Week 2: Review and Background

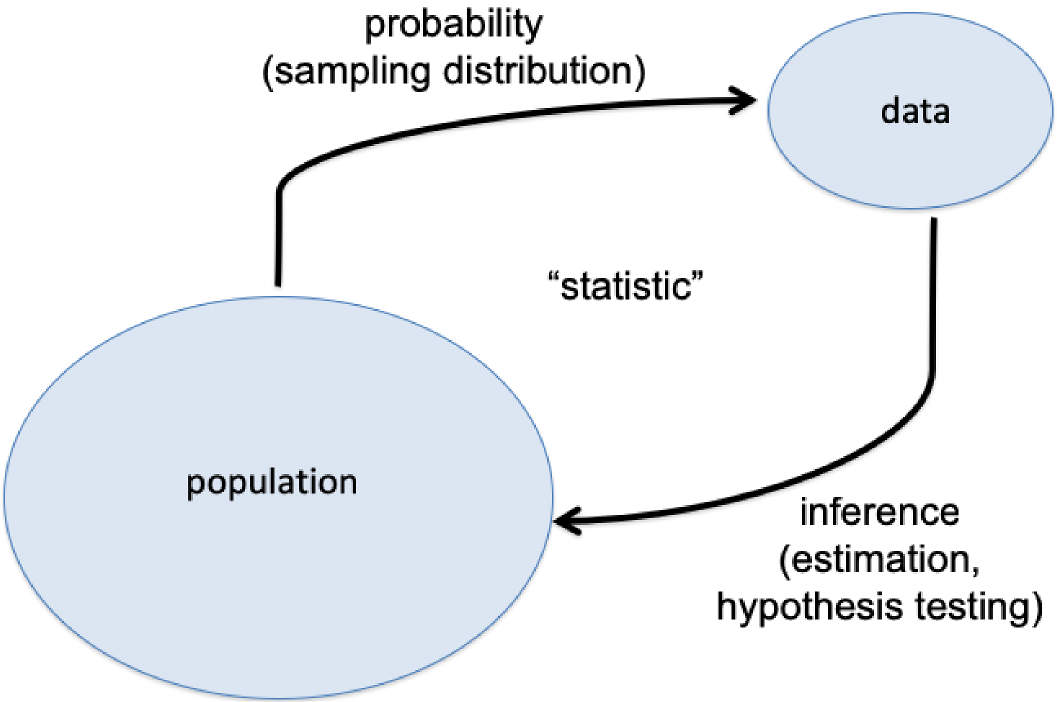
Professor Rodrigo Targino

Resources

Look at the resources folder in cloud for

- A fantastic probability review sheet
- Probability density information
- Bayes Rules: Chapters 1 and 2

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
    - 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

Constants

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

Setup

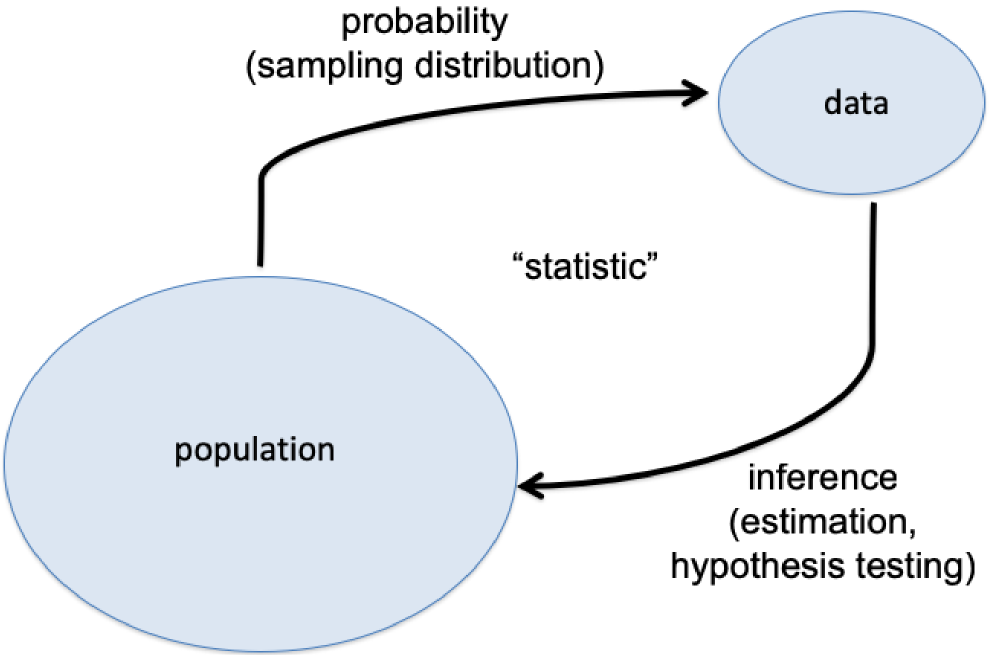
- The *sample space* $\mathcal{Y}$ is the set of all possible datasets we could observe. We observe *one* dataset, y, from which we hope to learn about the world.

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

- 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


probability
(sampling distribution)
data
"statistic"
population
inference
(estimation,
hypothesis testing)

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 $y = (y_1, y_2, \ldots, y_n)$

Independent Random Variables

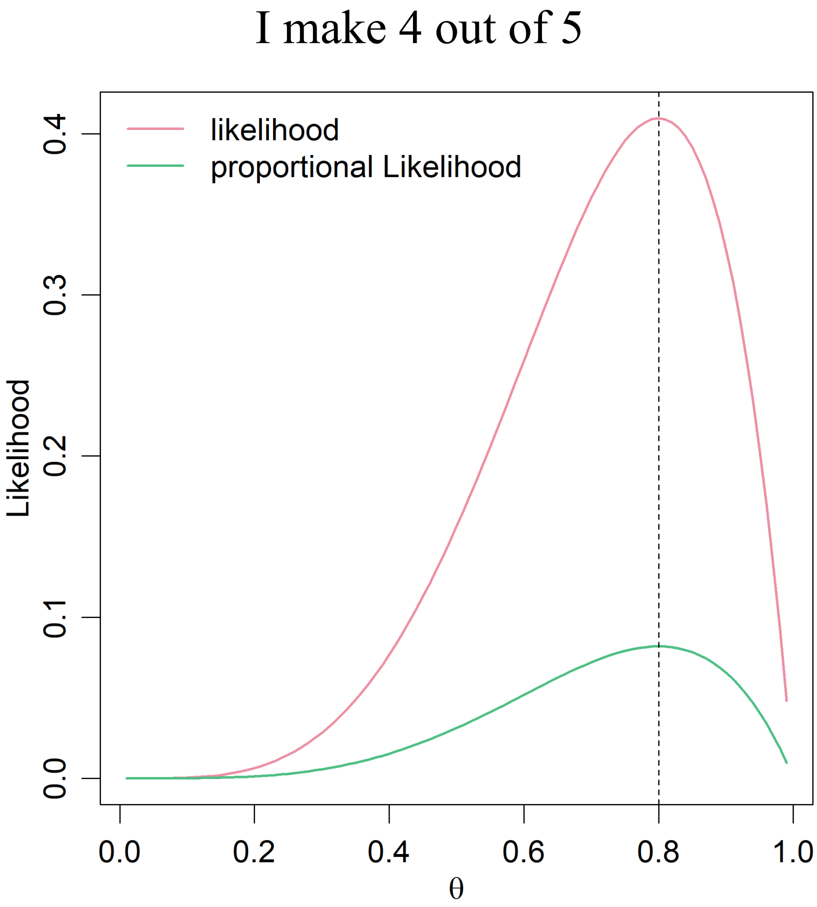
- $Y_1, \ldots, Y_n$ are random variables
- We say that $Y_1, \ldots, Y_n$ are *conditionally* independent given θ if ...
- Conditional independence means that Y_i gives no additional information about Y_j beyond that in knowing θ

Example: A binomial model

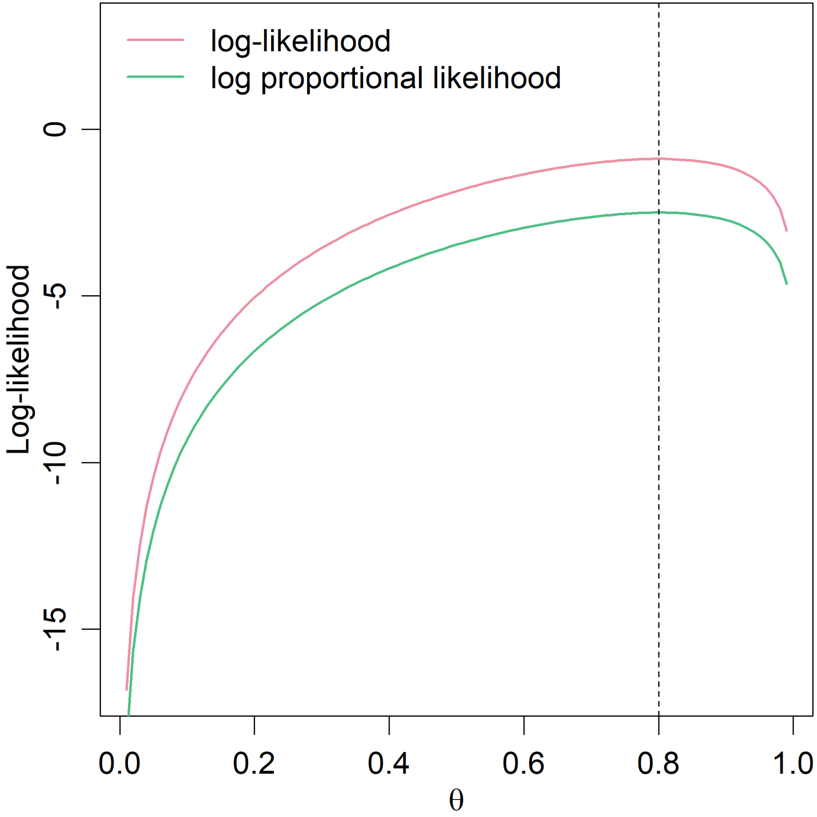
- Assume I go to the basketball court and takes 5 free throw shots
- Model the number of made shots I make using a $\mathrm{Bin}(5, \theta)$
  - What are the key assumptions that make this a reasonable model?
- θ represents my true skill (the fraction of shots I make)
- How can we estimate my true skill?

Likelihood:

The binomial likelihood



The log-likelihood



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:

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

  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

Example: Binomial

- Assume we are polling the presidential race in the next election
- We poll 25 random students in the class $Y_1, \ldots Y_n$ from $n = 25$
- Y_i is either 0 (Republican) or 1 (Democrat)
- $Y_i \sim \text{Bern}(\theta)$, 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?

Example: the likelihood for independent Bernoulli's

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

Sufficient Statistics

- Let $L(\theta) = p(y_1, \dots y_n \mid \theta)$ be the likelihood and $s(y_1, \dots y_n)$ be a statistic

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

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

  - 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* (Fisher–Neyman)

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

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?

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 $\mathrm{Binomial}(1, \theta)$: $\hat{\theta}(Y) = \frac{\sum_i Y_i}{n}$
  - An estimator is a random variable
  - Interested in properties of estimators (e.g. mean and variance)

Estimators and Estimates

- $\hat{\theta}(y)$ as a function of realized data is called an **estimate**
  - Plug in observed data $y = (y_1, \ldots y_n)$ 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.

Bias and Variance

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

Bias and Variance

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

- Bias: $E[\hat{\theta}] - \theta = 0$

  - $E[\hat{\theta}] - \theta = 0$ means the estimator is unbiased

  - E.g. expectation of the binomial MLE: $E[\hat{\theta}] = E[\frac{Y}{n}] = \theta$

- $\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$

E.g. variance of the binomial MLE is

$$\text{Var}[\hat{\theta}] = \text{Var}(\frac{Y}{n}) = \frac{\theta(1-\theta)}{n}$$

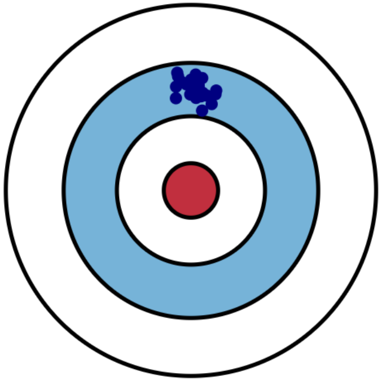
Bias and Variance

- Want estimators that have low bias and variance because this implies low overall error

- Mean squared error equals bias^2 + variance

Bias

The average difference between the prediction and the response

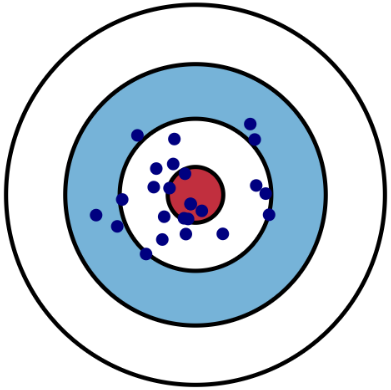


Statistical definition of bias:

$$E[\hat{\theta} - \theta]$$

Variance

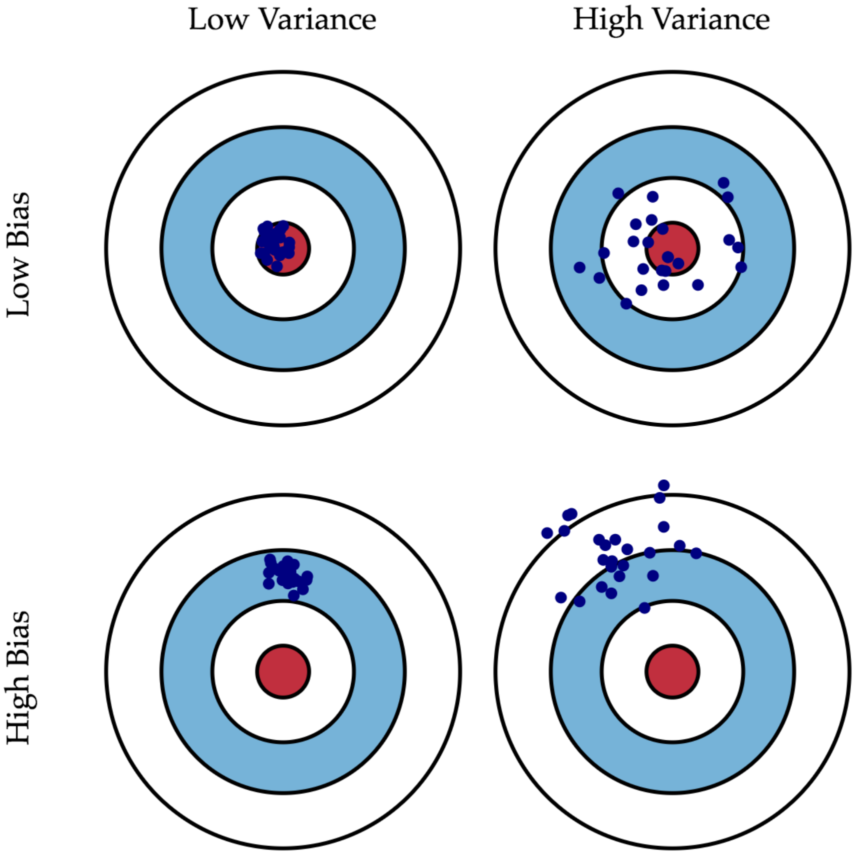
How variable is the prediction about its mean?



Statistical definition of variance:

$$E[\hat{\theta} - E[\hat{\theta}]]^2$$

Bias and Variance



Maximum Likelihood Estimators

Under relatively weak conditions:

- The MLE is *consistent*. It converges to the true value as the sample size goes to infinity.
  - Need bias and variance to go to 0 as sample size increases
- The MLE is *asymptotically optimal*. For "large" sample sizes is has the lowest variance.
- *Equivariance*: if $\hat{\theta}$ is the MLE for θ then $g(\hat{\theta})$ is the MLE for $g(\theta)$

Confidence Interval Simulations

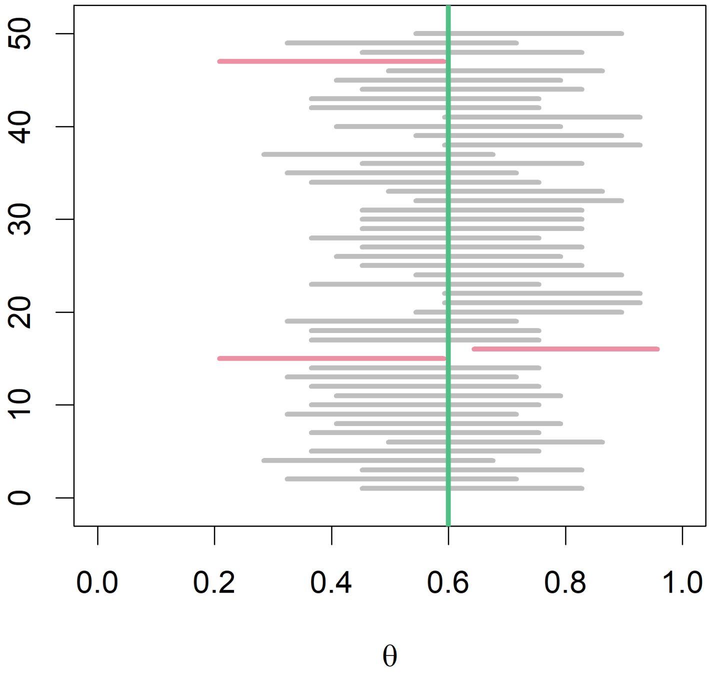
Let's do 50 hypothetical replications to illustrate confidence intervals

```
for i in 1 to 50:
    - Draw Y_i from Bin(25, 0.6)
    - Compute and plot the 95% confidence interval
```

- Will have 50 confidence intervals based on 50 simulated datasets.
- A 95% interval means that on average 95% of these 50 intervals will cover the true value

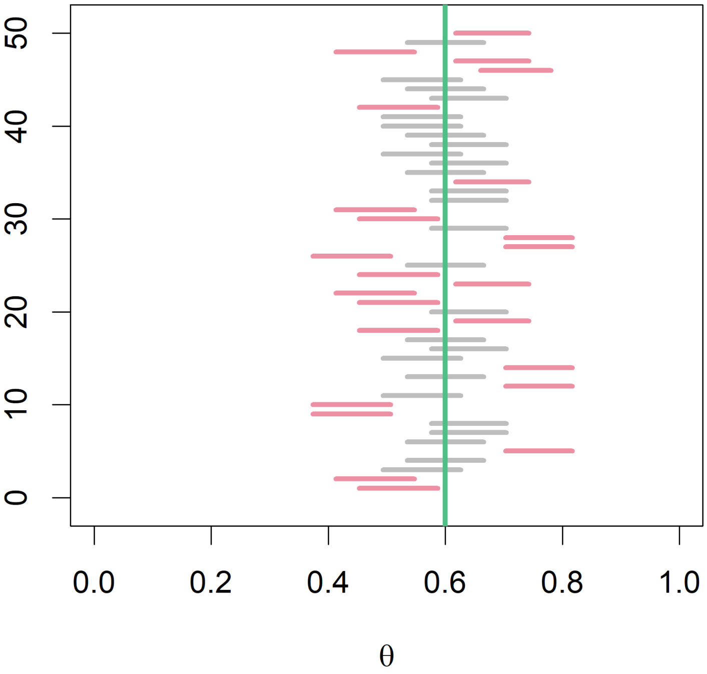
95% Confidence Intervals

In truth, 60% of the population will vote for "candidate 1"



We expect $0.05 \times 50 = 2.5$ of the intervals to *not* cover the true parameter, $p = 0.6$, on average

50% Confidence Intervals



We expect $0.50 \times 50 = 25$ of the intervals to *not* cover the true parameter, 0.6

Data Generating Process (DGP)

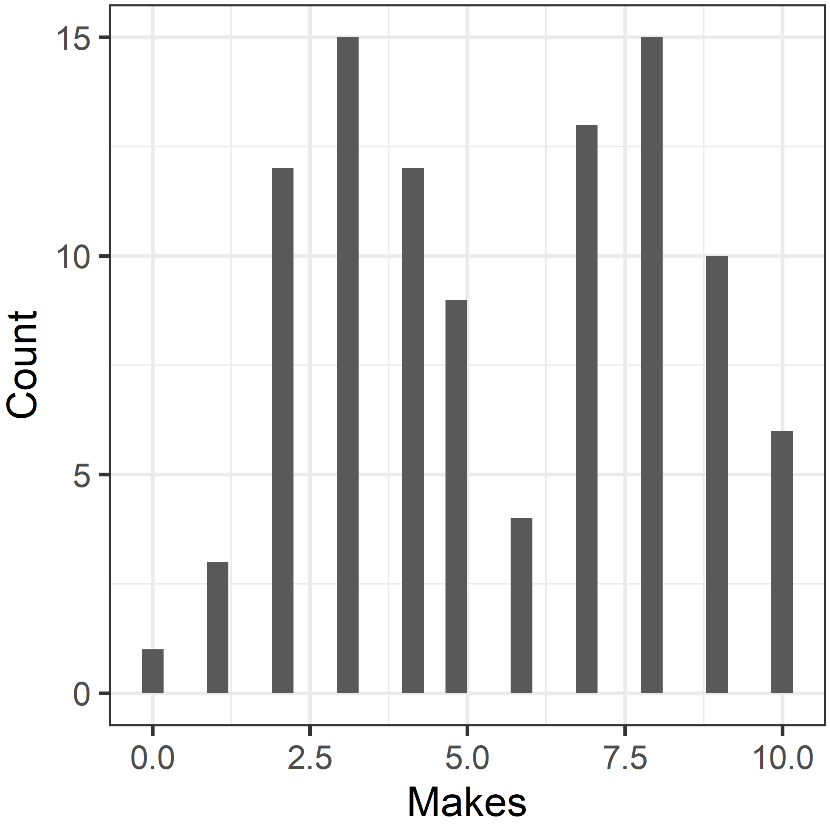
- DGP: a statistical model for how the observed data might have been generated
- Often write the DGP using pseudo-code:

```
for (i in 1:N)
  - Generate y_i from a Normal(0, 1)
return y = (y_1, ... y_N)
```

- The DGP should tell a story about how the data came to be
- Can translate the DGP into a statistical model

Data Generating Process (DGP)

Assume everybody in this class goes to a basketball court and takes 10 free throw shots:



Data Generating Process (DGP)

Tell a plausible story: some students play basketball and some don't. Before you take your shots we record whether or not you have played before.

```
assume theta_1 > theta_0
for (i in 1:100)
  - Generate z_i from Bin(1, phi)
    - p_i = theta_0 if z_i=0
    - p_i = theta_1 if z_i=1
  - Generate y_i from a Binom(10, p_i)
return y = (y_1, ... y_100) and z = (z_1, ..., z_100)
```

Is this a reasonable model?

Mixture Models

$$Z_i = \begin{cases} 0 & \text{if the } i^{th} \text{ if student doesn't play basketball} \\ 1 & \text{if the } i^{th} \text{ if student does play basketball} \end{cases}$$

$$Z_i \sim Bin(1, \phi)$$

$$Y_i \sim \begin{cases} \text{Bin}(10, \theta_0) & \text{if } Z_i = 0 \\ \text{Bin}(10, \theta_1) & \text{if } Z_i = 1 \end{cases}$$

Mixture Models

$$Z_i = \begin{cases} 0 & \text{if the } i^{th} \text{ if student doesn't play basketball} \\ 1 & \text{if the } i^{th} \text{ if student does play basketball} \end{cases}$$

$$Z_i \sim Bin(1, \phi)$$

$$Y_i \sim \begin{cases} \text{Bin}(10, \theta_0) & \text{if } Z_i = 0 \\ \text{Bin}(10, \theta_1) & \text{if } Z_i = 1 \end{cases}$$

- ϕ is the fraction of students that have experience playing basketball
- θ_1 is the probability of making a shot for an experienced player
- θ_0 is the probability of making a shot for an inexperienced player

Table of relevant quantities

- Can be a fixed constant (no variability) or a random variable (has variability)
- Can be observed (known) or unobserved (unknown)
- Helpful for to keep track of all of the relevant quantities

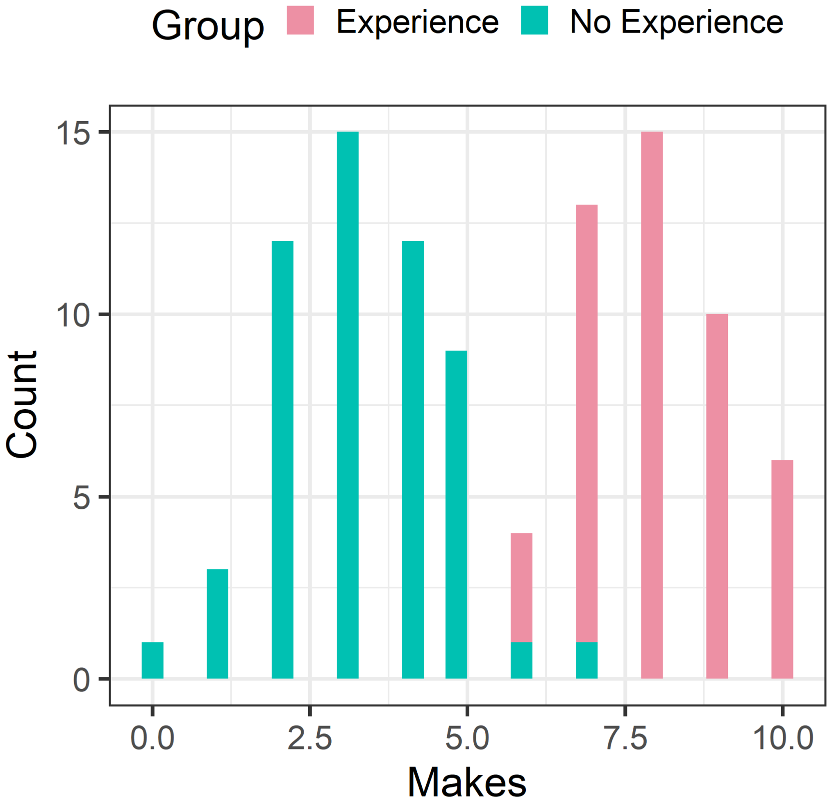
Mixture models

- A mixture model is a probabilistic model for representing the presence of subpopulations
- The subpopoluation to which each individual belongs is not necessarily known
    - e.g. do we ask: "have you played basketball before?"

Mixture models

- A mixture model is a probabilistic model for representing the presence of subpopulations
- The subpopoluation to which each individual belongs is not necessarily known
    - e.g. do we ask: "have you played basketball before?"
- When z_i is not observed, we sometimes refer to it as a clustering model
    - *unsupervised* learning

A Mixture Model



Note: z is observed

Mixture Model Likelihood

Z is observed

Sufficient statistics When Z_i is observed

Together, the following quantities are sufficient for $(\theta_0, \theta_1, \phi)$

- $\sum y_i z_i$ (total number of shots made by experienced players)
- $\sum y_i (1 - z_i)$ (total number of shots made by inexperienced players)
- $\sum z_i$ (total number experienced players)

Data Generating Process (DGP)

```
for (i in 1:100)
  - Generate z_i from Bin(1, phi)
    - p_i = theta_1 if z_i=1
    - p_i = theta_0 if z_i=0
  - Generate y_i from a Binom(10, p_i)
return y = (y_1, ... y_100)
```

This time we don't record who has experience with basketball.

A Mixture Model

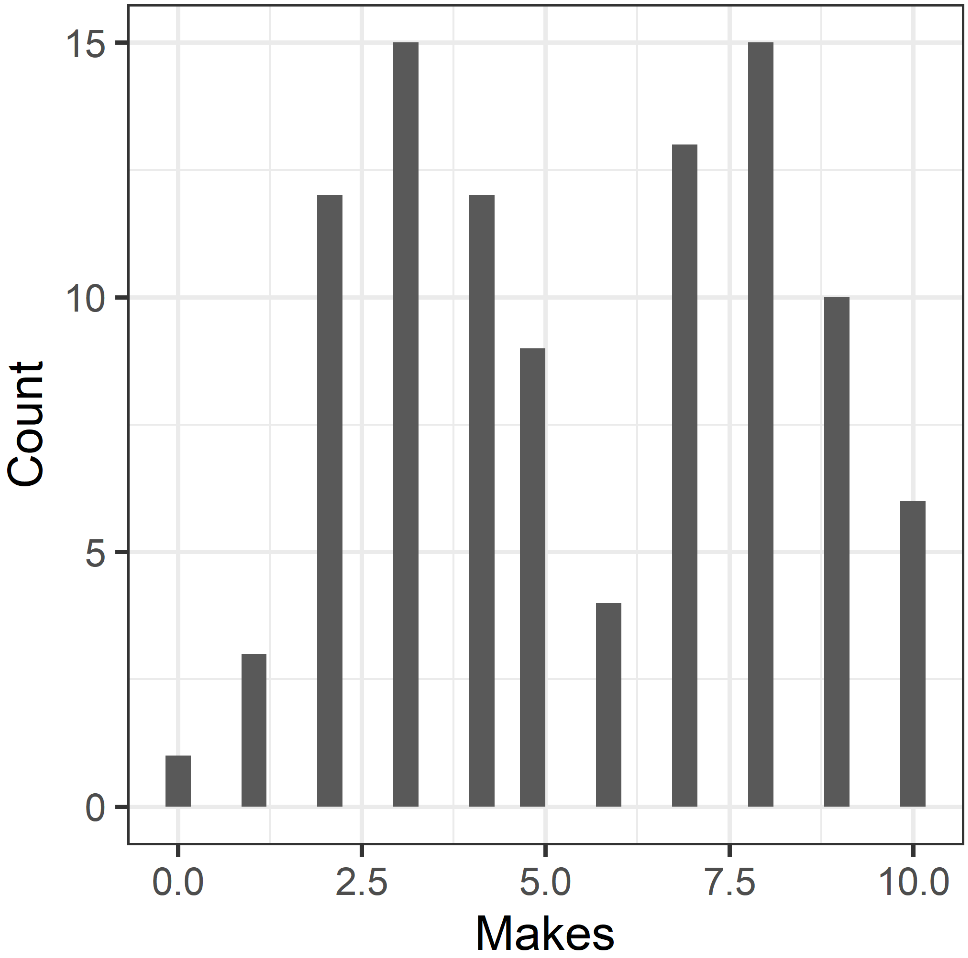


Table of Relevant Quantities

A finite mixture model

- Even if we don't observe Z, it's often useful to introduce it as a *latent* variable

- Write the *observed data likelihood* by integrating out the latent variables from the *complete data likelihood*

$$\begin{aligned} p(Y \mid \theta) &= \sum_z p(Y, Z = z \mid \theta) \\ &= \sum_z p(Y \mid Z = z, \theta) p(Z = z \mid \theta) \end{aligned}$$

In general we can write a K component mixture model as:

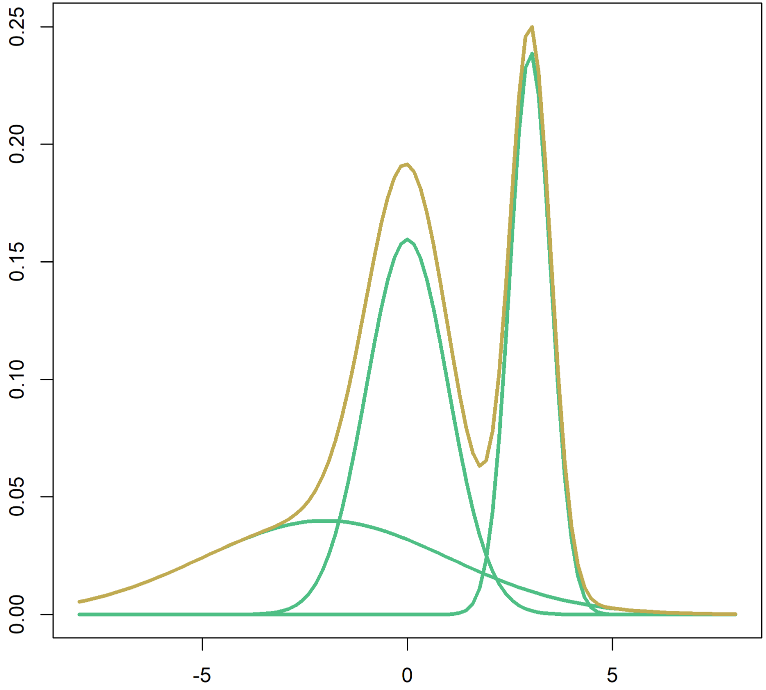
$$p(Y) = \sum_k^K \pi_k p_k(Y)$$

with $\sum \pi_k = 1$

Mixture Model Likelihood

Z unobserved

Finite Mixture models



Infinite Mixture Models

- In the previous example the latent variable had finitely many outcomes
- Latent varibles can have infinitely many outcomes in which case we have any infinite mixture
- Example:

$$\mu \sim N(0, \tau^2)$$
$$Y \sim N(\mu, \sigma^2)$$

$$p(Y \mid \sigma^2, \tau^2) = \int p(Y, \mu \mid \sigma^2, \tau^2) d\mu$$

What is the *marginal* distribution of Y?

Bayesian Inference

- In frequentist inference, θ is treated as a fixed unknown constant
- In Bayesian inference, θ is treated as a random variable
- Need to specify a model for the joint distribution $p(y, \theta) = p(y \mid \theta)p(\theta)$

Bayesian Inference in a Nutshell

1. The *prior distribution* $p(\theta)$ describes our belief about the true population characteristics, for each value of $\theta \in \Theta$.

Bayesian Inference in a Nutshell

1. The *prior distribution* $p(\theta)$ describes our belief about the true population characteristics, for each value of $\theta \in \Theta$.

2. Our *sampling model* $p(y \mid \theta)$ describes our belief about what data we are likely to observe if θ is true.

Bayesian Inference in a Nutshell

1. The *prior distribution* $p(\theta)$ describes our belief about the true population characteristics, for each value of $\theta \in \Theta$.

2. Our *sampling model* $p(y \mid \theta)$ describes our belief about what data we are likely to observe if θ is true.

3. Once we actually observe data, y, we update our beliefs about θ by computing *the posterior distribution* $p(\theta \mid y)$. We do this with Bayes' rule!

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A) PAB)}{P(B)}$$

- $P(A \mid B)$ is the conditional probability of A given B
- $P(B \mid A)$ is the conditional probability of B given A
- $P(A)$ and $P(B)$ are called the marginal probability of A and B (unconditional)

Bayes' Rule for Bayesian Statistics

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{P(y)}$$

- $P(\theta \mid y)$ is the posterior distribution
- $P(y \mid \theta)$ is the likelihood
- $P(\theta)$ is the prior distribution
- $P(y) = \int_\Theta p(y \mid \tilde{\theta})p(\tilde{\theta})d\tilde{\theta}$ is the model evidence

Bayes' Rule for Bayesian Statistics

$$\begin{aligned} P(\theta \mid y) &= \frac{P(y \mid \theta)P(\theta)}{P(y)} \\ &\propto P(y \mid \theta)P(\theta) \end{aligned}$$

- Start with a subjective belief (prior)
- Update it with evidence from data (likelihood)
- Summarize what you learn (posterior)

The posterior is proportional to the likelihood times the prior!

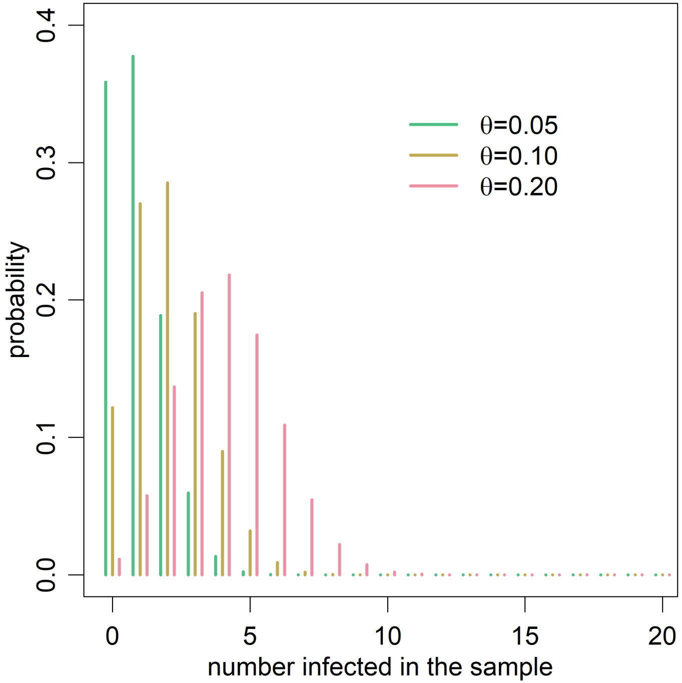
Example: Estimating COVID Infection Rates

- We need to estimate the prevalence of a COVID in Isla Vista
- Get a small random sample of 20 individuals to check for infection

Example: Estimating Infection Rates

- θ represents the population fraction of infected
- Y is a random variable reflecting the number of infected in the sample
- $\Theta = [0, 1] \quad \mathcal{Y} = \{0, 1, \ldots, 20\}$
- Sampling model: $Y \sim \text{Binom}(20, \theta)$

Example: Estimating Infection Rates



Example: Estimating Infection Rates

- Assume *a priori* that the population rate is low
  - The infection rate in comparable cities ranges from about 0.05 to 0.20
- Assume we observe $Y = 0$ infected in our sample
- What is our estimate of the true population fraction of infected individuals?

Example: Estimating Infection Rates

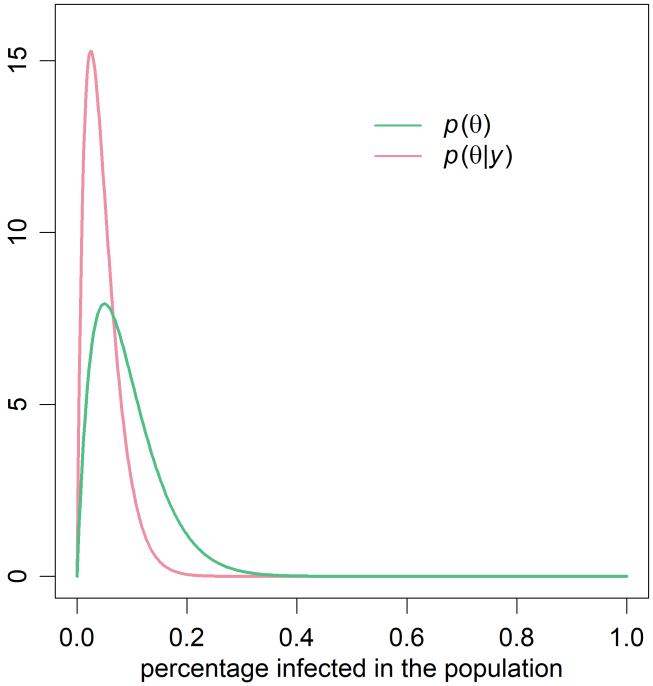


Table of Relevant Quantities

Summary

- Likelihood, log likehood in MLE
- Confidence intervals (how they are defined in frequentist inference)
- Sufficient statistics
- Mixture models

Summary

- In frequentist inference, unknown parameters treated as constants
    - Estimators are random (due to sampling variability)
    - Asks: "how would my results change if I repeated the experiment?"

Summary

- In Bayesian inference, unknown parameters are random variables.
    - Need to specify a prior distribution for θ (not easy)
    - Asks: "what do I *believe* are plausible values for the unknown parameters?"
    - Who cares what might have happened, focus on what *did* happen!

Assignments

- Read chapters 1 and 2 of BR