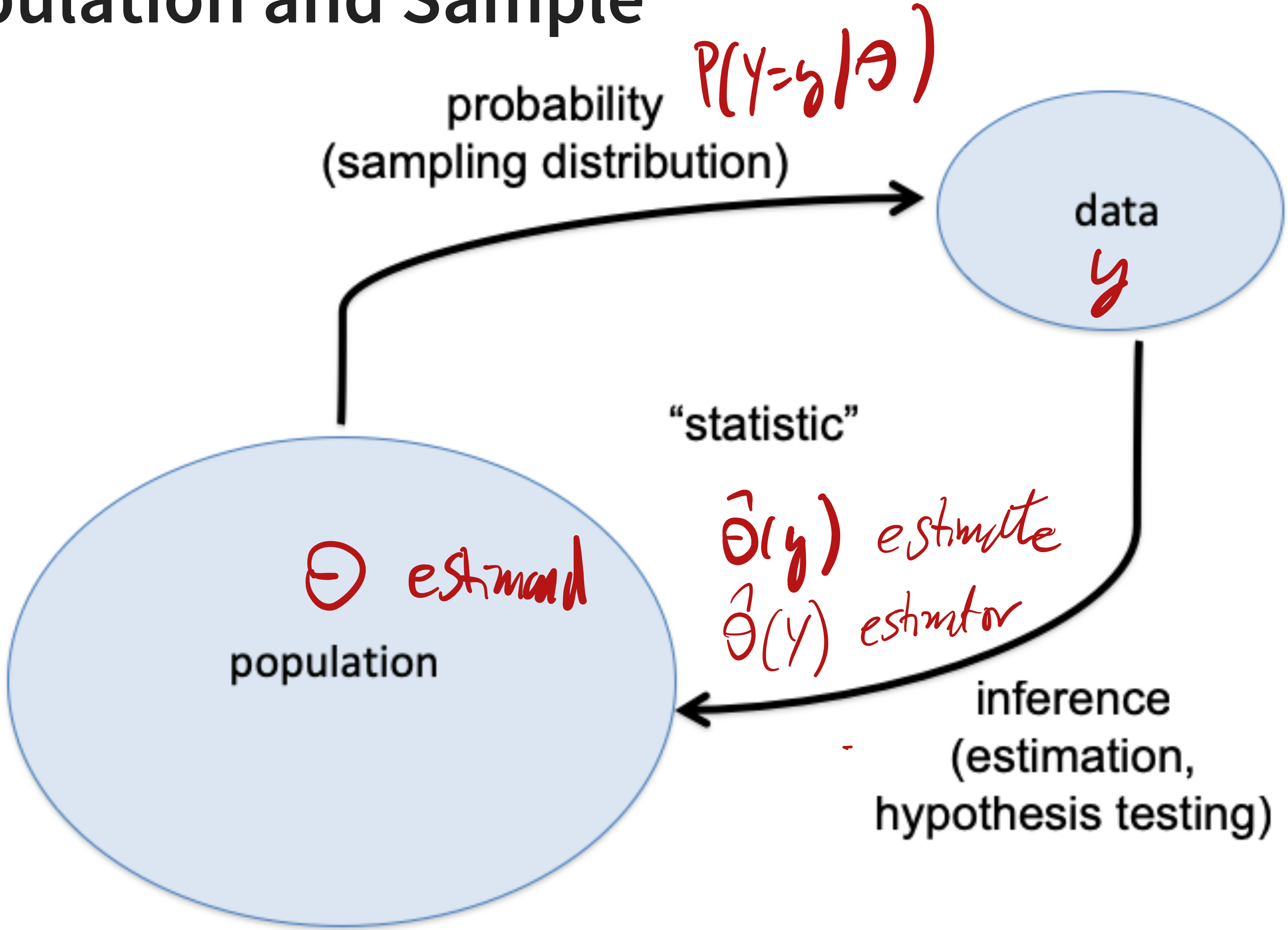
Lecture ~~1~~ 2: Likelihood Review

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
- No Midterm.

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

Exchangeable:

$P(y_1, \ldots, y_n) = P(y_{\pi(1)}, \ldots, y_{\pi(n)})$

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

$$L(\theta) \underset{\propto}{=} P(Y_1 = y_1, \ldots, Y_n = y_n \mid \theta)$$

If iid: $L(\theta) \propto \prod_{i=1}^{n} P(Y_i = y_i \mid \theta)$

Sufficient Statistics

vector

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.

$$P(Y_1, \ldots Y_n \mid S(Y_1, \ldots Y_n), \theta) = P(Y_1, \ldots Y_n \mid S(Y_1, \ldots Y_n))$$

$$Y_1, \ldots Y_n \overset{iid}{\sim} N(\theta, 1)$$

$$\longrightarrow \quad Y_i \mid \bar{Y} \sim N(0, 1 - 1/n) \quad \text{(exercise)}$$

Bayesian: $P(\theta \mid y_1, \ldots y_n) = P(\theta \mid S(y_1, \ldots y_n))$

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

$$L(\theta) = \underset{\text{no } \theta}{h(y_1, \dots y_n)}\underbrace{g(s(y), \theta)}_{\text{only } s(y) \,\&\, \theta}$$

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

$$P(\theta \mid y_1, \dots, y_n) \propto L(\theta) P(\theta)$$

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

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

$Y \sim \text{Bin}(12, \theta)$ obs: $y = 3$

$L(\theta; y=3) \propto \cancel{\binom{12}{3}} \theta^3 (1-\theta)^9$

$X \sim NB(3, \theta)$ (keep flipping until 3 heads)

obs $x = 9$

$L(\theta; x=9) = \cancel{\binom{11}{2}} \theta^3 (1-\theta)^9$

$H_0: \theta = .5$

$H_a: \theta < .5$

p-values

Bin: pbinom(3, 12, $\theta = 1/2$)

" .073

NB: 1 − pnbinom(8, 3, 1/2)

" 0.033

Score and Fisher Information

- **The score function:** $\frac{d\ell(\theta;y)}{d\theta}$
  - $E[\frac{d\ell(\theta;Y)}{d\theta} \mid \theta] = 0$ (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: $I(\theta) = -E_{Y|\theta}\left[\frac{d^2\ell(\theta;Y)}{d^2\theta}\right]$

Observed Info: $-\frac{d^2\ell(\theta;y)}{d^2\theta}$ obs data

Fisher Information

$$Y_1, \ldots Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

known

$$L(\mu) \propto \prod^{n} \cancel{\frac{1}{\sqrt{2\pi\sigma^2}}} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}}$$

$$\left.\propto e^{\frac{-(\bar{y}-\mu)^2}{2\sigma^2/n}}\right\} g(s(y), \theta)$$

$$\ell(\mu) = \frac{-(\bar{y}-\mu)^2}{2\sigma^2/n}$$

$$\ell'(\mu) = \frac{(\bar{y}-\mu)}{\sigma^2/n}, \quad \ell'' = \frac{-n}{\sigma^2} \Rightarrow I(\mu) = \frac{n}{\sigma^2}$$

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.

```
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

Experience

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

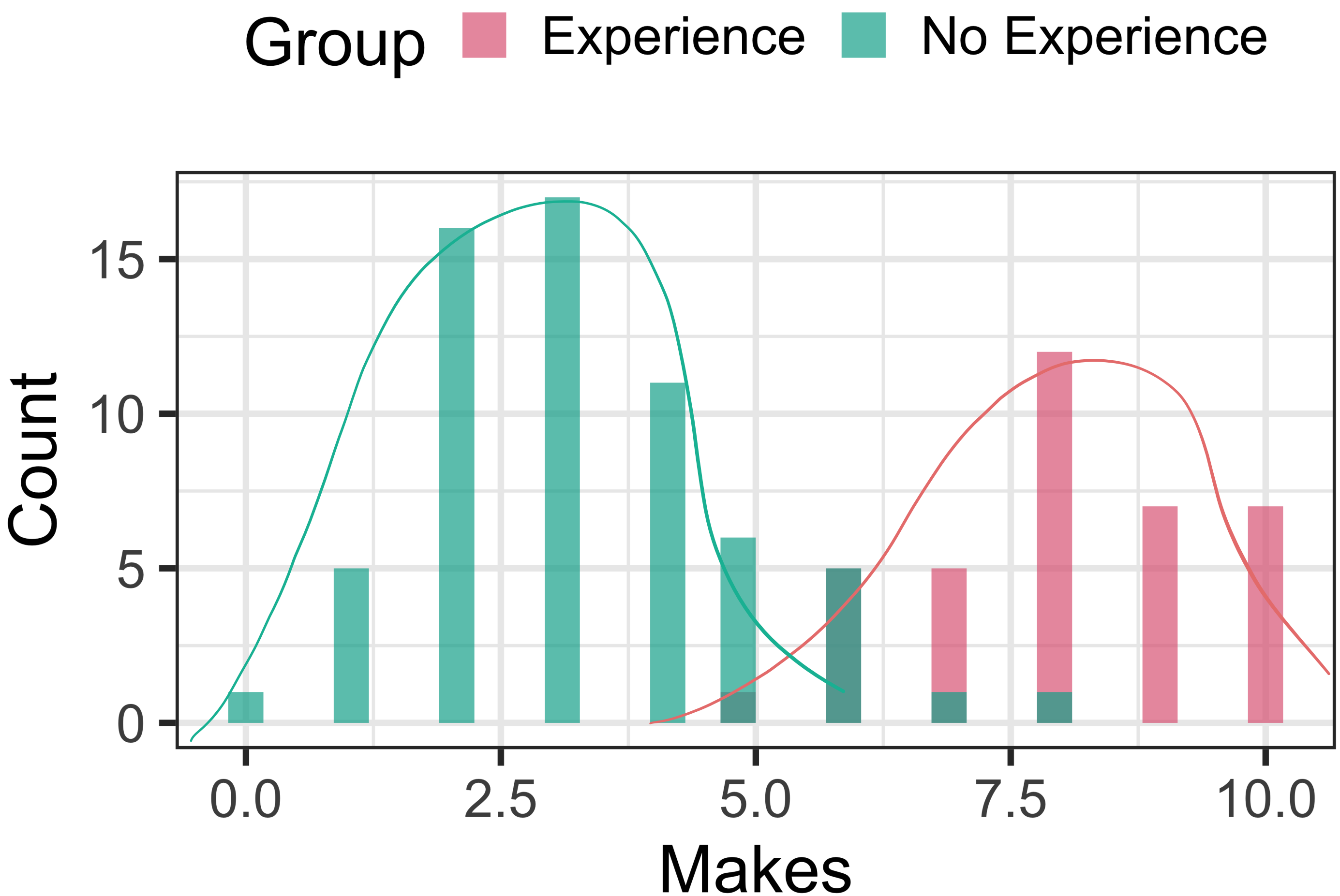
$$Z_i \sim Bin(1, \phi)$$

Fraction Experienced

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

Make Probabilities

A Mixture Model



Note: z is observed

$$L(\phi, \theta_0, \theta_1) \propto \prod_{i=1}^{n} P(y_i, z_i \mid \theta_0, \theta_1, \phi)$$

$$\propto \prod_{i=1}^{n} \left(\left[\phi \cancel{\binom{10}{y_i}} \theta_1^{y_i} (1-\theta_1)^{10-y_i} \right]^{z_i} \times \left[(1-\phi) \cancel{\binom{10}{y_i}} \theta_0^{y_i} (1-\theta_0)^{10-y_i} \right]^{1-z_i} \right)$$

$$\propto \phi^{\sum z_i} \theta_1^{\sum y_i z_i} (1-\theta_1)^{\sum (10-y_i) z_i} (1-\phi)^{\sum (1-z_i)} \theta_0^{\sum y_i (1-z_i)} (1-\theta_0)^{\sum (-y_i)(1-z_i}$$

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

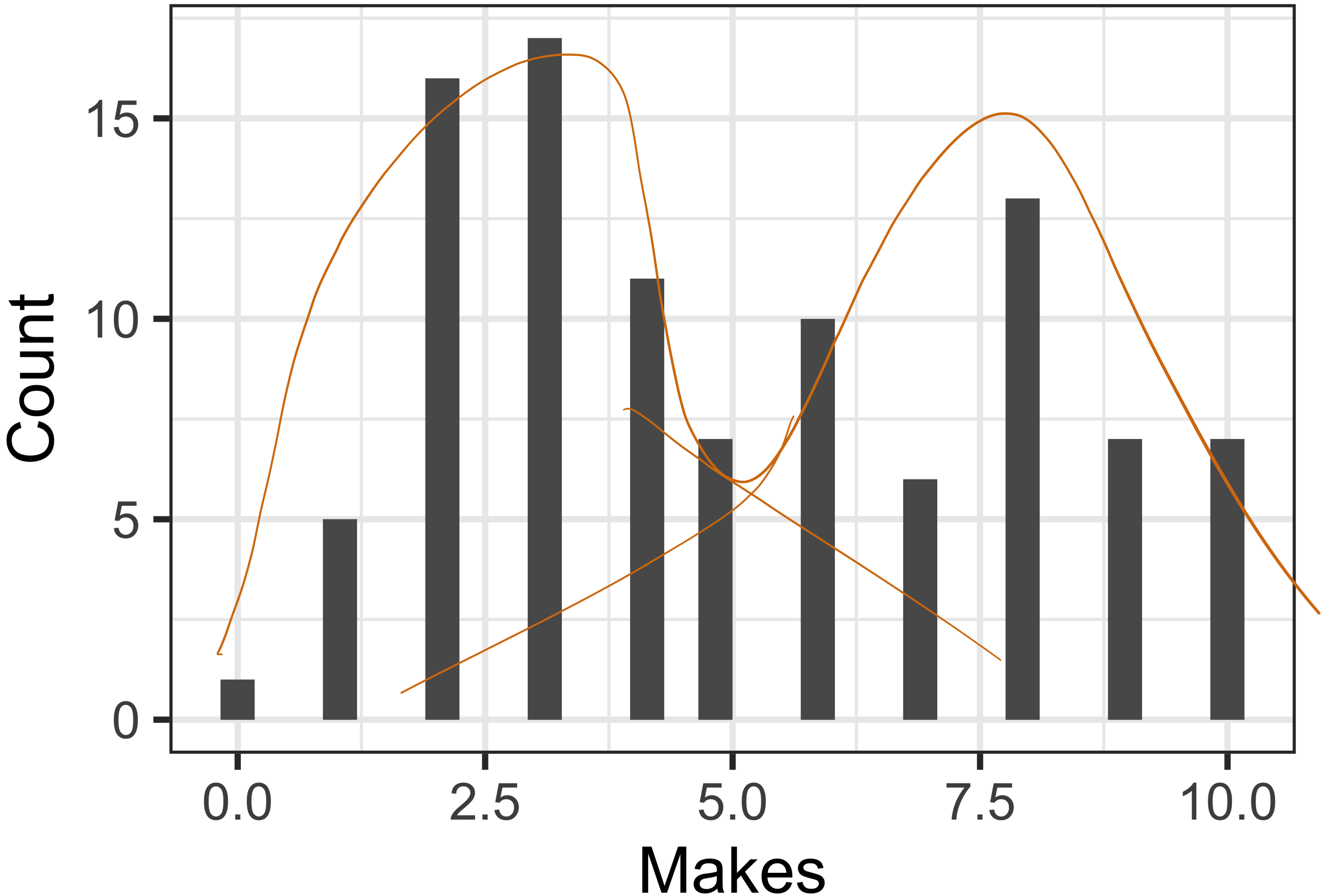
- A mixture model is a probabilistic model for representing the presence of subpopulations
- The subpopoluation to which each individual belongs is not necessarily known
  - e.g. do we ask: “have you played basketball before?”
- When z_i is not observed, we sometimes refer to it as a clustering model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



$$L(\theta_1, \theta_0, \phi) \propto \prod_{i=1}^{n} P(y_i \mid \theta_0, \theta_1, \phi)$$

$P(X,Y) = P(X|Y)P(Y)$

$$\propto \prod_{i=1}^{n} \sum_{z=0}^{z=1} P(y_i, z_i \mid \theta_1, \theta_0, \phi)$$

$$\propto \prod^{n} \left[\sum_{z} P(y_i \mid z_i, \theta_1, \theta_0, \phi) \, P(z_i \mid \phi) \right]$$

$$\propto \prod^{n} \left[\phi \cancel{\binom{10}{y_i}} \theta_1^{y_i} (1-\theta_1)^{10-y_i} + (1-\phi) \cancel{\binom{10}{y_i}} \theta_0^{y_i} (1-\theta_0)^{10-y_i} \right]$$

A finite mixture model

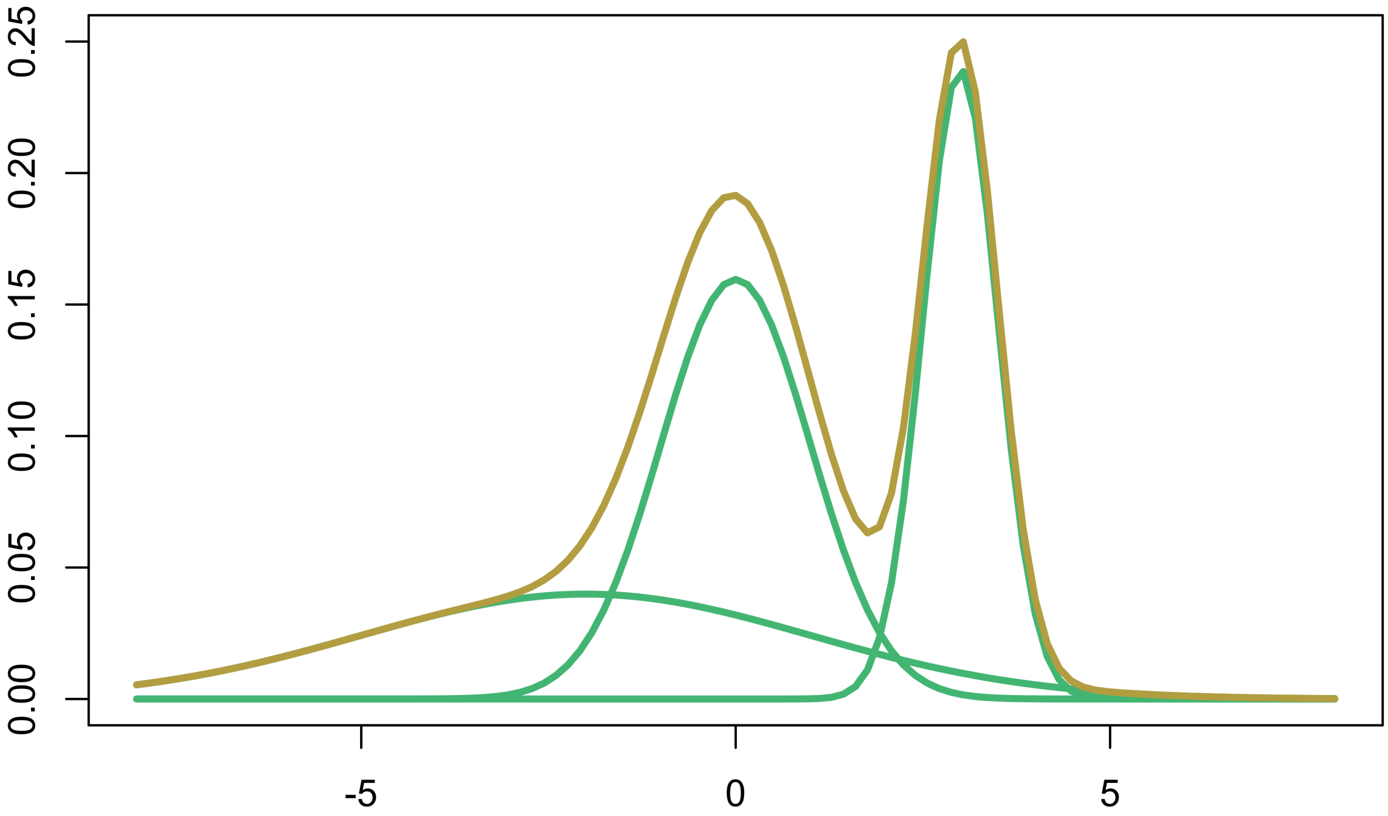
- Often crucial to understand the complete data generating process by introducing *latent* variables
- Write the *observed data likelihood* by integrating out the latent variables from the *complete data likelihood*

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

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

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

Finite mixture models



Infinite Mixture Models

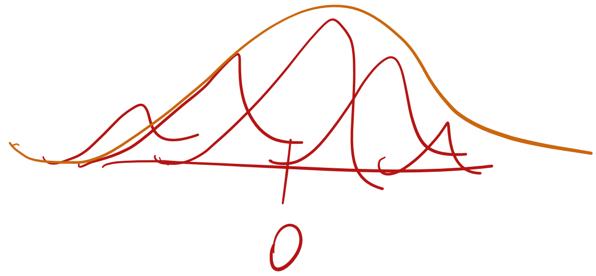
- Often helpful to think about infinite mixture models
- Example 1: normal observations with normally distributed mean

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

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

$P(Y_i \mid \tau^2, \sigma^2) = \int P(Y_i \mid \mu_i, \sigma^2) P(\mu_i \mid \tau^2) d\mu$

- Calculus
- MGFs
- Representation



$$\varepsilon \sim N(0, \sigma^2)$$

$$Y_i = \mu_i + \varepsilon \qquad \mu_i \sim N(0, \tau^2)$$

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

Infinite Mixture Models

Example 2: Poisson observations with random rates

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

$$P(Y \mid \alpha, \beta) = \int_{\lambda} P(Y \mid \lambda) P(\lambda \mid \alpha, \beta) \, d\lambda$$

$$= \int_{\lambda} \frac{\lambda^{y} e^{-\lambda}}{y!} \frac{\beta^{\alpha} \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \, d\lambda$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)\Gamma(y+1)} \int \underbrace{\lambda^{y+\alpha-1} e^{-(\beta+1)\lambda}}_{\substack{\text{kernel of Gamma} \\ Gam(\alpha+y,\, \beta+1)}} d\lambda$$

$$\frac{(\beta+1)^{\alpha+y}}{\Gamma(y+\alpha)}$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)\Gamma(y+1)} \frac{(\beta+1)^{\alpha+y}}{\Gamma(y+\alpha)}$$

Negative Binomial

Infinite Mixture Models

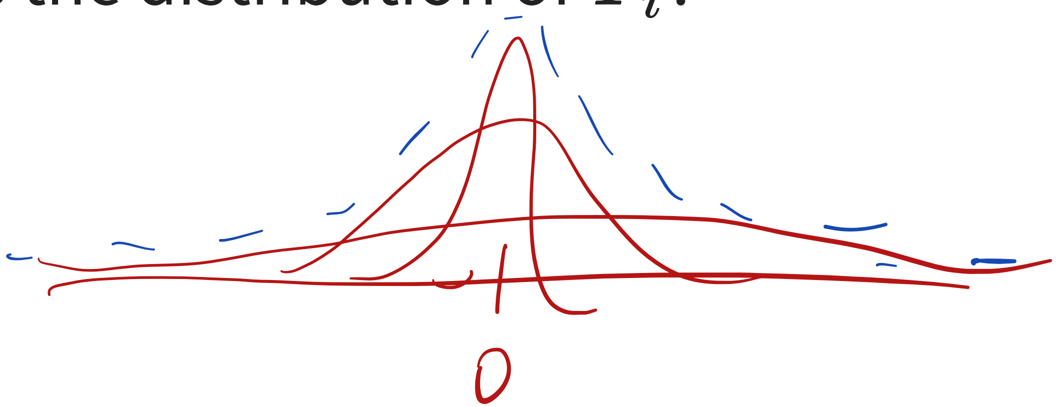
- Example 3: normal observations with exponentially distributed scale

Inv-Gam$(\nu/2, \nu/2)$

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

Laplace

What is the distribution of Y_i?



$$y_i \sim N(0, \sigma_i^2)$$

$$\sigma_i = 1/|z_i|$$

$$z_i \sim N(0,1)$$

$$x_i \sim N(0,1)$$

$$z_i \sim N(0,1)$$

$$y_i = \frac{x_i}{|z_i|}$$

$\downarrow$

Cauchy

Summary

- Likelihood, log likehood
- Sufficient statistics
- Fisher information
- Mixture models

Assignments

- Read chapter 1-2 BDA3