Lecture 1: Likelihood Review

Professor Alexander Franks

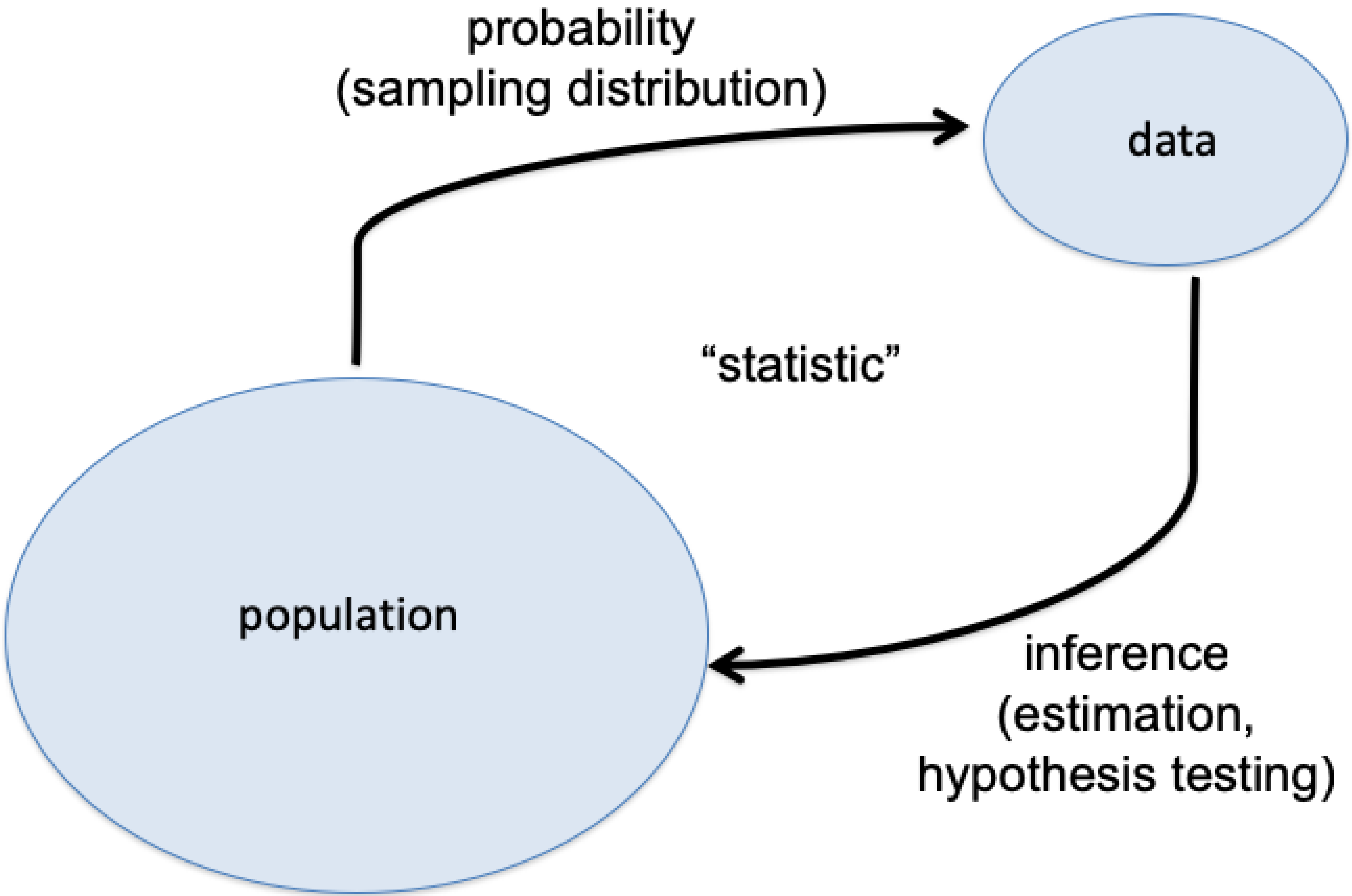
Logistics

- Read: BDA Chapters 1-2

Logistics

- Use this link to pull all course content into your environment
- Link is on the course website. Will be used to sync all assignments.

Population and Sample



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 $P(y_1, \ldots, y_n \mid \theta) = \prod_i P(y_i \mid \theta)$
- Conditional independence means that Y_i gives no additional information about Y_j beyond that in knowing θ

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 $y = (y_1, y_2, \ldots, y_n)$
- Two likelihood functions are equivalent if one is a scalar multiple of the other

Sufficient Statistics (Frequentist Definition)

A statistic s(Y) is sufficient for underlying parameter θ if the conditional probability distribution of the Y, given the statistic s(Y), does not depend on θ.

Sufficient Statistics

- Let $L(\theta) = p(y_1, \ldots y_n \mid \theta)$ be the likelihood and $s(y_1, \ldots y_n)$ be a statistic
- *Factorization theorem*: $s(y)$ is a sufficient statistic if we can write:

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

  - g is only a function of s(y) and θ only
  - h is *not* a function of θ
- $L(\theta) \propto g(s(y), \theta)$

The Likelihood Principle

- **The likelihood principle**: All information from the data that is relevant to inferences about the value of the model parameters is in the equivalence class to which the likelihood function belongs
- Two likelihood functions are equivalent if one is a scalar multiple of the other
- Frequentist testing and some design based estimators violate the likelihood principle

Binomial vs Negative Binomial

Data Generating Process (DGP)

- I select 100 random students at UCSB to 10 free throw shots at the basketball court
- Assume there are two groups: experienced and inexperienced players
- Skill is identical conditional on experience level

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)
```

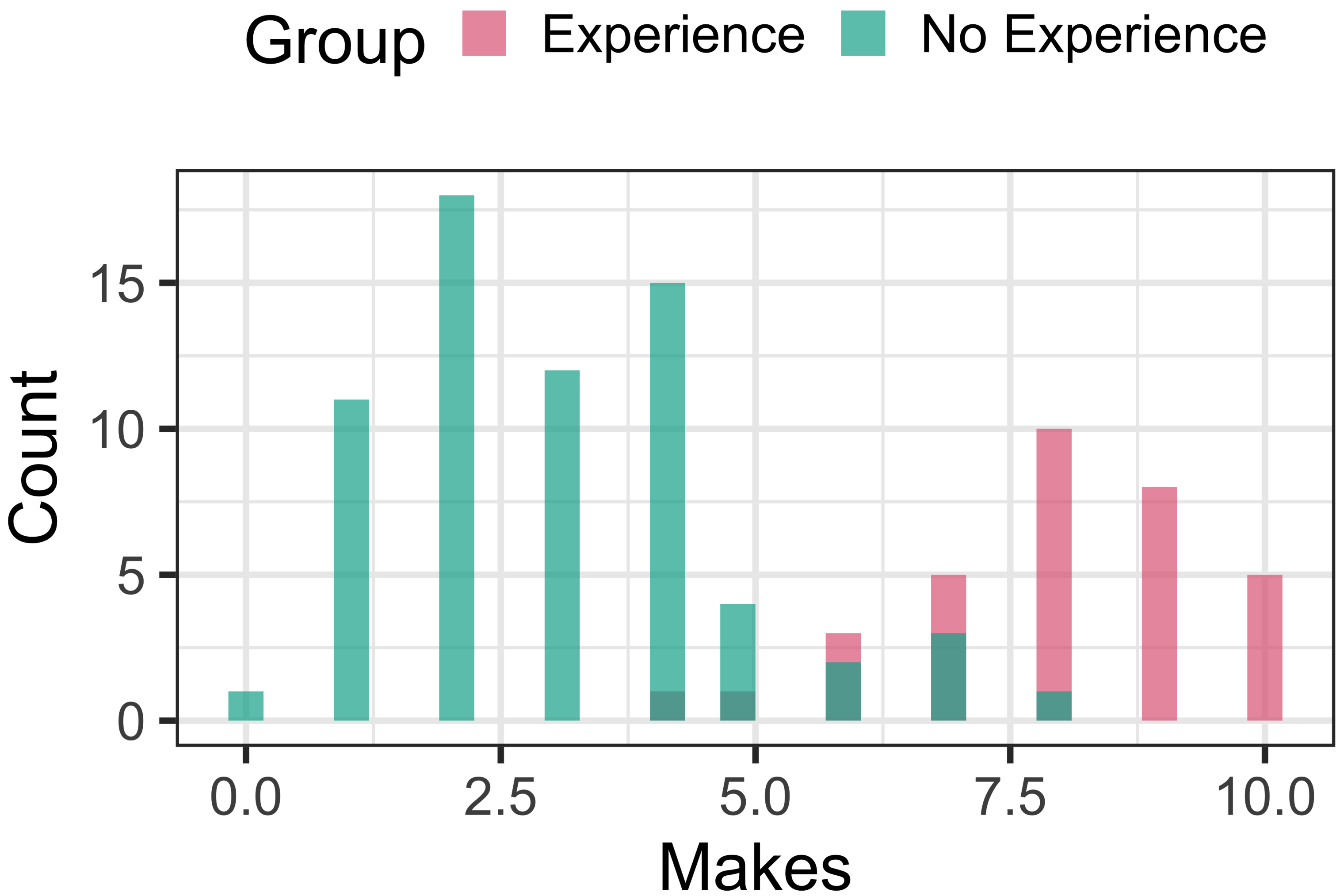
Mixture models

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

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

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

A Mixture Model



Note: 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)

Mixture models

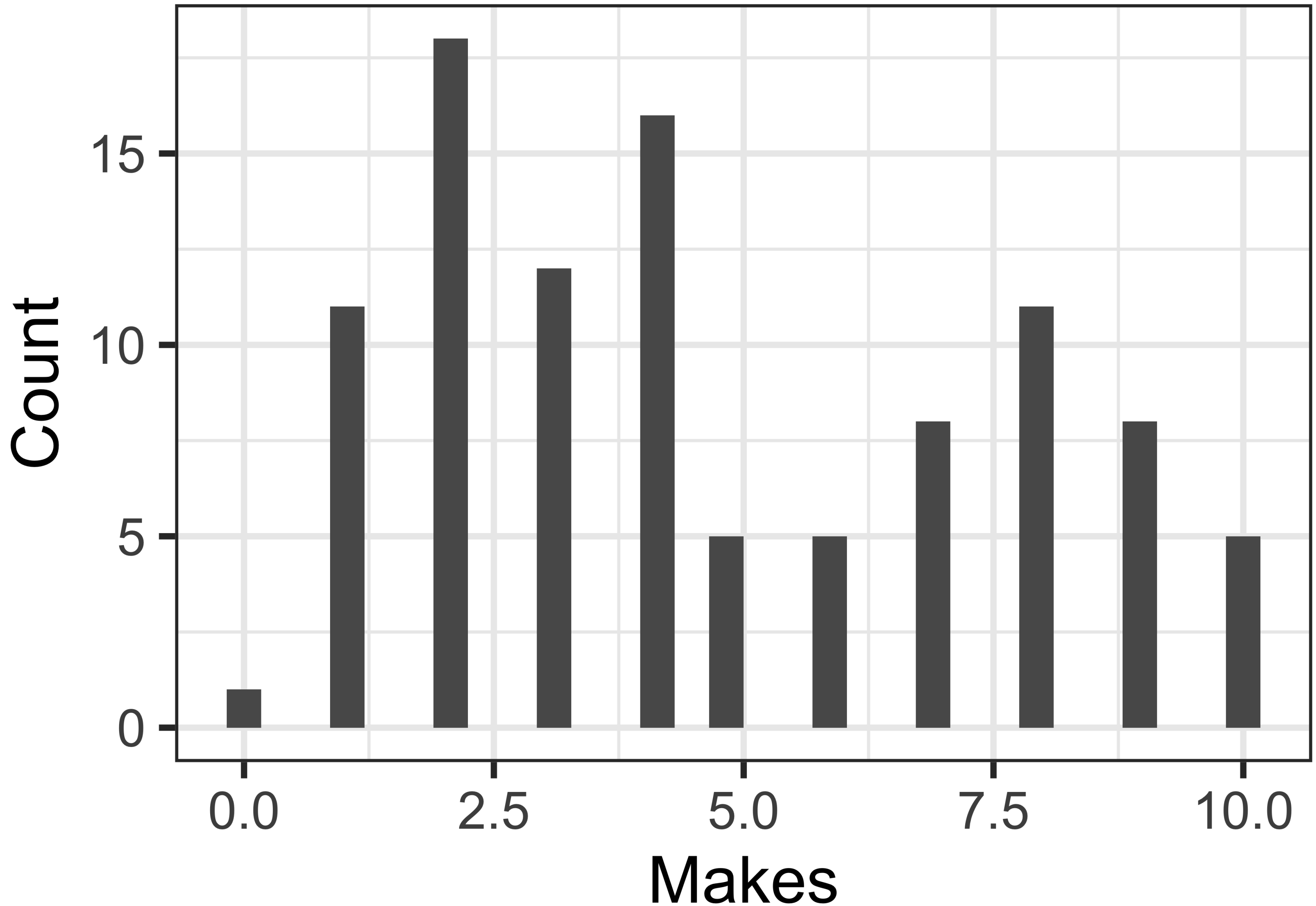
- A mixture model is a probabilistic model for representing the presence of subpopulations
- The subpopoluation to which each individual belongs is not necessarily known
  - e.g. do we ask: “have you played basketball before?”
- When z_i is not observed, we sometimes refer to it as a clustering model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

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



A finite mixture model

- Often crucial to understand the complete data generating process by introducing *latent* variables
- 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:

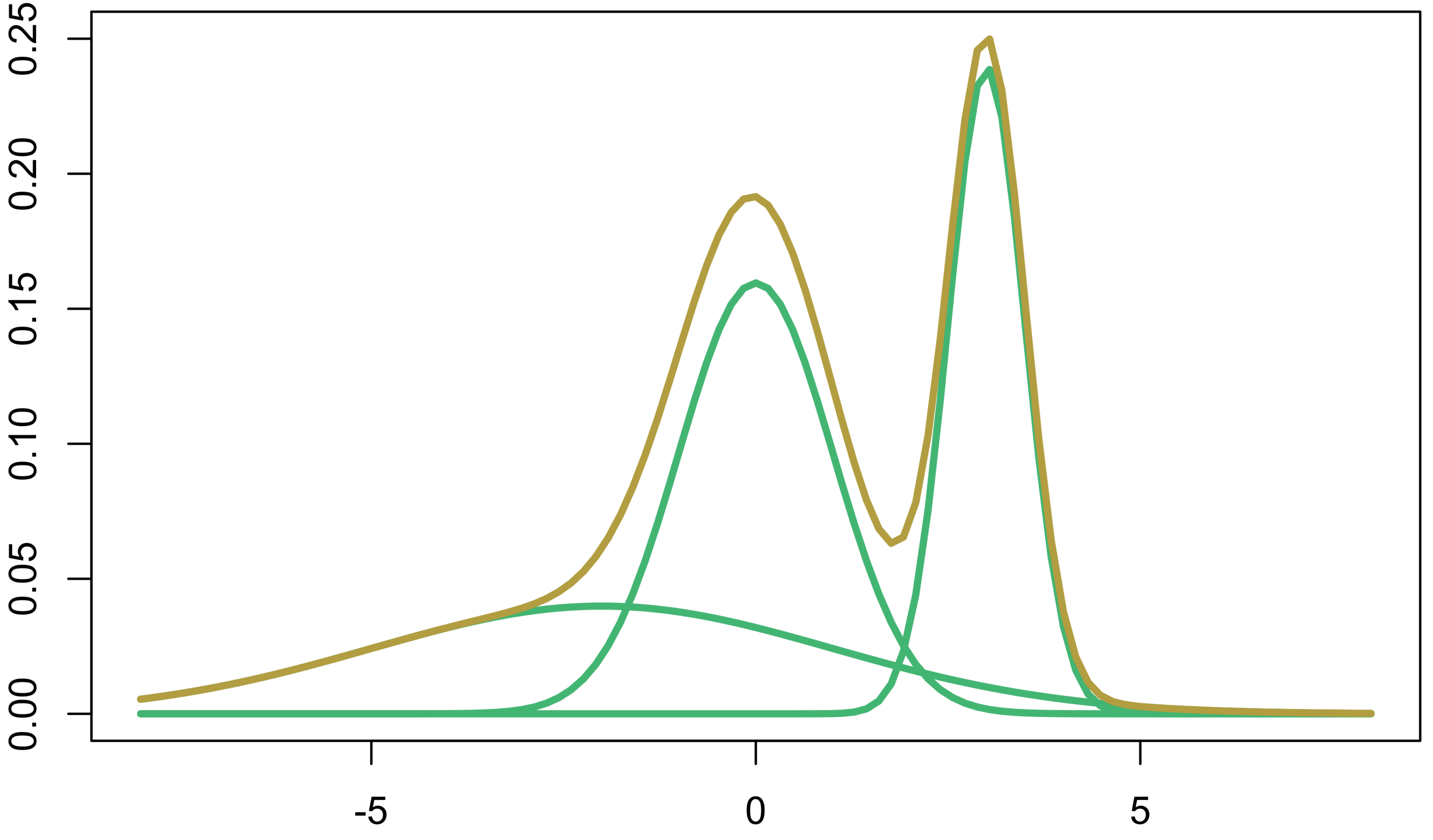
$$p(Y) = \sum_{k}^{K} \pi_k p_k(Y)$$

with $\sum \pi_k = 1$

Mixture Model Likelihood

Z unobserved

Finite mixture models



Infinite Mixture Models

- Often helpful to think about infinite mixture models
- Example 1: normal observations with normally distributed mean

$$
\begin{aligned}
\mu_i &\sim N(0, \tau^2) \\
Y_i &\sim N(\mu_i, \sigma^2)
\end{aligned}
$$

What is the distribution of Y_i given τ^2 and σ^2 (integrating over μ)?

Infinite Mixture Models

- Example 2: normal observations with exponentially distributed scale

$$\begin{aligned} \sigma_i^2 &\sim Exponential(1/2) \\ Y_i &\sim N(\mu, \sigma_i^2) \end{aligned}$$

What is the distribution of Y_i given μ?

Summary

- Likelihood, log likehood in MLE
- 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?”

Look ahead

- In Bayesian inference, unknown parameters are random variables.
    - Need to specify a prior distribution for all parameters (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 chapter 1-2 BDA3