Lecture 1: Review and Background

Professor Alexander Franks

2020-10-07

Logistics

- First homework is out, due October 18 at 11:59pm
- Lab begins this week, Tuesday/Wednesday
- Try pstat115.lsit.ucsb.edu
  - Cloud based rstudio service
  - Log in with your UCSB NetID

Resources

Look at the resources folder in cloud for

- A fantastic probabilility review sheet
- Probability density information
- Hoff textbook

Rstudio in the cloud

- Please post on piazza if you notice and issues
- After several hours of inactivity you will be logged out automatically (save your work if you are going to stop working for a while)
- All packages required for this course are pre-installed. If there is a package you like to use that is not installed let us know (on piazza)

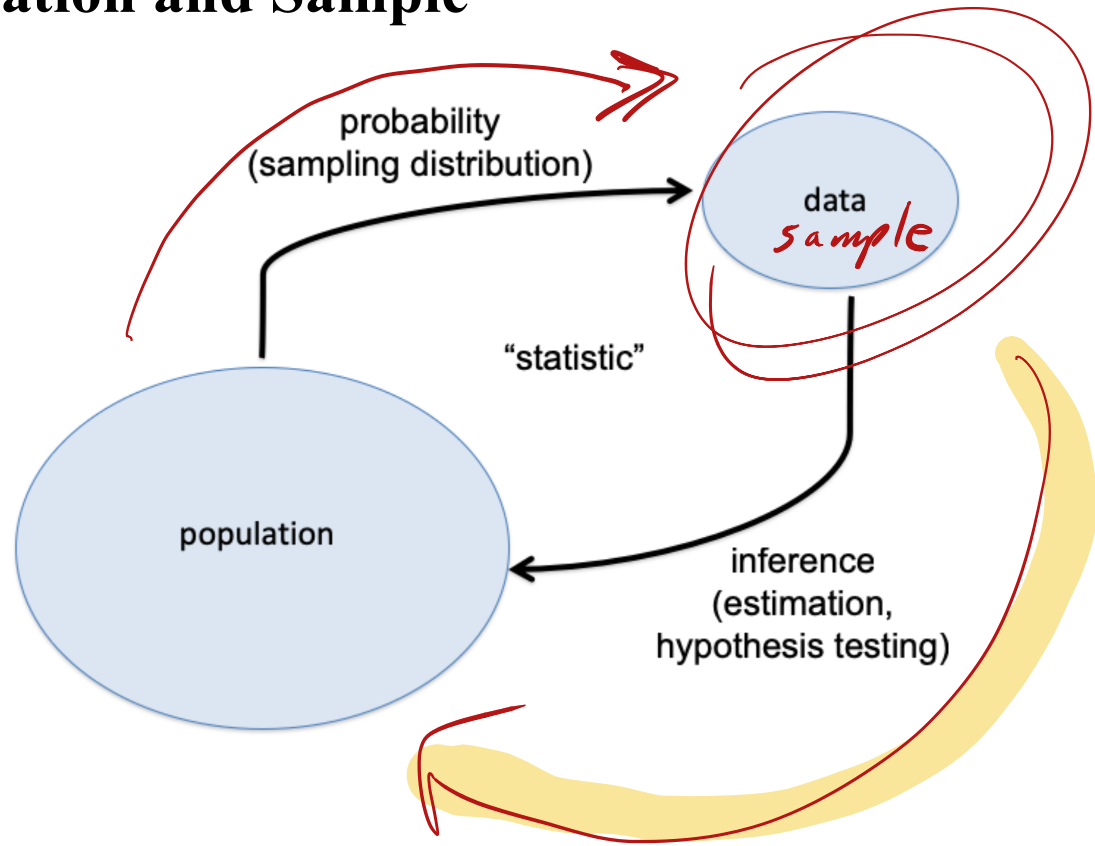
Homework 1 is out

- Use this link to pull the assignment into your environment https://bit.ly/3kZ2sVr
- Link is pinned to piazza page. Will be used to sync all assignments.
- File will be in `fall20/homework1` (homework1.Rmd)
- Knit file to create pdf
- Do not change the name of the file or the directory
- Autograding
  - Leave code cells that look like `. = ottr::check("tests/q1a.R")`

Staff and Office Hours

- Prof. Franks: Wed. 2pm
- TAs:
  - Dorothy Li: Wednesday 5-6pm, Thursday 4-5pm
  - Xubo Liu: Thursday 5-6pm, Friday 1-2pm
- ULA:
  - Matthew Coleman: Thursday: 11am-12pm, Friday: 10am-12pm

Population and Sample



Population and Sample

- The *population* is the group or set of items relevant to your question
  - Usually very large (often conceptualize a population as infinite)
- Sample: a finite subset of the population

Modeling

  - How is the sampling collected (representative?)
  - Denote the sample size with n

Population and Sample

- Our goal is (usually) to learn about the population from the sample
  - Population parameters encode relevant quantities
  - The **estimand** is the thing we what to infer and is usually a function of the population parameters

Random variables

- A random variable, Y, has variability, can take on several different values (possibly infinitely may), and is associated with a distribution.
    - The distribution determines the probability that the r.v. will take a specific value.
- Notation:
    - Y (uppercase) denotes a random variable
    - y (lowercase) is a *realization* of that random variable and is not random

$$Y \sim \mathrm{Bin}(\underset{\substack{\| \\ 10}}{n}, \underset{\substack{\| \\ 0.5}}{\theta})$$

$y = 5$ (got 5 heads)

Constants

- Constants: quantities with 0 variance. data
  - Constants can be *known* (e.g. observed data)
  - Constants can be *unknown* (not observed)

θ, the weight of the coin

estimand

(in frequentist thinking)

Setup

- The *sample space* $\mathcal{Y}$ is the set of all possible datasets we could observe. We observe *one* dataset, y, from which we hope to learn about the world.

  5 heads

- The *parameter space* Θ is the set of all possible parameter values θ

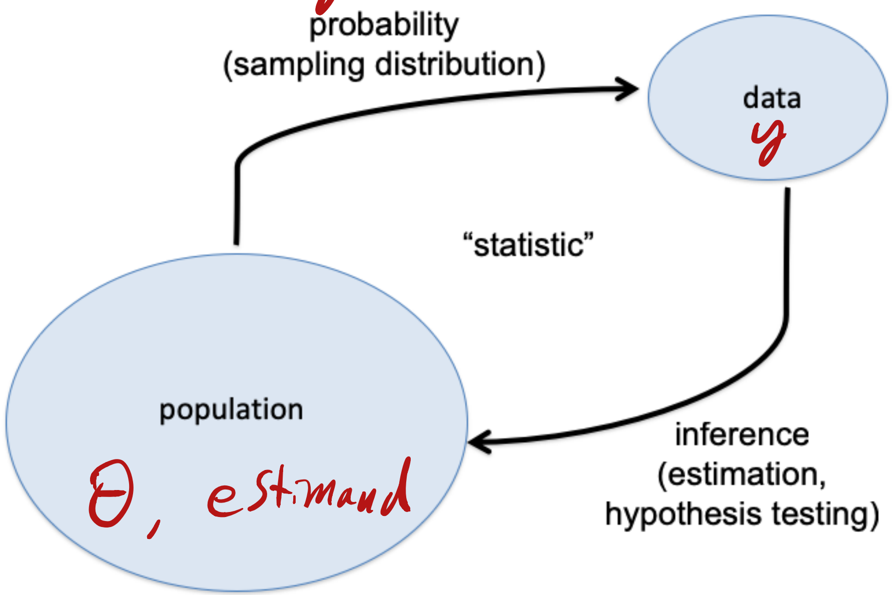
  e.g. $\theta \in [0, 1]$ (any probability)

- θ encodes the population characteristics that we want to learn about

- Our *sampling model* $p(y \mid \theta)$ describes our belief about what data we are likely to observe for a given value of θ.

Notation

$P(y|\theta)$



(observed data)

"hat"
estimate/
estimator

$\hat{\theta}(y)$ estimate

depends on data

The Likelihood Function

- The likelihood is the "probability of the observed data" expressed as a function of the unknown parameter:
- A function of the unknown constant θ.
- Depends on the observed data $y = (y_1, y_2, \ldots, y_n)$

$$L(\theta) = P(\underbrace{y}_{\text{known, data}} \mid \overbrace{\theta}^{\text{unknown}})$$

Independent Random Variables

- $Y_1, \ldots, Y_n$ are random variables

# of trials

- We say that $Y_1, \ldots, Y_n$ are *conditionally* independent given θ if
- Conditional independence means that Y_i gives no additional information about Y_j beyond that in knowing θ

$$P(Y_1, Y_2, \ldots Y_n \mid \theta) = \prod_{i=1}^{n} P(Y_i \mid \theta)$$

$$\overset{?}{\Downarrow}$$

$$? \left[P(Y_1, Y_2, \ldots Y_n) = \prod_{i=1}^{n} P(Y_i) \right] ?$$

$$\overset{?}{\Downarrow}$$

Conditional Indep $\not\Longrightarrow$ Indep.

Example: A binomial model

- Assume I go to the basketball court and takes 5 free throw shots
- Model the number of made shots I make using a $\text{Bin}(5, \theta)$ → Each shot is independent
    - What are the key assumptions that make these a reasonable emodel?
- θ represents my true skill (the fraction of shots I make) → If I were to take ∞ shots.
- How can we estimate my true skill?

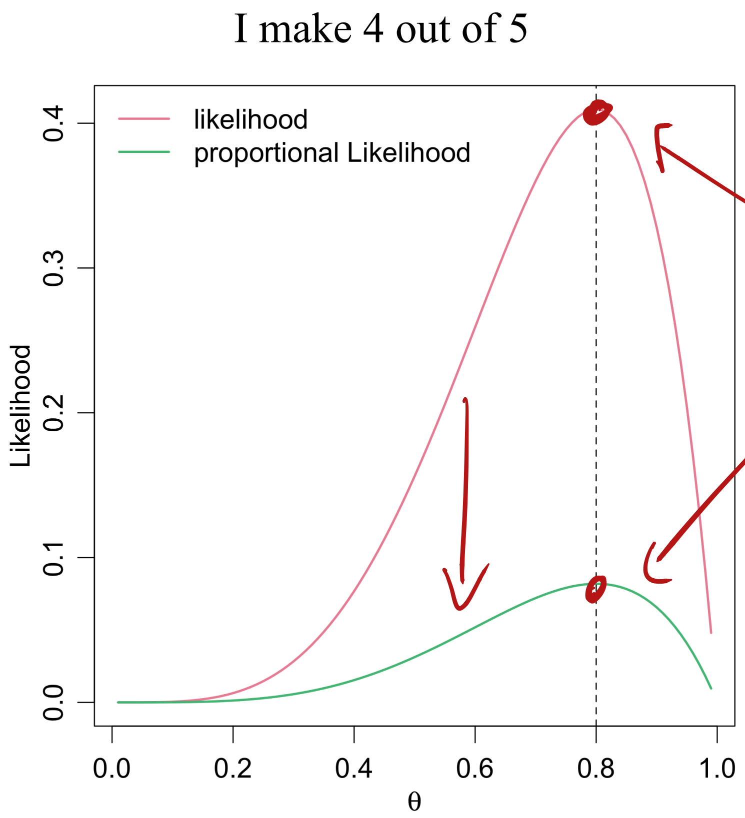
Likelihood:

$$L(\theta) = P(y|\theta) = \overbrace{\binom{5}{y}}^{\text{const. in } \theta} \theta^{y} (1-\theta)^{5-y}$$

$y = 4$ shots (plug-in)

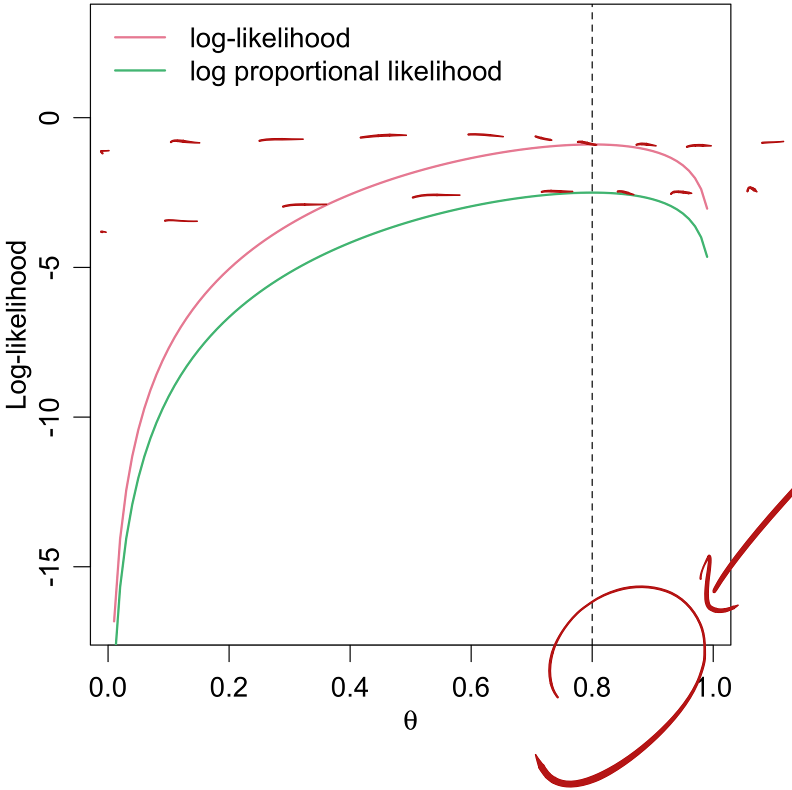
$$\propto \theta^{y} (1-\theta)^{5-y}$$

The binomial likelihood



Max. doesn't change.

The log-likelihood



Doesn't change where the max is.

Maximum Likelihood Estimation

- The *maximum likelihood estimate* (MLE) is the value of θ that makes the data the most "likely", that is, the value that maximizes $L(\theta)$

- To compute the maximum likelihood estimate:

  1. Write down the likelihood and take its log:

  $$\log(L(\theta)) = \ell(\theta)$$

  script for log.

  2. Take the derivative of $\ell(\theta)$ with respect to θ:

  $$\ell'(\theta) = \frac{d\ell(\theta)}{d\theta}$$

  3. Solve for $\hat{\theta}$ such that $\ell'(\theta) = 0$

Maximum Likelihood Estimation

1. $L(\theta) \propto \theta^{y}(1-\theta)^{n-y}$

$$\log(L(\theta)) = \ell(\theta) = y \log \theta + (n-y)\log(1-\theta)$$

2. $\dfrac{d\ell(\theta)}{d\theta} = \dfrac{y}{\theta} - \dfrac{n-y}{1-\theta} = 0$

$$\frac{y}{\theta} = \frac{n-y}{1-\theta} \rightarrow (n-y)\theta = y(1-\theta)$$

$$\hat{\theta}_{MLE} = \frac{y}{n}$$

$$\log(a^b) = b \log a$$

$$\log(ab) = \log(a) + \log(b)$$

Example: Binomial

- Assume we are polling the presidential race in the upcoming election
- We poll 25 random students in the class $Y_1, \ldots Y_n$ from $n = 25$
- Y_i is either 0 (Trump) or 1 (Biden)
- $Y_i \sim \text{Bern}(\theta)$, where $\text{Bern}(\theta)$ is equivalent to Bin(1, θ)
  - Bernoulli random variables is a binomial with one trial
  - Assume our class is a simple random sample of the population
- How do we estimate θ for multiple observations?

$$L(\theta) = P(y_1, y_2, \ldots y_{25} | \theta)$$

$$= \prod_{i=1}^{25} P(y_i | \theta)$$

$$= \prod_{i=1}^{25} \binom{1}{y_i} \theta^{y_i} (1-\theta)^{1-y_i}$$

Example: the likelihood for independent Bernoulli's

$$
\begin{aligned}
p(y_1, y_2, \ldots, y_n|1, \theta) &= p(y_1, y_2, \ldots, y_n|\theta) \\
&= p(y_1|\theta)p(y_2|\theta)\ldots p(y_n|\theta) \\
&= \prod_{i=1}^{n} p(y_i|\theta) \\
&= \prod_{i=1}^{n} \binom{1}{y_i} \theta^{y_i}(1-\theta)^{(1-y_i)} \\
&= \left[\cancel{\prod_{i=1}^{n} \binom{1}{y_i}}\right] \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i} \\
&= L(\theta)
\end{aligned}
$$

$$\hat{\theta}_{MLE} = \frac{\sum y_i}{n} = \bar{y}$$

$$\propto \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}$$

n indep. Berns(θ) is same as one Bin(n, θ)

Sufficient Statistics

- Let $L(\theta) = p(y_1, \ldots y_n \mid \theta)$ be the likelihood and $\underbrace{s(y_1, \ldots y_n)}_{\text{function of the data}}$ be a statistic

- $s(y)$ is a sufficient statistic if we can write:

$$L(\theta) = h(y_1, \ldots y_n) g(s(y), \theta)$$

$$\binom{5}{y} \underbrace{\theta^{y}(1-\theta)^{n-y}}$$

  - g is only a function of s(y) and θ only
  - h is *not* a function of θ

- This is known as the *factorization theorem*

- $L(\theta) \propto g(s(y), \theta)$

AH view

$z_1, \ldots z_5$ = make/miss each shot.

$s(z_1, \ldots z_5) = \sum z_i$

$L(\theta) \propto \theta^{\sum z_i} (1-\theta)^{n - \sum z_i}$

$$\prod_{i=1}^{5} \theta^{z_i} (1-\theta)^{1-z_i} \rightarrow \theta^{\sum z_i} (1-\theta)^{\sum(1-z_i)}$$

$$\prod_i a^{b_i} = a^{\sum b_i}$$

Sufficient Statistics

- Intuition: a sufficient statistic contains all of the information about θ
  - Many possible sufficient statistics
  - Often seek a statistic of the lowest possible dimension (minimal sufficient statistic)
  - What are some sufficient statistics in the previous binomial example?

$$\boxed{\frac{\sum z_i}{n}, \quad \sum z_i, \quad (z_1, \ldots, z_5)}$$

$$L(\theta) = h(\ldots)\, g(s, \theta)$$

Is z_1 sufficient?

Estimators and Estimates

- In classical (frequentist) statistics, θ is an unknown constant
- An **estimator** of a parameter θ is a function of the random variables, Y
  - E.g. for Binomial$(1, \theta)$: $\hat{\theta}(Y) = \frac{\sum_i Y_i}{n}$ capital
  - An estimator is a random variable
  - Interested in properties of estimators (e.g. mean and variance)

Estimator: $\frac{\sum Y_i}{n}$ (random)

Estimate: $\frac{\sum y_i}{n}$ (constant)

Estimators and Estimates

- $\hat{\theta}(y)$ as a function of realized data is called an **estimate**
  - Plug in observed data $y = (y_1, \dots y_n)$ to estimate θ
  - An estimate is a non-random constant (it is has 0 variability)
  - E.g. in the binomial(1, θ), $\hat{\theta} = \bar{y} = \frac{\sum_i y_i}{n}$ is the maximum likelihood estimate for the binomial proportion.

lowercase

Bias and Variance

- Estimators are random variables. What are some r.v. properties that are desirable?

(asymptotically)

– Unbiased

Accurate

– Small Variance

– Sufficient?

– ** Consistent **