** framework
  - The **causal graphical models** framework

The potential-outcomes framework

- In epidemiology and many other areas of statistics, causal inference was developed out of the idea of **potential outcomes** (Neyman 1923, Rubin 1974)
- Consider an outcome, Y, and a potential **treatment** A
- **Example:**

Y: an individual survives to the end of the year (0: no, 1: yes)

A: an individual with heart disease receives a heart transplant (0: no, 1: yes)

Potential-outcome random variables

- Suppose that A is discrete; for this case, $A \in \{0,1\}$
- The **potential outcomes**, or **counterfactual outcomes**, are random variables for Y for each potential value of A

$Y^{a=0}$	The value that Y would take if A were 0
$Y^{a=1}$	The value that Y would take if A were 1

- **Counterfactual risk** is the **expected value** of each counterfactual-outcome random variable:

$$E[Y^{a=0}] \qquad\qquad E[Y^{a=1}]$$

- Expected value, or expectation, is defined as follows:

$$E[X] = \sum_x xP(X = x)$$

- So we are interested in (and likewise for $Y^{a=1}$):

$$E[Y^{a=0}] = \sum_y yP(Y^{a=0} = y) = 0 \times P(Y^{a=0} = 0) + 1 \times P(Y^{a=0} = 1) = \boxed{P(Y^{a=0} = 1)}$$

Counterfactual data and causal effects

- Suppose we knew **what would happen** for each individual in the population under each value of the treatment
- Then we could compute the counterfactual risks:

$$E[Y^{a=0}] = 0.5 \qquad E[Y^{a=1}] = 0.5$$

- The **average causal effect** of treatment A is defined as the difference of counterfactual risks:

$$E[Y^{a=1}] - E[Y^{a=0}] = 0$$

- Here, treatment is **ineffective**

(Hernan & Robins, 2020, Table 1.1)

	$Y^{a=0}$	$Y^{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Leto	0	1
Ares	1	1
Athena	1	1
Hephaestus	0	1
Aphrodite	0	1
Cyclope	0	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0
$P(Y^{a=*}) = 1$	0.5	0.5

Estimating causal effects

Remember, $E[Y^{a=i}] = P(Y^{a=i} = 1)$

	L	A	Y	Y^0	Y^1
Rheia	0	0	0	0	?
Kronos	0	0	1	1	?
Demeter	0	0	0	0	?
Hades	0	0	0	0	?
Hestia	0	1	0	?	0
Poseidon	0	1	0	?	0
Hera	0	1	0	?	0
Zeus	0	1	1	?	1
Artemis	1	0	1	1	?
Apollo	1	0	1	1	?
Leto	1	0	0	0	?
Ares	1	1	1	?	1
Athena	1	1	1	?	1
Hephaestus	1	1	1	?	1
Aphrodite	1	1	1	?	1
Polyphemus	1	1	1	?	1
Persephone	1	1	1	?	1
Hermes	1	1	0	?	0
Hebe	1	1	0	?	0
Dionysus	1	1	0	?	0

- Naively, we might estimate the counterfactual risks $P(Y^{a=i} = 1)$ directly from observed A and Y:

$$\hat{P}_{MLE}(Y = 1 \mid A = 0) = \frac{3}{7} \quad \hat{P}_{MLE}(Y = 1 \mid A = 1) = \frac{7}{13}$$

- But under what circumstances $\hat{P}_{MLE}(Y \mid A = i) = \hat{P}_{MLE}(Y^{a=i} = 1)$?
- The following is certainly true:

$$\hat{P}_{MLE}(Y = 1 \mid A = i) = \frac{\text{Count}(Y = 1 \wedge A = i)}{\text{Count}(A = i)}$$

CONSISTENCY: when $A = i,\ Y = Y^{a=i}$

$$= \frac{\text{Count}(Y^{a=1} = 1 \wedge A = i)}{\text{Count}(A = i)}$$

Crucial step; make sure you understand it!

$$= \hat{P}_{MLE}(Y^{a=i} = 1 \mid A = i)$$

- So, the following condition suffices:

$$P(Y^{a=i} = 1 \mid A = i) = P(Y^{a=i} = 1)$$

- This is called **EXCHANGEABILITY**:

$$Y^a \perp A \mid \{\}$$

Exchangeability and randomization

Goal: $\hat{P}(Y^a = 1)$

- Why is a randomized experiment so powerful?
- Recap of exchangeability criterion:

$$Y^a \perp A \,|\, \{\}$$

- If we ourselves determine A in a way that is *truly blind to* Y^a, it **imposes** exchangeability!
- We can now go ahead and estimate

$$\hat{P}(Y^{a=i} = 1) = \hat{P}(Y = 1 \,|\, A = i)$$

- Hooray!!!

Rheia
Kronos
Demeter
Hades
Hestia
Poseidon
Hera
Zeus
Artemis
Apollo
Leto
Ares
Athena
Hephaestus
Aphrodite
Polyphemus
Persephone
Hermes
Hebe
Dionysus

Does loss of randomization make things hopeless?

- In the real world, many datasets are ***not*** randomized this way
- **Example:** let's imagine some other variable that might affect whether treatment A is applied; e.g., $L =$ whether the patient was in critical condition (1=yes, 0=no)
- In general, L will be related to Y^a
  - E.g., in this example, patients in critical condition are surely more likely to die overall!

$$\cancel{A \perp Y^a \mid \{\}}$$

	L
Rheia	0
Kronos	0
Demeter	0
Hades	0
Hestia	0
Poseidon	0
Hera	0
Zeus	0
Artemis	1
Apollo	1
Leto	1
Ares	1
Athena	1
Hephaestus	1
Aphrodite	1
Polyphemus	1
Persephone	1
Hermes	1
Hebe	1
Dionysus	1