Lecture 1: Likelihood Review
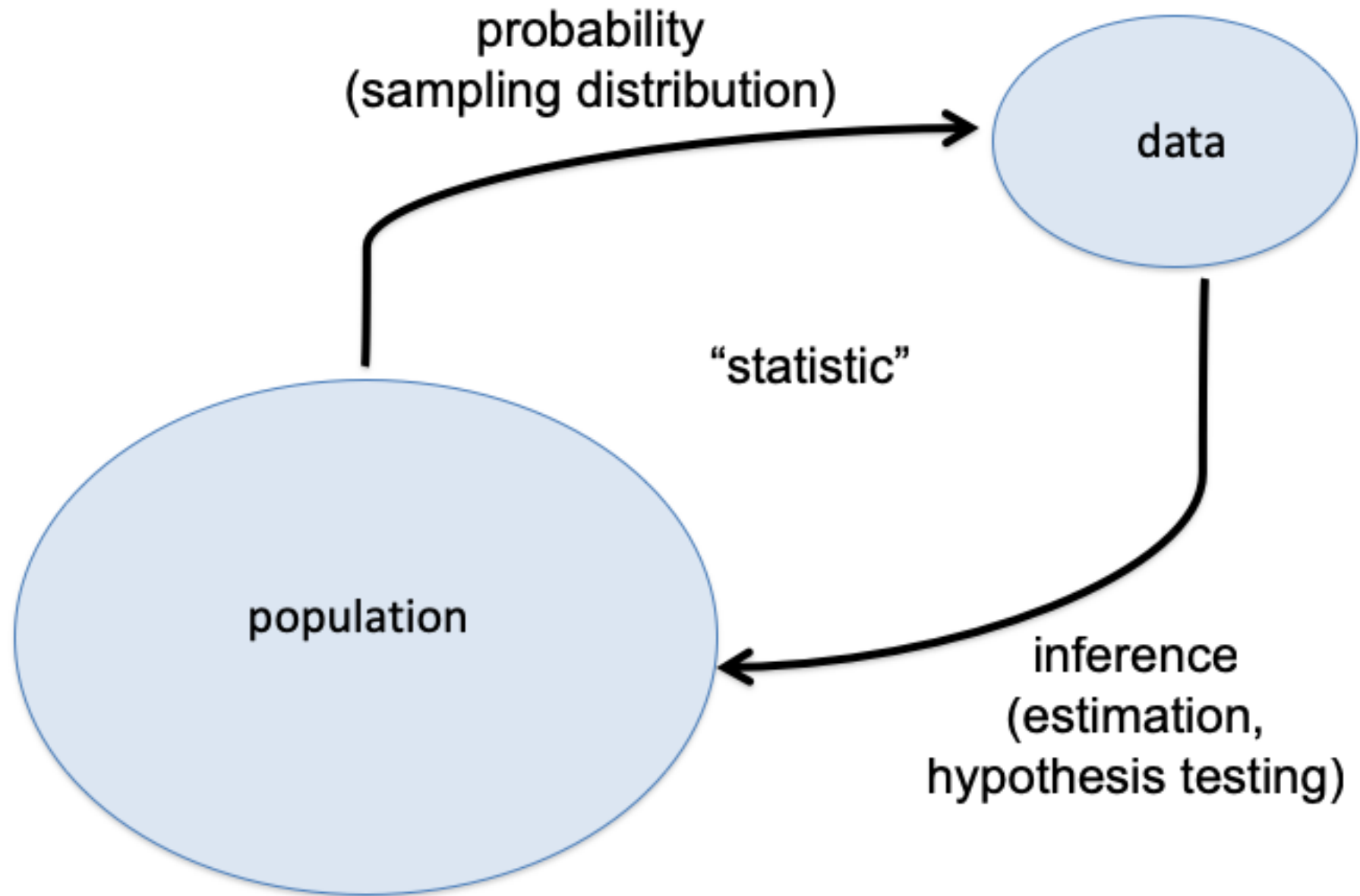
Professor Alexander Franks

Logistics

- Read: BDA Chapters 1-2

- Sync content using link on course website:
  https://bit.ly/3NXxk9H

- Annotated lecture slides appear after class

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 function is the probability density function of the observed data expressed as a function of the unknown parameter (conditional on observed data):

- 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

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

Score and Fisher Information

- **The score function:** $\frac{d\ell(\theta;y)}{d\theta}$

  - $E\left[\frac{d\ell(\theta;Y)}{d\theta} \mid \theta\right] = 0$ (under certain regularity conditions)

- **Fisher information** is a measure of the amount of information a random variable carries about the parameter

  - $I(\theta) = E\left[\frac{(d\ell(\theta;Y)}{d\theta})^2 \mid \theta\right]$ (variance of the score)

  - Equivalently: $I(\theta) = -E\left[\frac{d^2\ell(\theta;Y)}{d^2\theta}\right]$

Fisher Information

Data Generating Process

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.

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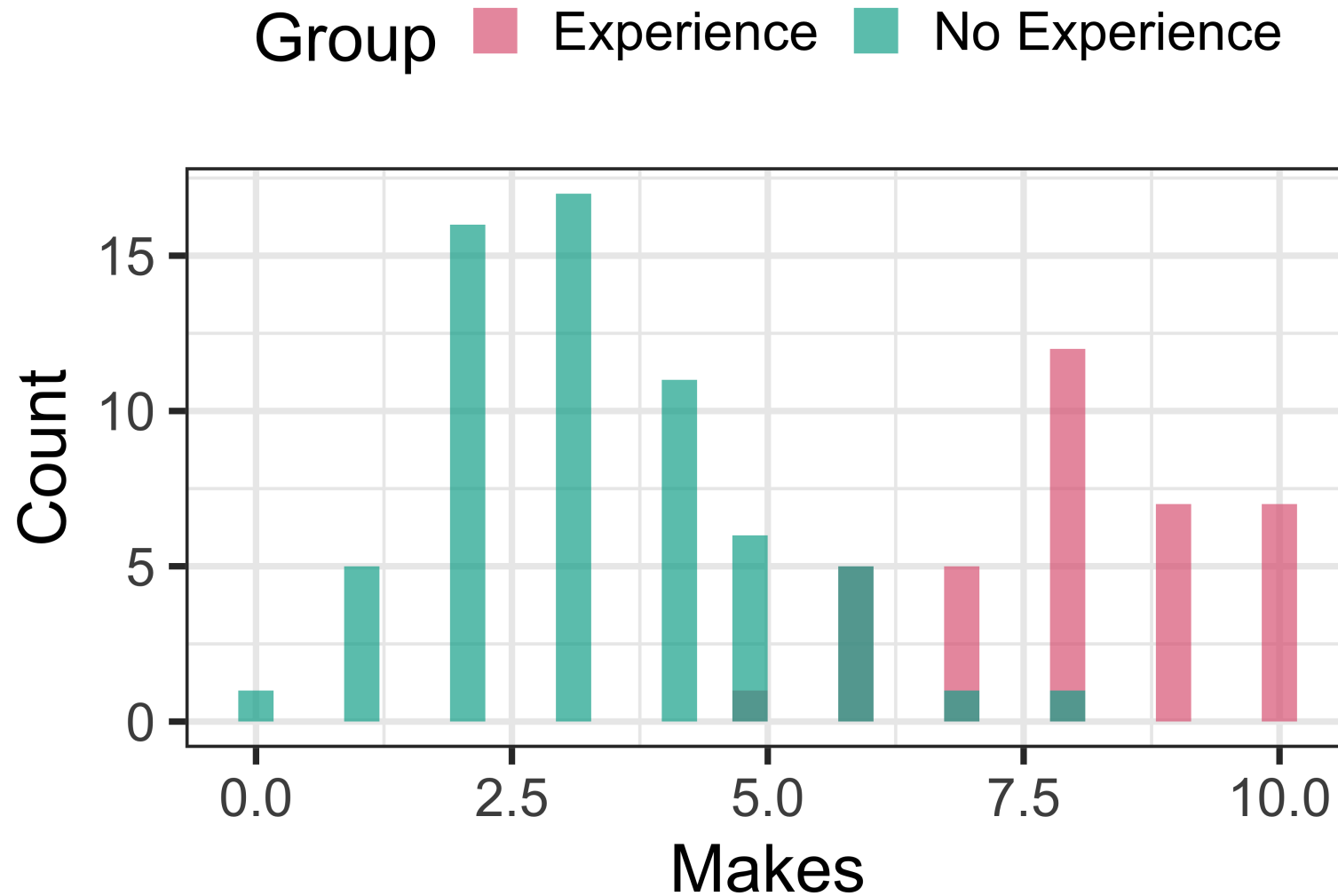
Mixture models

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

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

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

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)

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

This time we don't record who has experience with basketball.
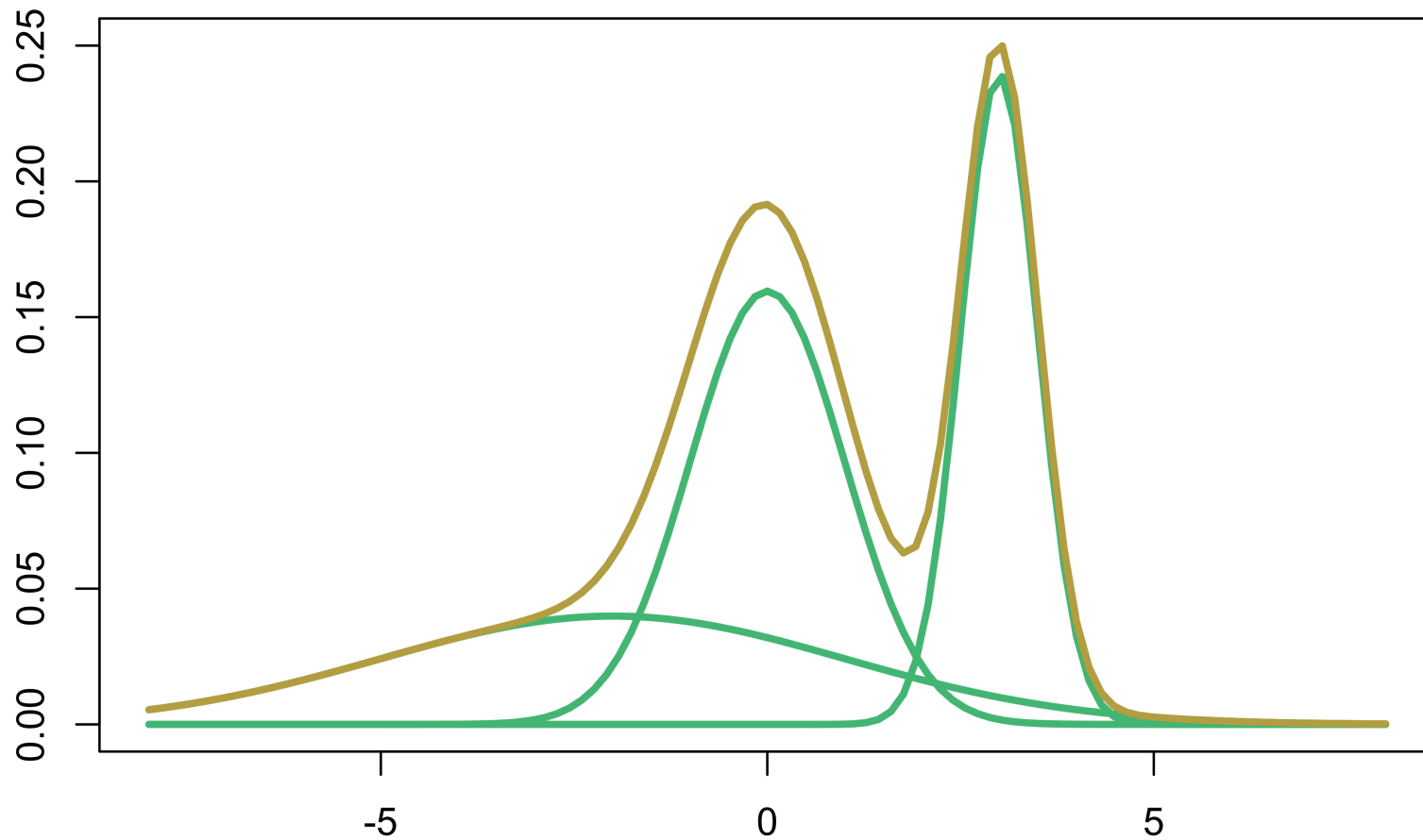
A Mixture Model
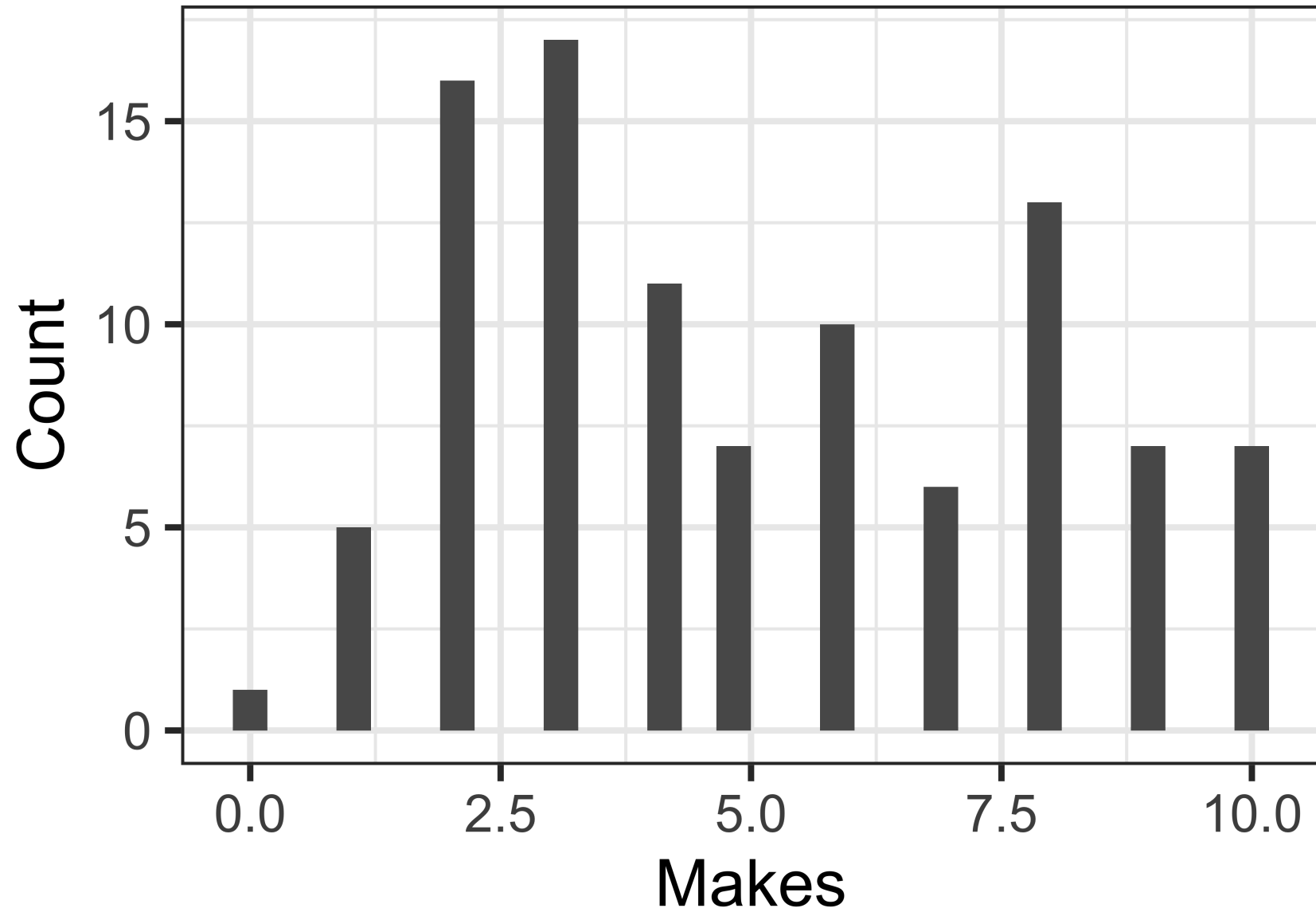
A finite mixture model

- 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*

$$p(Y \mid \theta) = \sum_z p(Y, Z = z \mid \theta)$$
$$= \sum_z p(Y \mid Z = z, \theta) p(Z = z \mid \theta)$$

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

$$p(Y) = \sum_k^K \pi_k p_k(Y) \text{ with } \sum \pi_k = 1$$

Finite mixture models

Infinite Mixture Models

- Often helpful to think about infinite mixture models

- Example 1: normal observations with normally distributed mean

$$\mu_i \sim N(0, \tau^2)$$
$$Y_i \sim N(\mu_i, \sigma^2)$$

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

Infinite Mixture Models

Example 2: Poisson observations with random rates

$$\lambda \sim Gamma(\alpha, \beta)$$
$$Y \sim Pois(\lambda)$$

Infinite Mixture Models

- Example 3: normal observations with exponentially distributed scale

$$\sigma_i^2 \sim Exponential(1/2)$$
$$Y_i \sim N(0, \sigma_i^2)$$

What is the distribution of Y_i?

Summary

- Likelihood, log likehood

- Sufficient statistics

- Fisher information

- Mixture models

Assignments

- Read chapter 1-2 BDA3