Lecture 1: Likelihood Review

Professor Alexander Franks

Logistics

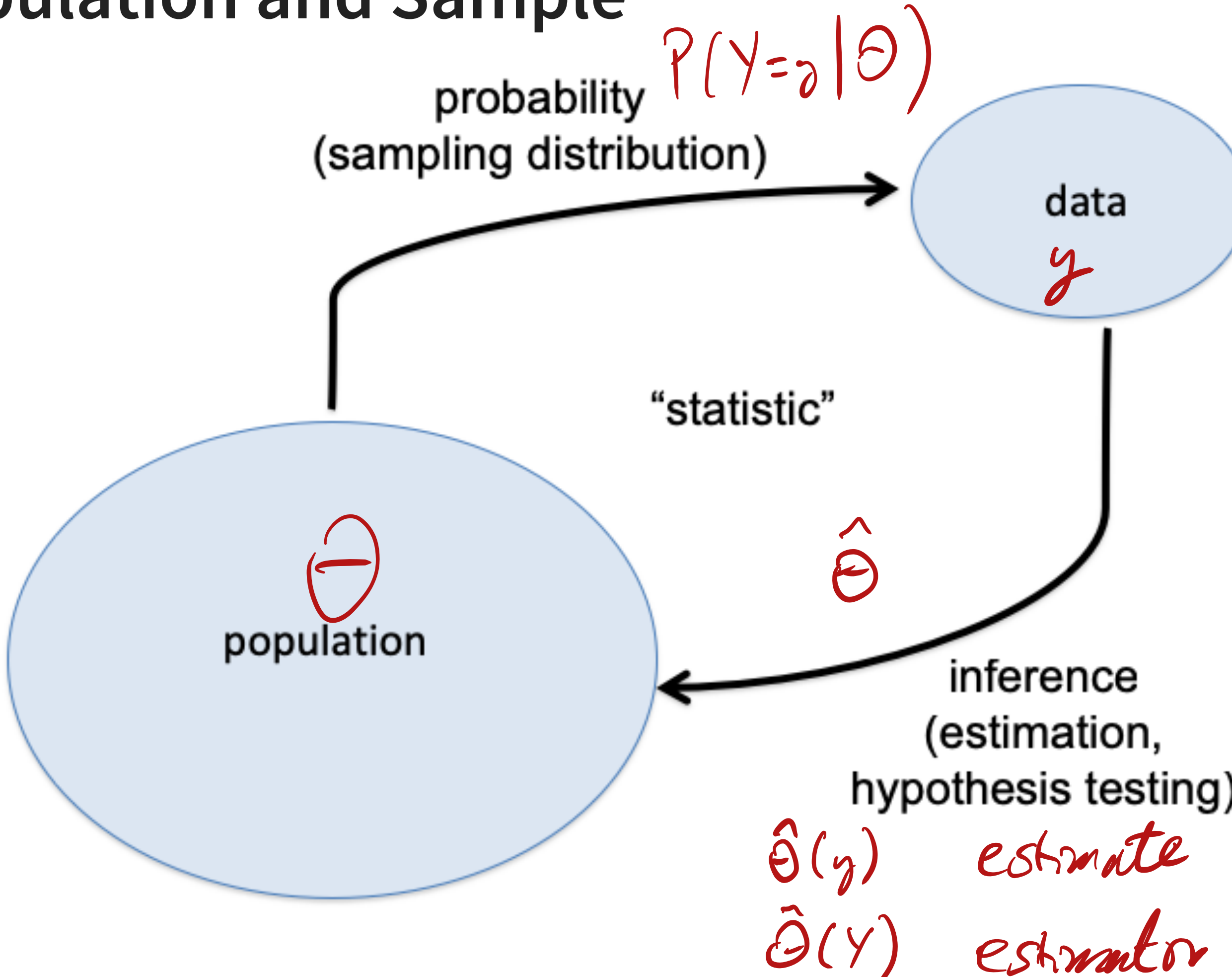
- Read: BDA Chapters 1-2

  - Nectir
  - Jupyterhub + sync Content.

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

Exchangeability

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

$$Y_i \overset{iid}{\sim} P(Y|\theta)$$

$$L(\theta) \propto \underbrace{\prod_{i}^{n} P(Y_i = y_i | \theta)}$$

Sufficient Statistics

vector $(Y_1, \dots Y_n)$

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.

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

E.g. $Y_1, \dots Y_n \mid \bar{Y} \sim N(0, 1 - 1/n)$ (exercise)

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

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:

$$L(\theta) = \underbrace{\cancel{h(y_1, \ldots y_n)}}_{\text{No } \theta}\,\underbrace{g(s(y), \theta)}_{\text{Has } \theta \text{ and sufficient stat.}}$$

  - 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

$Y \sim Bin(12, \theta)$ Obs: $y = 3$

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

$X \sim NB(3, \theta)$. Obs: $x = 9$

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

$H_0: \theta = 1/2$

$H_a: \theta < 1/2$

Bin: $\text{pbinom}(3, 12, \theta = 1/2)$ " 0.073

NB: $1 - \text{PNBinom}(8, 3, \theta = 1/2)$

Score and Fisher Information

0.033

L : likelihood

ℓ : log-likelihood

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

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

$Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ known

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

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

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

$$I(\mu) = \frac{n}{\sigma^2}$$

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

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.

```
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 w/ experience.

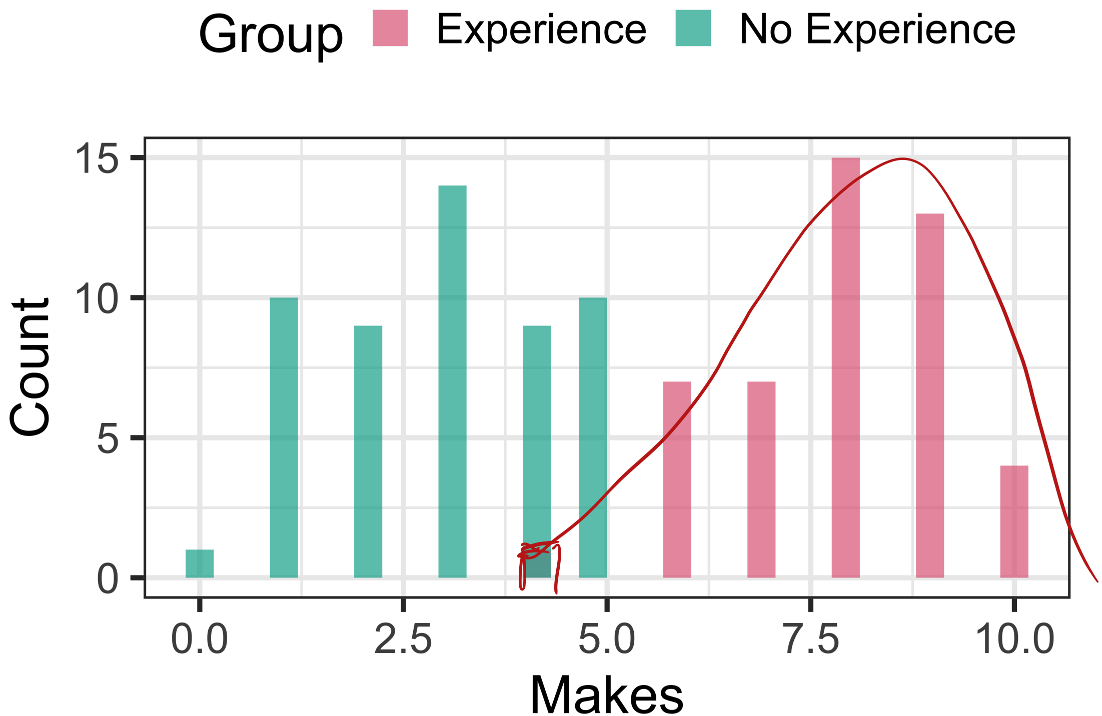
Makes (out of 10)

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

Make probabilities.

Assume $\theta_1 > \theta_0$

A Mixture Model



Note: z is observed

$$\prod_{i=1}^{n} P(Y_i, Z_i \mid \theta_1, \theta_0, \phi) = \prod_{i=1}^{n} \left[P(Y_i \mid Z_i, \theta_0, \theta_1)\, P(Z_i \mid \phi) \right]$$

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

simplify

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

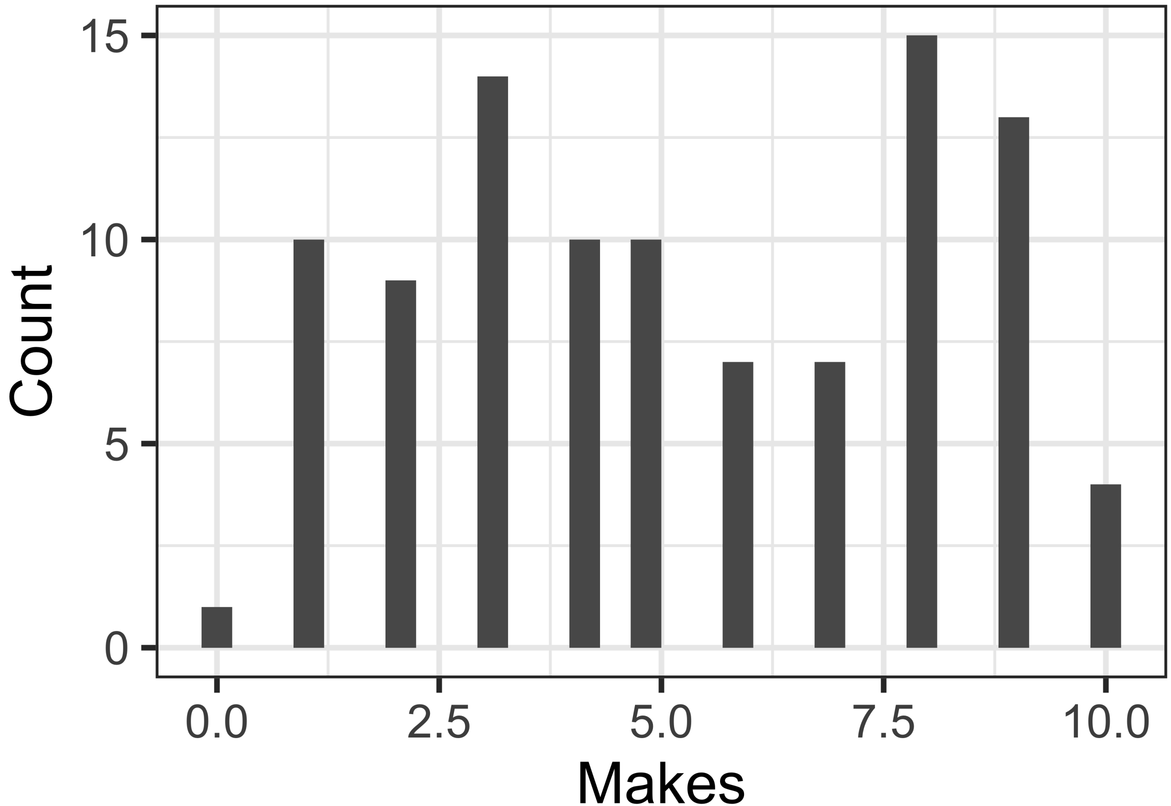
- A mixture model is a probabilistic model for representing the presence of subpopulations
- The subpopoluation to which each individual belongs is not necessarily known
  - e.g. do we ask: “have you played basketball before?”
- When z_i is not observed, we sometimes refer to it as a clustering model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

Mixture Weights

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

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

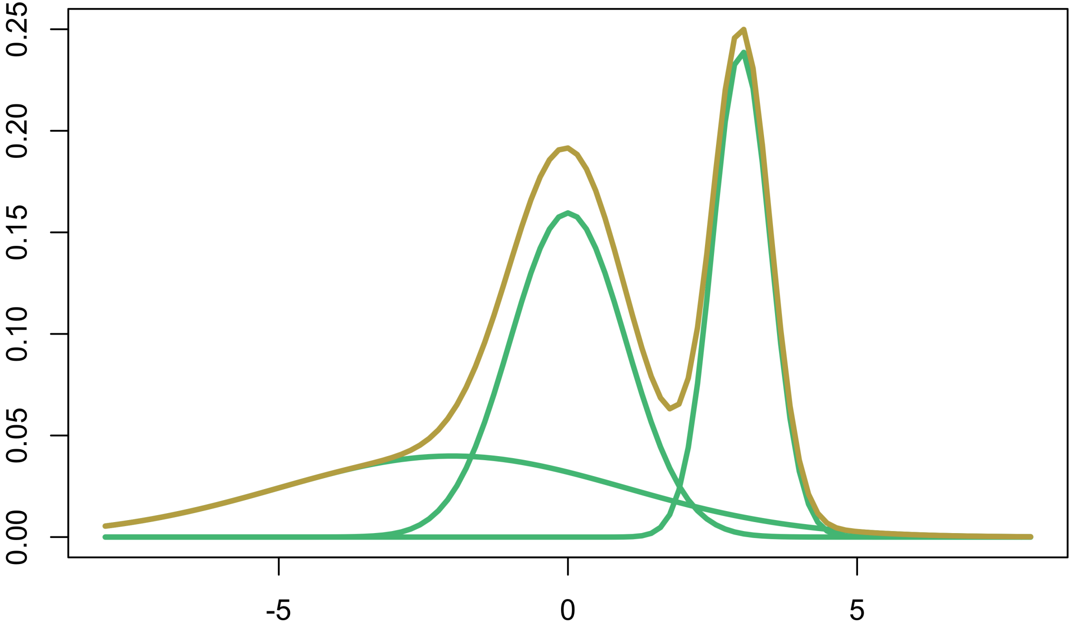
$$
\underset{P(Y|\theta)}{p(Y)} = \sum_k^K \pi_k \underset{P_k(Y|\theta)}{p_k(Y)} \text{ with } \sum \pi_k = 1
$$

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

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

$$\propto \prod_{i=1}^{100} \left[\phi \cancel{\binom{10}{y_i}} \theta_1^{y_i} (1-\theta_1)^{10-y_i} + (1-\phi) \cancel{\binom{10}{y_i}} \theta_0^{y_i} (1-\theta_0)^{10-y_i} \right]$$

Finite mixture models



Infinite Mixture Models

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

Calculus, MGFs

$$P(Y|\tau^2,\sigma^2) = \int_{-\infty}^{\infty} P(Y|\mu,\sigma^2)P(\mu|\tau^2)\,d\mu$$

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

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

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

$$\boxed{Y_i = \mu_i + \varepsilon, \quad \varepsilon \sim N(0,\sigma^2)}$$

Infinite Mixture Models

$\theta = \frac{1}{\beta} \rightarrow Var = \alpha\theta^2$

Example 2: Poisson observations with random rates

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

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

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

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

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

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

$\text{Gam}(\alpha+y, \beta+1)$

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

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

Infinite Mixture Models

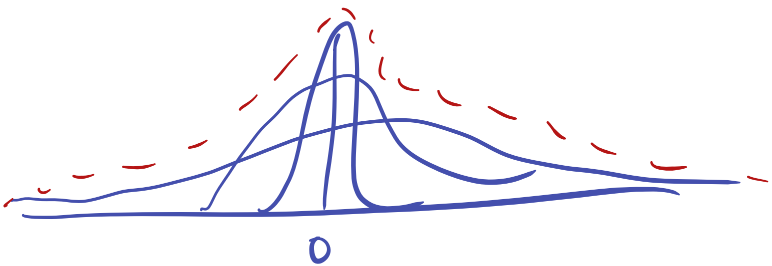
- Example 3: normal observations with exponentially distributed scale

$$\sigma_i^2 \sim Exponential(1/2)$$
$$Y_i \sim N(\cancel{\mu}\,0, \sigma_i^2)$$

Inv-Gam

What is the distribution of Y_i ~~given μ~~?

$$\frac{Z_1}{|Z_2|}$$



Summary

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

Assignments

- Read chapter 1-2 BDA3