

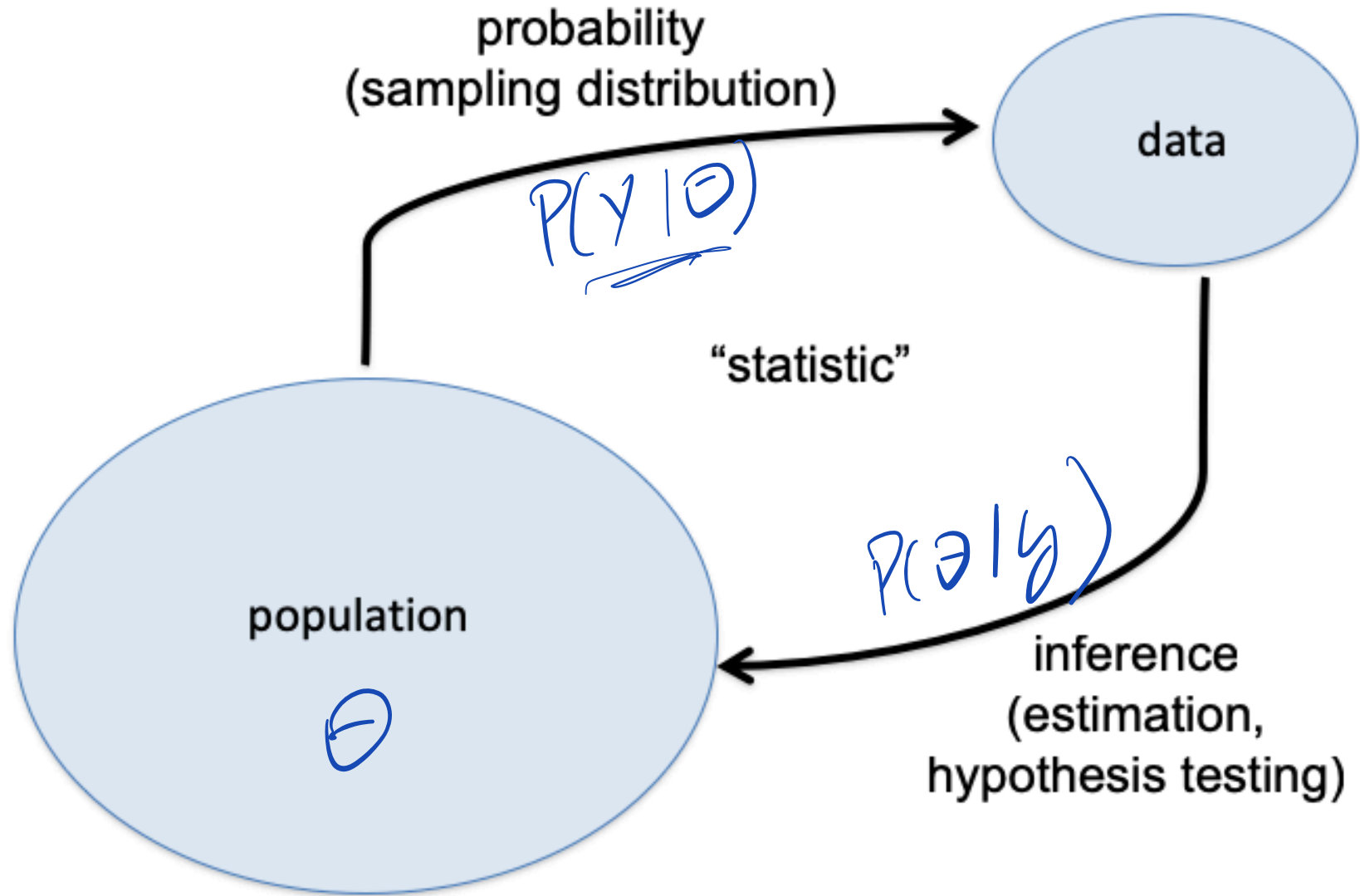
# Lecture 1: Likelihood Review

Professor Alexander Franks

# Logistics

- Read: BDA Chapters 1-2
- Sync content using link on course website:  
~~<https://bit.ly/3NXyk9H>~~ [tinyurl.com/pstat2/5a](https://tinyurl.com/pstat2/5a)
- Annotated lecture slides appear after class
- Homework 1 out
- OH: Wed 2pm SH 5522

# Population and Sample

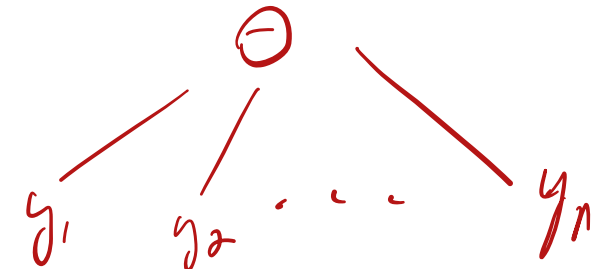


# Independent Random Variables

- $Y_1, \dots, Y_n$  are random variables
- We say that  $Y_1, \dots, Y_n$  are conditionally independent given  $\theta$  if  $P(y_1, \dots, y_n \mid \theta) = \prod_i P(y_i \mid \theta)$
- Conditional independence means that  $Y_i$  gives no additional information about  $Y_j$  beyond that in knowing  $\theta$

~~$$P(y_1, \dots, y_n) \stackrel{??}{=} \prod P(y_i)$$~~

→ "Exchangeable":



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

# 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  $\theta$ .
- Depends on the observed data  $y = (y_1, y_2, \dots, y_n)$
- Two likelihood functions are equivalent if one is a scalar multiple of the other

$$y_i \stackrel{\text{iid}}{\sim} p(y/\theta), \quad L(\theta; y_1, \dots, y_n) = \prod_{i=1}^n p(y=y_i | \theta)$$

# Sufficient Statistics

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

$$Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\theta, 1)$$

$$\bar{Y} \text{ is sufficient} \Rightarrow \underline{Y_i} \mid \underline{\bar{Y}} \stackrel{\text{iid}}{\sim} \overset{\text{"No } \theta \text{"}}{N(\bar{y}, 1 - \frac{1}{n})}$$

Bayesian:  $P(\theta | y_1, \dots, y_n) = P(\theta | s(y))$

# 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:

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

*Handwritten notes:* "No  $\theta$ " above the crossed-out  $h$ ; " $\theta$  &  $s(y)$  only" above the  $g$  term.

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

# 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

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

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

$$X \sim \text{NB}(3, \theta) \quad \text{obs } \underline{x = 9}$$

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

$$H_0: \theta = 1/2$$

$$H_a: \theta < 1/2$$

$$\text{Bin: } p_{\text{binom}}(3, 12, \theta = 1/2) = .073$$

$$\text{NB: } 1 - p_{\text{binom}}(8, 3, \theta = 1/2) = .033$$

# Score and Fisher Information

$L$ : likelihood  
 $l$ : log-likelihood.

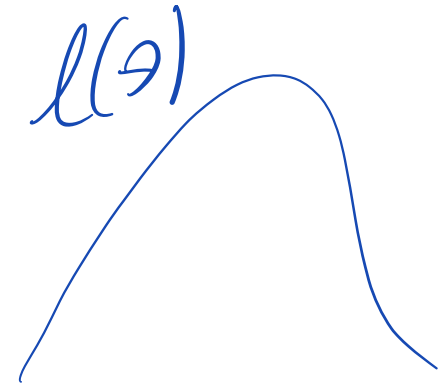
- The score function:  $\frac{d\ell(\theta; y)}{d\theta}$   
 $MLE$  is  $\hat{\theta}$  s.t.  $l'(\hat{\theta}) = 0$ 
  - $E\left[\frac{d\ell(\theta; Y)}{d\theta} \mid \theta\right] = 0$  (under certain regularity conditions)

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

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

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

"How peaked/curved  
the likelihood is"



# Fisher Information

$$y_1, \dots, y_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$$

known value.

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

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

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

$$l''(\mu) = -\frac{n}{\sigma^2}$$

$$I(\mu) = -E_y[l''(\mu)] = \frac{n}{\sigma^2}$$

Cramer-Rao  
Bound

$$\text{Var}(\hat{\theta}) \geq 1/I(\theta)$$

for  $\hat{\theta}$   
unbiased  
estimator.

# 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)
```

$\phi$  chance of experience

# Mixture models

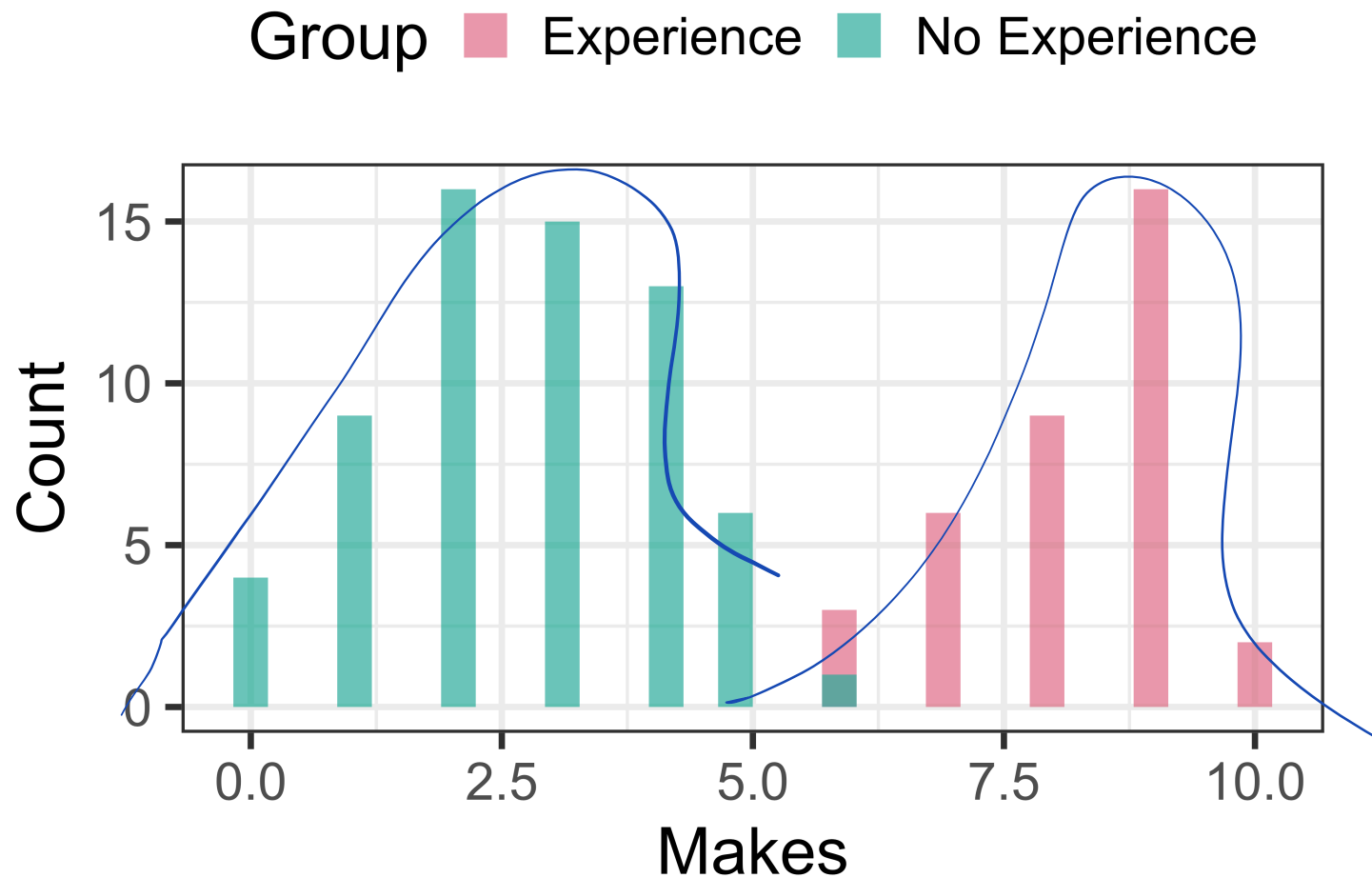
$$Z_i = \begin{cases} 0 & \text{if the } i^{th} \text{ student doesn't play basketball} \\ 1 & \text{if the } i^{th} \text{ student does play basketball} \end{cases}$$

$$Z_i \sim \text{Bin}(1, \phi)$$

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

$$\mathcal{L}(\phi, \theta_0, \theta_1) =$$

# A Mixture Model



Note:  $z$  is observed

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

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

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

$$\begin{aligned} & \phi^{\sum z_i} \theta_1^{\sum y_i z_i} (1-\theta_1)^{\sum (10-y_i) z_i} \\ & (1-\phi)^{\sum (1-z_i)} \theta_0^{\sum y_i (1-z_i)} (1-\theta_0)^{\sum (10-y_i)(1-z_i)} \end{aligned}$$

$$P(y|z, \theta_1, \theta_0, \phi) = P(y|z, \theta_1, \theta_0)$$

# 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 subpopulation 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)

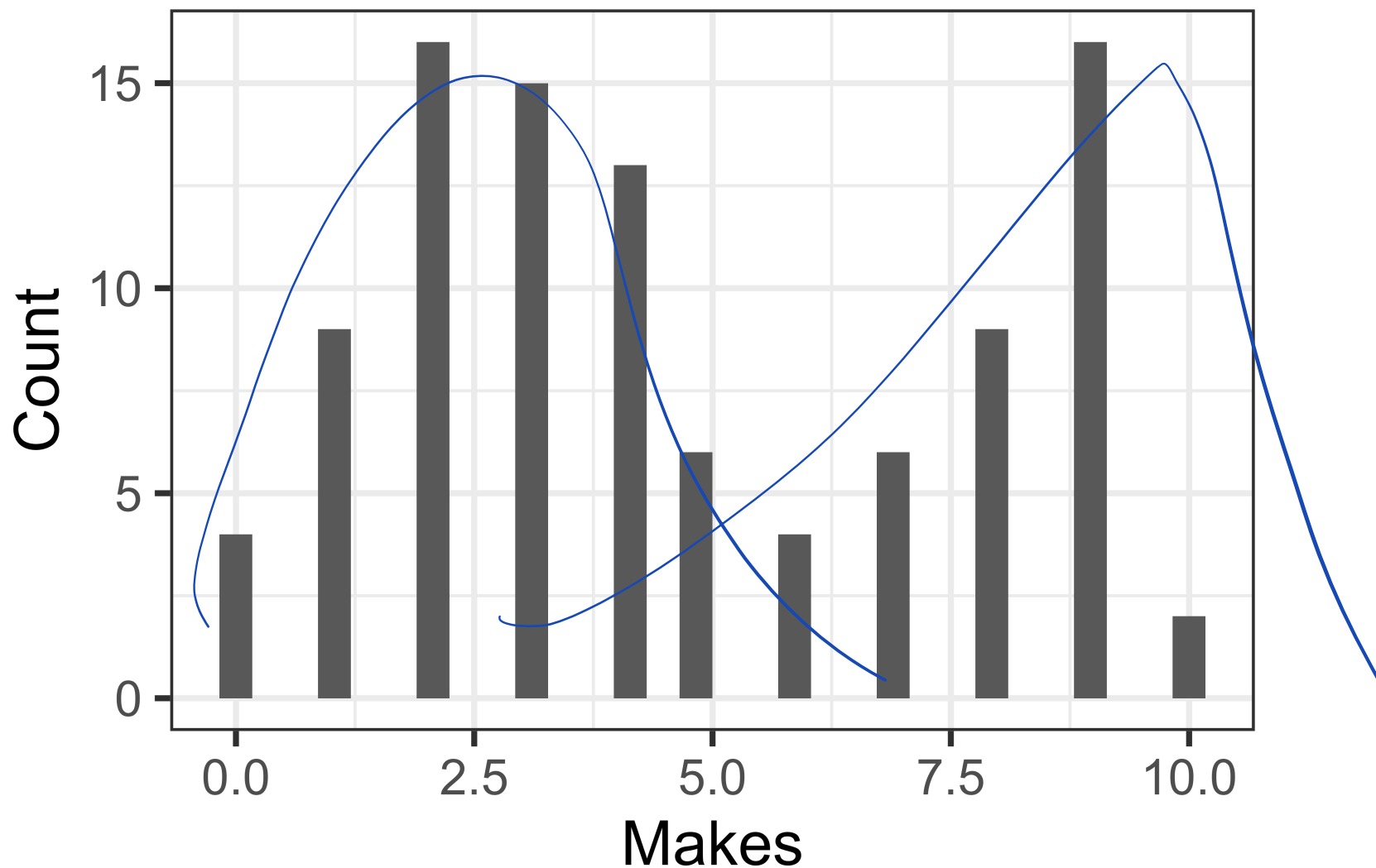
```
1 for (i in 1:100)
2   - Generate z_i from Bin(1, phi)
3   - p_i = theta_1 if z_i=1
4   - p_i = theta_0 if z_i=0
5   - Generate y_i from a Binom(10, p_i)
6 return y = (y_1, ... y_100)
```

*No  $z_i$*

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

# A Mixture Model

$$\mathcal{L}(\theta, \theta_0, \phi; y_1, \dots, y_{100})$$



# A finite mixture model

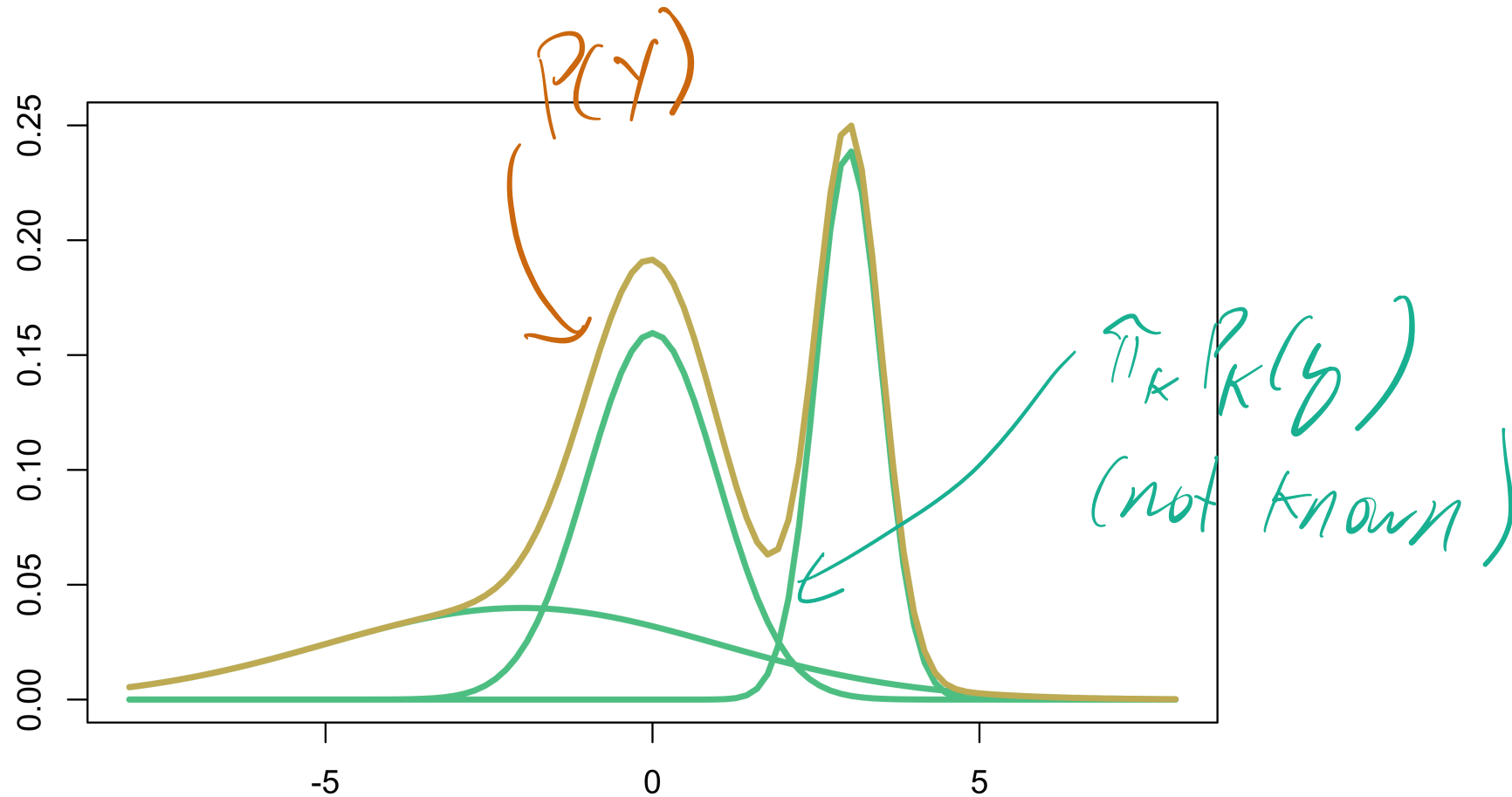
- Often crucial to understand the complete data generating process by introducing *latent* variables
- Write the *observed data likelihood* by integrating out the latent variables from the *complete data likelihood*

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

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



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

$$\propto \prod_{i=1}^n \left[ \sum_{z=0}^1 P(y_i | z_i = z_i, \theta_0, \theta_1) P(z_i = z_i | \phi) \right]$$

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

Can't simplify

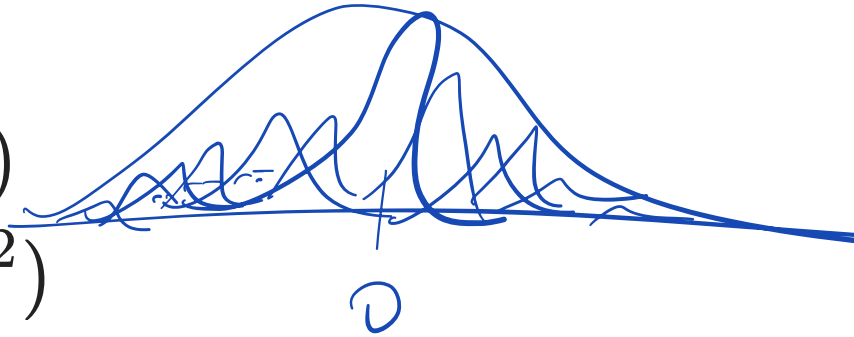
$(y_{11}, \dots, y_{100})$  is minimal sufficient.

# 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  $\tau^2$  and  $\sigma^2$  (integrating over  $\mu$ )?

$$P(Y_i | \mu_i) = Y_i \sim N(\mu_i, \sigma^2)$$

$$P(\mu_i) \quad \mu_i \sim N(0, \tau^2)$$

$$P(Y_i) = \int_{-\infty}^{\infty} P(Y_i, \mu_i) d\mu_i$$

$$= \int \underbrace{P(Y_i | \mu_i)} \underbrace{P(\mu_i)} d\mu_i$$

$$\left| \begin{array}{l} Y_i = \mu_i + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \\ \mu_i \sim N(0, \tau^2) \end{array} \right.$$

$$\begin{array}{c} \text{Representation} \\ \hline Y_i \sim N(0, \sigma^2 + \tau^2) \end{array}$$

# Infinite Mixture Models

Example 2: Poisson observations with random rates

$$\left| \begin{array}{l} \lambda_i \sim \text{Gamma}(\alpha, \beta) \\ Y_i \sim \text{Pois}(\lambda_i) \end{array} \right.$$

$$E[\lambda] = \frac{\alpha}{\beta}$$

$$\text{Var}(\lambda) = \frac{\alpha}{\beta^2}$$

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

$$= \int_0^{\infty} \underbrace{\frac{\lambda^y e^{-\lambda}}{y!}}_{\text{pois}} \underbrace{\frac{\beta^{\alpha} \lambda^{\alpha-1} e^{-\beta \lambda}}{\Gamma(\alpha)}}_{\text{gamma}} d\lambda$$

A Gamma-Poisson Mixture is  
a Negative-Binomial.

(Model count data where  $\text{Var} > \text{Mean}$ )

$$E[Y] \stackrel{\text{iterated Expectation}}{=} E[E[Y|\lambda]] = E[\lambda] = \frac{\alpha}{\beta}$$

$$\text{Var}(Y) \stackrel{\substack{\text{Law of total} \\ \text{Variance}}}{=} \underbrace{E[\text{Var}(Y|\lambda)]}_{\text{EVVEs law}} + \underbrace{\text{Var}(E[Y|\lambda])}$$

$$= E[\lambda] + \text{Var}(\lambda)$$

$$= \frac{\alpha}{\beta}$$

$$+ \frac{\alpha}{\beta^2}$$

extra  
dispersion

# Infinite Mixture Models

- Example 3: normal observations with exponentially distributed scale

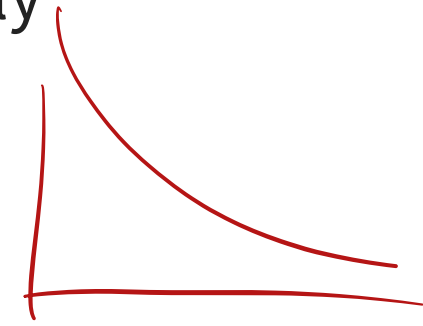
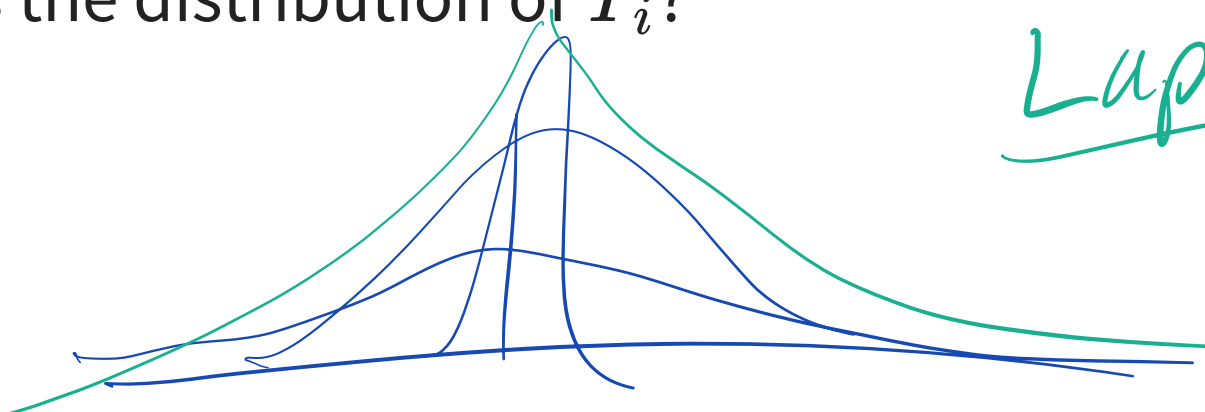
*Scale-Mixture  
of  
Normals*

$$\sigma_i^2 \sim \text{Exponential}(1/2)$$

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

What is the distribution of  $Y_i$ ?

*Laplace*



# Summary

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

# Assignments

- Read chapter 1-2 BDA3