

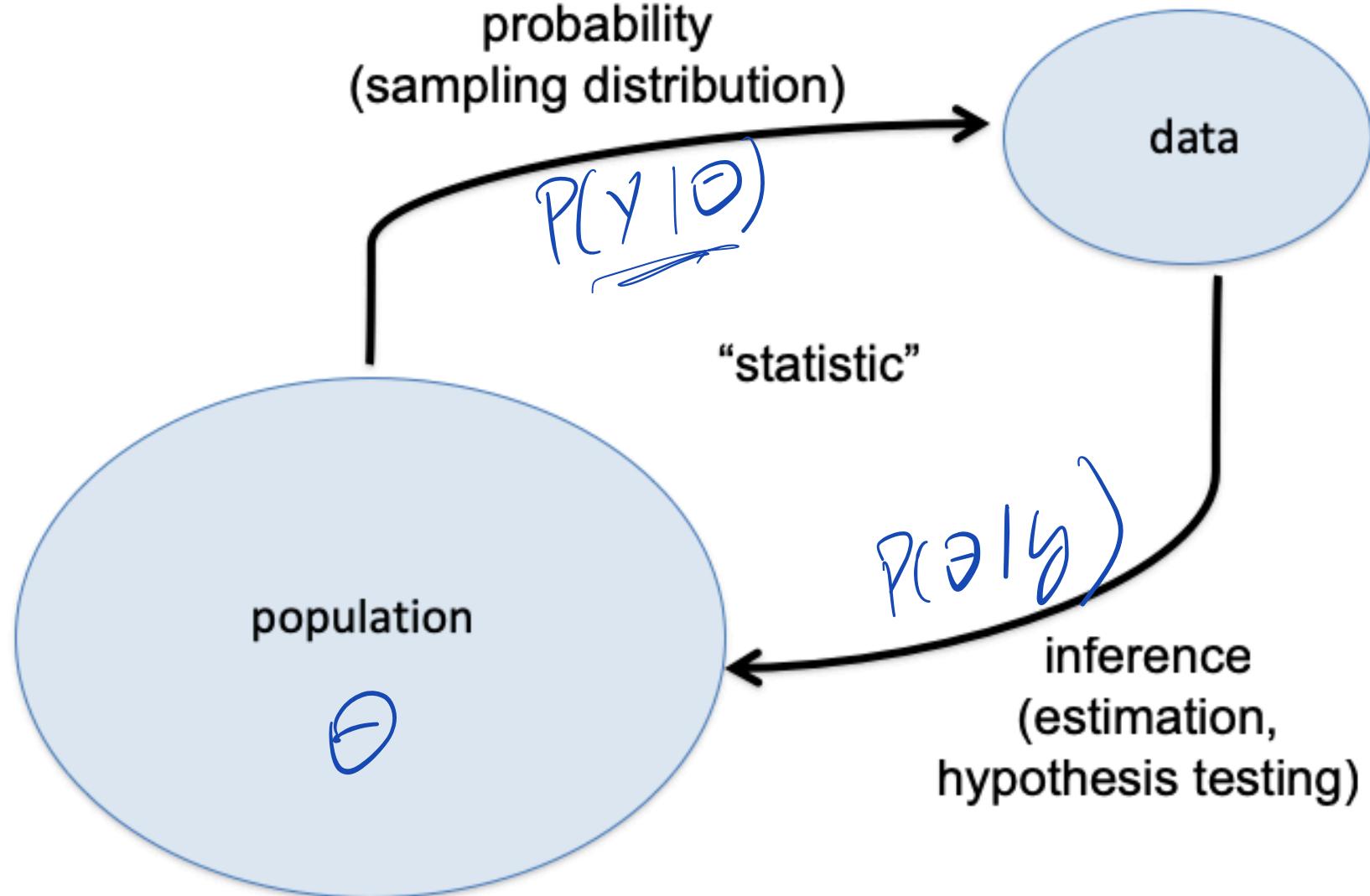
Lecture 1: Likelihood Review

Professor Alexander Franks

Logistics

- Read: BDA Chapters 1-2
- Sync content using link on course website:
~~<https://bit.ly/2NIVyk9H>~~ tinyurl.com/pstat215a
- 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 θ if $P(y_1, \dots, y_n | \theta) = \prod_i P(y_i | \theta)$
- Conditional independence means that Y_i gives no additional information about Y_j beyond that in knowing θ

$$\cancel{P(y_1, \dots, y_n)} = \prod_i P(y_i) \quad ??$$

\Rightarrow "Exchangeable":

$$\begin{array}{c} \theta \\ \diagdown \quad \diagup \\ y_1 \quad y_2 \quad \dots \quad y_n \end{array}$$
$$P(y_1, \dots, y_n) = P(y_{\sigma(1)}, \dots, y_{\sigma(n)})$$

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 $y = (y_1, y_2, \dots, y_n)$
- 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 θ if the conditional probability distribution of the Y , given the statistic $s(Y)$, does not depend on θ .

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

\bar{Y} is sufficient \Rightarrow

$$\underbrace{Y_i}_{\text{---}} \mid \bar{Y} \stackrel{iid}{\sim} N\left(\bar{y}, 1 - \frac{1}{n}\right) \quad \text{"No } \theta\text{"}$$

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

Sufficient Statistics

- Let $L(\theta) = p(y_1, \dots, y_n | \theta)$ 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:
 $\theta \notin s(y)$ only

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

No θ

$\theta \notin s(y)$ only

- 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

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

$$L(\theta; y=3) = \frac{\cancel{(12)}_3 \theta^3 (1-\theta)^9}{\cancel{(3)!}}$$

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

$$L(\theta; X=9) = \frac{\cancel{(11)}_2 \theta^3 (1-\theta)^9}{\cancel{(2)!}}$$

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

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

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

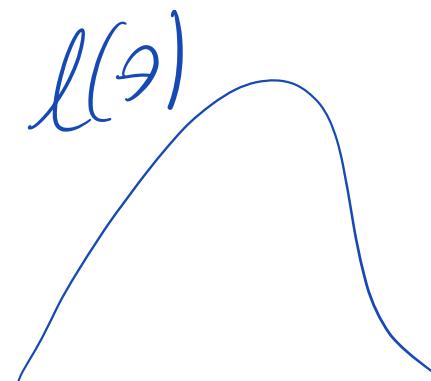
$$NB: 1 - p\text{binom}(8, 3, \theta=1/2) = .033$$

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. $\ell'(\hat{\theta}) = 0$
 - $E\left[\frac{d\ell(\theta; Y)}{d\theta} \mid \theta\right] = 0$ (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: $\underline{I(\theta)} = -E_Y\left[\frac{d^2\ell(\theta; Y)}{d^2\theta}\right]$

"How peaked/curved the likelihood is"



Fisher Information

$y_1, \dots, y_n \stackrel{\text{ird}}{\sim} N(\mu, \sigma^2)$ known value.

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

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

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

$$I(\mu) = -E_y[\ell''(\mu)] = \frac{1}{\sigma^2}$$

Cramer-Rao
Bound

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

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) ← φ chance of experienced_
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

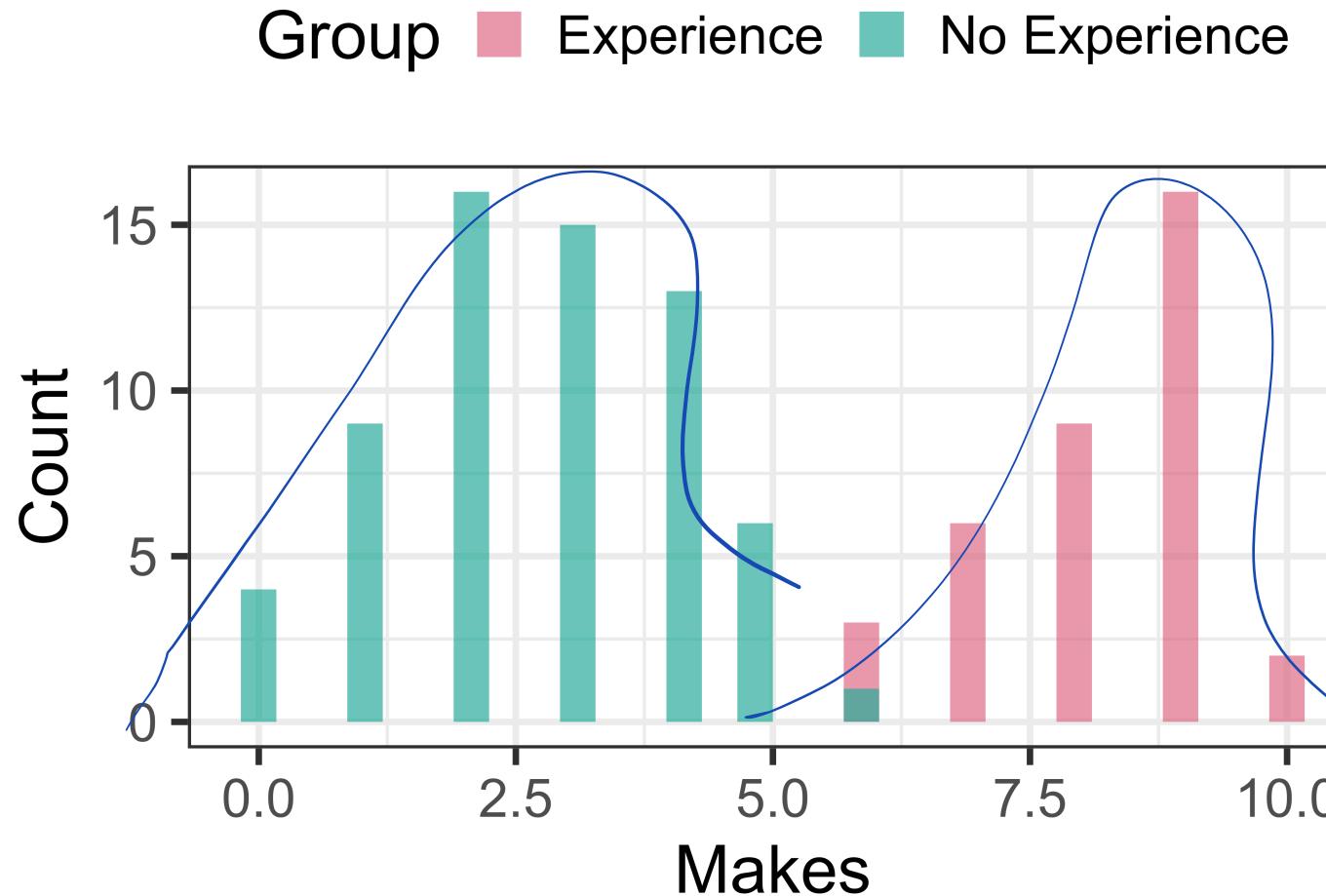
$$Z_i = \begin{cases} 0 & \text{if the } i^{\text{th}} \text{ if student doesn't play basketball} \\ 1 & \text{if the } i^{\text{th}} \text{ if student does play basketball} \end{cases}$$

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

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

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

A Mixture Model



Note: z is observed

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

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

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

$$= \phi^{\sum z_i} \theta_1^{\sum y_i z_i} (1-\theta_1)^{\sum (1-y_i) z_i} \times \\ (1-\phi)^{\sum (1-z_i)} \theta_0^{\sum y_i (1-z_i)} (1-\theta_0)^{\sum (1-y_i) (1-z_i)}$$

$$P(Y|Z, \theta_1, \theta_0, \phi) = P(Y|Z, \theta_1, \theta_0)$$

Sufficient statistics When Z_i is observed

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

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

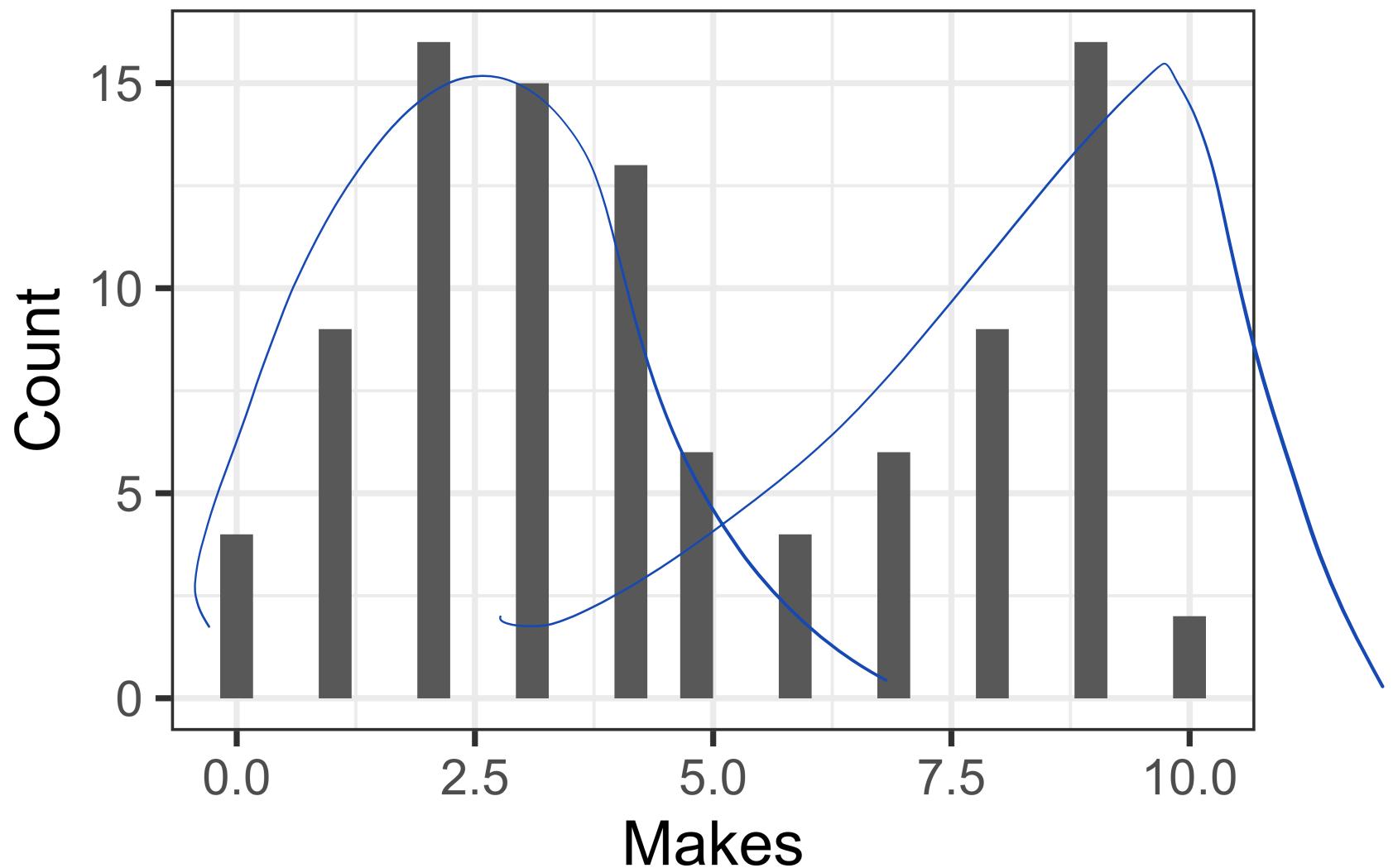
```
1 for (i in 1:100)
2   - Generate z_i from Bin(1, phi)
3   - p_i = theta_1 if z_i=1
4   - p_i = theta_0 if z_i=0
5   - Generate y_i from a Binom(10, p_i)
6 return y = (y_1, ... y_100)
```

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

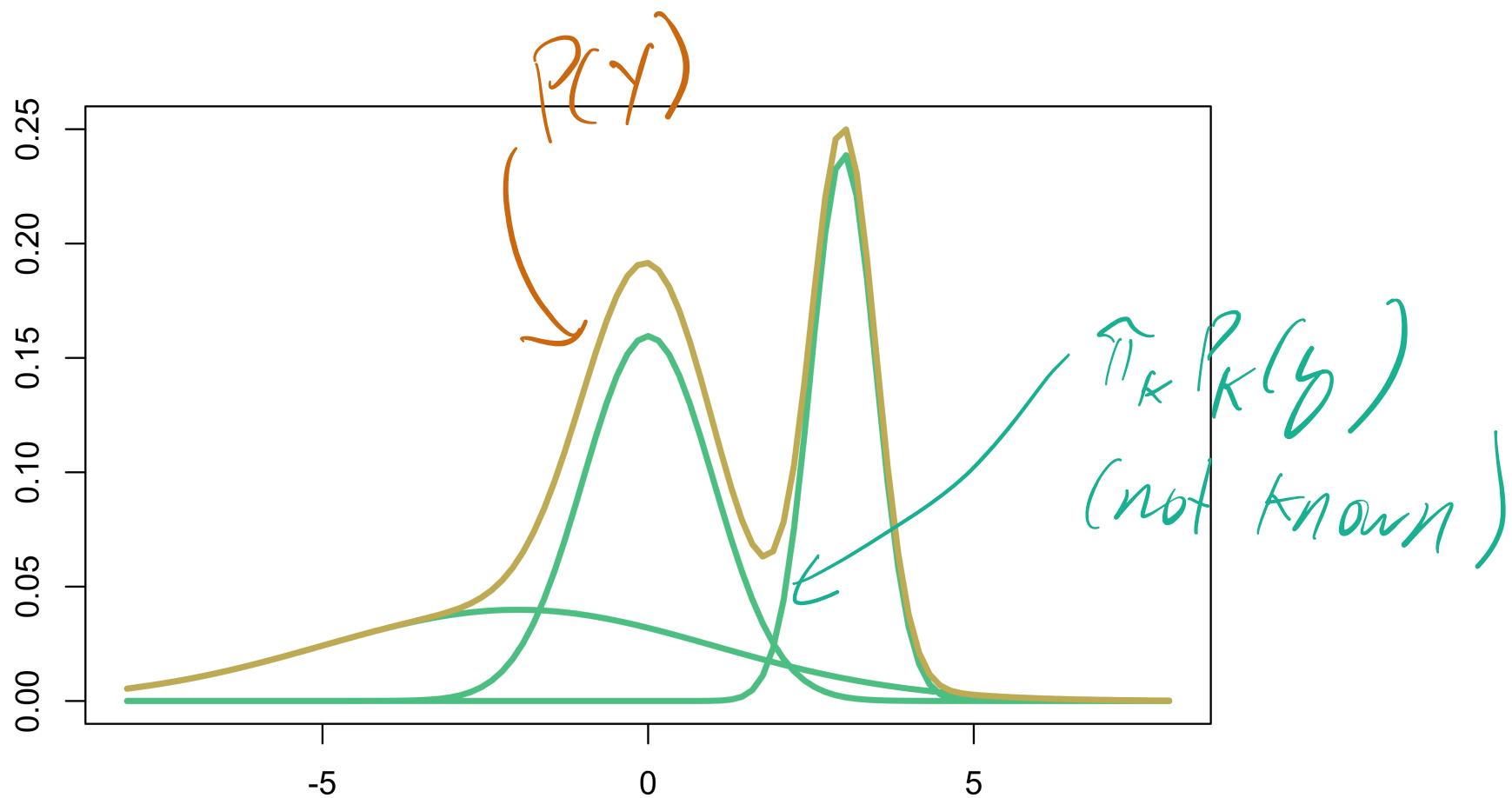
- Often crucial to understand the complete data generating process by introducing *latent* variables
- Write the *observed data likelihood* by integrating out the latent variables from the *complete data likelihood*

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

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



$$\begin{aligned}
 L(\theta_0, \theta_1, \phi) &\propto \prod_{i=1}^n P(y_i | z_i = z_i, \theta_0, \theta_1, \phi) \\
 &\propto \prod_{i=1}^n \left[\sum_{z_i=0}^1 P(y_i | z_i = z_i, \theta_0, \theta_1) P(z_i = z_i | \phi) \right] \\
 &\propto \prod_{i=1}^n \left[\frac{(10)}{y_i} \theta_1^{y_i} (1 - \theta_1)^{10 - y_i} \phi + \frac{(10)}{y_i} \theta_0^{y_i} (1 - \theta_0)^{10 - y_i} (1 - \phi) \right]
 \end{aligned}$$

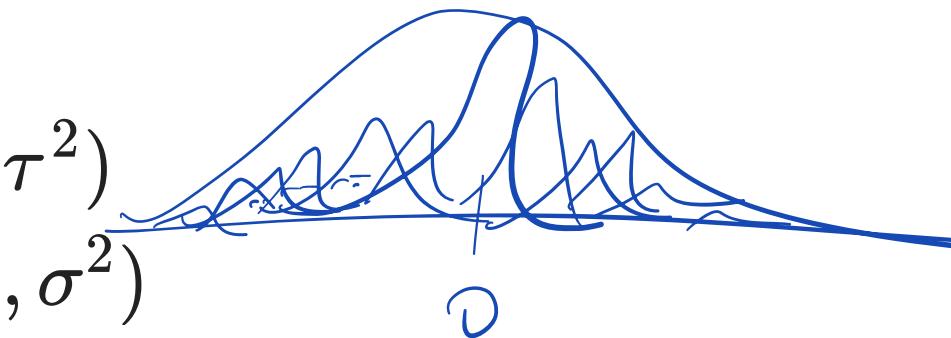
Can't Simplify

(y_1, \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 τ^2 and σ^2 (integrating over μ)?

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

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

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

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

$$\begin{cases} Y_i = \mu_i + \epsilon \\ \epsilon \sim N(0, \sigma^2) \\ \mu_i \sim N(0, \sigma^2) \end{cases}$$

Representation

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

Infinite Mixture Models

Example 2: Poisson observations with random rates

$$\left| \begin{array}{l} \lambda_i \sim \text{Gamma}(\alpha, \beta) \\ Y_i \sim \text{Pois}(\lambda_i) \end{array} \right. \quad \begin{array}{l} E[\lambda] = \frac{\alpha}{\beta} \\ \text{Var}(\lambda) = \frac{\alpha}{\beta^2} \end{array}$$

$$\begin{aligned} P(Y | \alpha, \beta) &= \int_0^\infty p(y|\lambda) p(\lambda | \alpha, \beta) d\lambda \\ &= \int_0^\infty \underbrace{\frac{\lambda^y e^{-\lambda}}{y!}}_{\text{Pois}} \underbrace{\frac{\beta^\alpha \lambda^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\lambda}}_{d\lambda} d\lambda \end{aligned}$$

A Gamma-Poisson Mixture is
a Negative-Binomial.

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

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

$$\begin{aligned} \text{Var}(Y) &= \underset{\substack{\text{Law of total variance} \\ (\text{EVVES law})}}{E[\text{Var}(Y|\lambda)] + \text{Var}(E[Y|\lambda])} \\ &= E[\lambda] + \text{Var}(\lambda) \end{aligned}$$

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

extra dispersion

Infinite Mixture Models

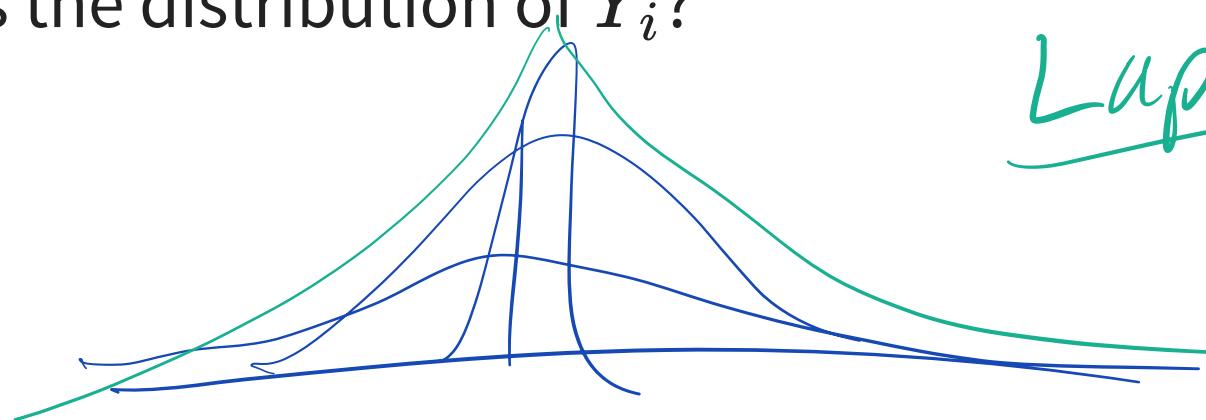
- Example 3: normal observations with exponentially distributed scale

Scale-Mixture
of
Normals

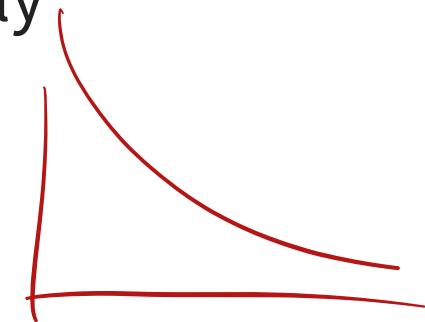
$$\sigma_i^2 \sim \text{Exponential}(1/2)$$

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

What is the distribution of Y_i ?



Laplace



Summary

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

Assignments

- Read chapter 1-2 BDA3