Frequentist Data Analysis II

Prof. Pradeep Ravikumar
pradeepr@cs.cmu.edu

Bernoulli distribution

Data, D = 

- P(Heads) = θ, P(Tails) = 1-θ

- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

<u>Choose θ that maximizes the probability of observed data</u>

Probability of one coin flip

Let's say we observe a coin flip $X \in \{0, 1\}$.

The probability of this coin flip,
given a Bernoulli distribution with parameter p:

$$p^X (1 - p)^{1-X}.$$

Equal to p when $X = 1$, and equal to $(1 - p)$ when $X = 0$.

Probability of Multiple Coin Flips

$$\text{Probability of Data} = \mathbb{P}(X_1, X_2, \ldots, X_n; \theta)$$

Probability of Multiple Coin Flips

$$\begin{aligned}\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\ &= P(X_1)\,P(X_2)\ldots P(X_n)\end{aligned}$$

…Independe
nce of
samples

Probability of Multiple Coin Flips

$$\begin{aligned}\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\ &= P(X_1)\, P(X_2) \ldots P(X_n) \\ &= \prod_{i=1}^{n} P(X_i)\end{aligned}$$

Probability of Multiple Coin Flips

$$\begin{aligned}
\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\
&= P(X_1)\, P(X_2) \ldots P(X_n) \\
&= \prod_{i=1}^{n} P(X_i) \\
&= \prod_{i=1}^{n} p^{X_i}\, (1-p)^{1-X_i}
\end{aligned}$$

…probability of a Bernoulli sample

Probability of Multiple Coin Flips

$$
\begin{aligned}
\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\
&= P(X_1)\, P(X_2) \ldots P(X_n) \\
&= \prod_{i=1}^{n} P(X_i) \\
&= \prod_{i=1}^{n} p^{X_i}\, (1-p)^{1-X_i} \\
&= p^{\sum_{i=1}^{n} X_i}\, (1-p)^{n-\sum_{i=1}^{n} X_i}
\end{aligned}
$$

$$\ldots\; p^a\, p^b = p^{a+b}$$

Probability of Multiple Coin Flips

$$
\begin{aligned}
\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\
&= P(X_1)\, P(X_2) \ldots P(X_n) \\
&= \prod_{i=1}^{n} P(X_i) \\
&= \prod_{i=1}^{n} p^{X_i}\, (1-p)^{1-X_i} \\
&= p^{\sum_{i=1}^{n} X_i}\, (1-p)^{n-\sum_{i=1}^{n} X_i} \\
&= p^{n_h}\, (1-p)^{n-n_h}.
\end{aligned}
$$

where n_h is the number of heads,
n is the total number of coin flips

Maximum Likelihood Estimator (MLE)

The MLE solution is then given by solving the following problem:

$$\begin{aligned}
\widehat{p} &= \arg\max_{p} \mathbb{P}(X_1, \ldots, X_n; p) \\
&= \arg\max_{p} \left\{ p^{n_h} (1-p)^{n-n_h} \right\}
\end{aligned}$$

Maximum Likelihood Estimator (MLE)

The MLE solution is then given by solving the following problem:

$$
\begin{aligned}
\widehat{p} &= \arg\max_p \mathbb{P}(X_1, \ldots, X_n; p) \\
&= \arg\max_p \left\{p^{n_h} (1-p)^{n-n_h}\right\} \\
&= \arg\max_p \left\{n_h \log p + (n - n_h) \log(1-p)\right\}
\end{aligned}
$$

…argmax_x f(x) = argmax_x log f(x)

MLE for coin flips

The MLE solution is then given by solving the following problem:

$$\widehat{p} = \arg\max_{p} \{n_h \log p + (n - n_h)\ \log(1 - p)\}$$

$$\implies \frac{n_h}{\widehat{p}} - \frac{n - n_h}{1 - \widehat{p}} = 0$$

$$\implies \widehat{p} = \frac{n_h}{n}.$$

Maximum Likelihood Estimation

Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} = 3/5$$

"Frequency of heads"

MLE of probability of head:

How many flips do I need?

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: $\theta = 3/5$, it is the MLE!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, it is the MLE!
- **He says: If you get the same answer,** would you prefer to flip 5 times or 50 times**?**
- You say: Hmm… The more the merrier???
- He says: Is this why I am paying you the big bucks???

SO FAR:
THE MLE IS A CLASS OF ESTIMATORS THAT ESTIMATE MODEL FROM DATA

KEY QUESTION: HOW GOOD IS THE MLE (OR ANY OTHER ESTIMATOR)?

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the “true” coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the “true” coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

Do we get that $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \to p$ in probability as $n \to \infty$?

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the “true” coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

Do we get that $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \to p$ in probability as $n \to \infty$?

By the Law of Large Numbers!

How good is this MLE: Infinite Trial Average

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

How good is this MLE: Infinite Trial Average

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

Formally: the estimator $\widehat{p}$ is random: it depends on the samples (i.e. coin flips) drawn from a Bernoulli distribution with parameter p.

What would the expectation of the estimator be?

How good is this MLE: Infinite Trial Average

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

Formally: the estimator $\widehat{p}$ is random: it depends on the samples (i.e. coin flips) drawn from a Bernoulli distribution with parameter p.

What would the expectation of the estimator be?

It would be great if this expectation be equal to the "true" coin flip probability.

How good is this MLE: Infinite Trial Average

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

Formally: the estimator $\widehat{p}$ is random: it depends on the samples (i.e. coin flips) drawn from a Bernoulli distribution with parameter p.

What would the expectation of the estimator be?

It would be great if this expectation be equal to the "true" coin flip probability.

This property is called **unbiasedness**.

How good is this MLE?

It would be great if this expectation be equal to the “true” coin flip probability.

This property is called **unbiasedness**.

$$
\begin{aligned}
\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)
\end{aligned}
$$

How good is this MLE?

It would be great if this expectation be equal to the "true" coin flip probability.

This property is called **unbiasedness**.

$$\begin{aligned}\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\ &= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i)\end{aligned}$$

…linearity of expectation:

E(a X + b Y) = a E(X) + b E(Y)

How good is this MLE?

It would be great if this expectation be equal to the “true” coin flip probability.

This property is called **unbiasedness**.

$$
\begin{aligned}
\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \\
&= \mathbb{E}X_1
\end{aligned}
$$

How good is this MLE?

It would be great if this expectation be equal to the "true" coin flip probability.

This property is called **unbiasedness**.

$$
\begin{aligned}
\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \\
&= \mathbb{E}X_1 \\
&= p.
\end{aligned}
$$

Summary: Classical/Frequentist Data Analysis

Parameter θ

Observation X

θ is a deterministic (i.e. not random) but unknown quantity

X is random, with distribution

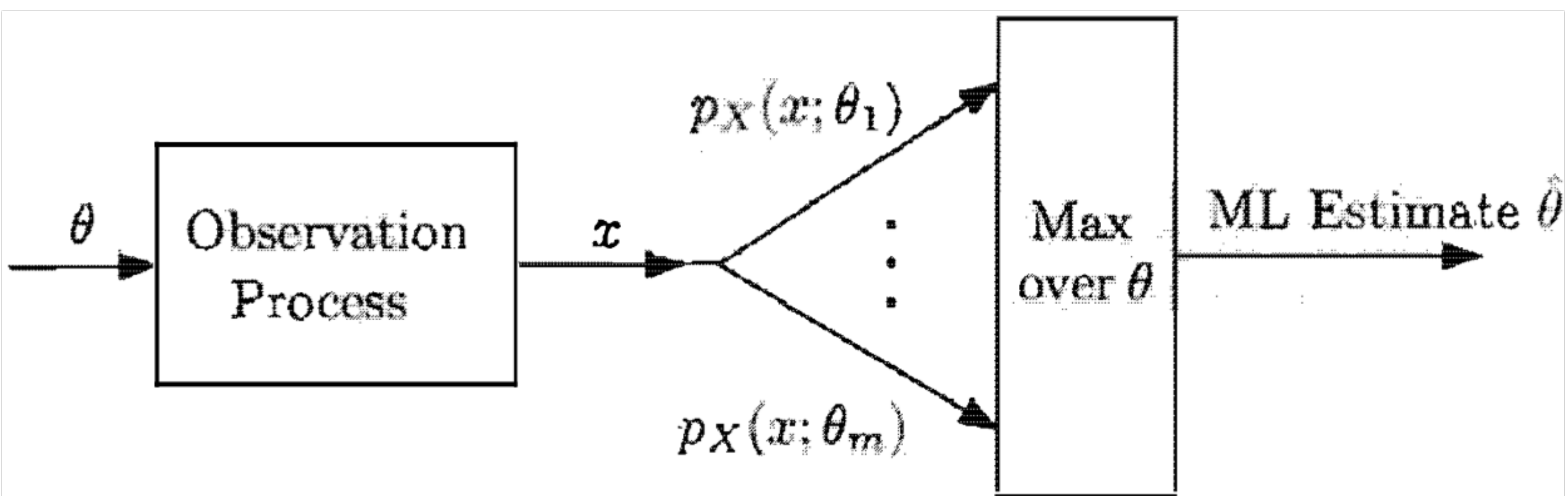
$$p_X(x;\theta) \text{ (if } X \text{ is discrete), or } f_X(x;\theta) \text{ (if } X \text{ is continuous)}$$

- These are NOT conditional probabilities;
  θ is NOT random
  - mathematically: many models,
    one for each possible value of θ

Summary: Maximum Likelihood Estimation

- Model, with unknown parameter(s): $X \sim p_X(x;\theta)$
- Pick θ that "makes data most likely"

$$\widehat{\theta}_{\mathsf{ML}} = \arg\max_{\theta} p_X(x;\theta)$$



Likelihood Function

- We refer to $p_X(x;\theta)$ [or $f_X(x;\theta)$ if X is continuous] as the **likelihood function**.
  - Note that this is a function of θ
- If the observations X_i are independent, the likelihood function takes the form

$$p_X(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} p_{X_i}(x_i;\theta)$$

- Log-likelihood function:

$$\log p_X(x_1,\ldots,x_n;\theta) = \log \prod_{i=1}^{n} p_{X_i}(x_i;\theta) = \sum_{i=1}^{n} \log p_{X_i}(x_i;\theta),$$

$$\log f_X(x_1,\ldots,x_n;\theta) = \log \prod_{i=1}^{n} f_{X_i}(x_i;\theta) = \sum_{i=1}^{n} \log f_{X_i}(x_i;\theta).$$

Log-Likelihood

- Log-likelihood function

$$\log p_X(x_1, \ldots, x_n; \theta) = \log \prod_{i=1}^{n} p_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log p_{X_i}(x_i; \theta),$$

$$\log f_X(x_1, \ldots, x_n; \theta) = \log \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log f_{X_i}(x_i; \theta).$$

- It might be analytically more convenient to maximize log-likelihood rather than likelihood --- though either would yield the same answer

$$\arg\max_{\theta}\{f(\theta)\} = \arg\max_{\theta}\{\log f(\theta)\}$$

Example

X is said to have exponential distribution with param. θ if

$$f_X(x;\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

$X_1, \ldots, X_n$: i.i.d., exponential(θ)

What is the ML Estimate of θ?

Example

X is said to have exponential distribution with param. θ if

$$f_X(x;\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

$X_1, \ldots, X_n$: i.i.d., exponential(θ)

What is the ML Estimate of θ?

$$\max_\theta \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

Example

X is said to have exponential distribution with param. θ if

$$f_X(x;\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

$X_1, \ldots, X_n$: i.i.d., exponential(θ)

What is the ML Estimate of θ?

$$\max_\theta \prod_{i=1}^n \theta e^{-\theta x_i}$$

$$\max_\theta \left(n \log \theta - \theta \sum_{i=1}^n x_i \right)$$

Example I

X is said to have exponential distribution with param. θ if

$$f_X(x;\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

$X_1, \ldots, X_n$: i.i.d., exponential(θ)

What is the ML Estimate of θ?

$$\max_\theta \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

$$\max_\theta \left(n \log \theta - \theta \sum_{i=1}^{n} x_i \right)$$

$$\boxed{\hat{\theta}_{\mathrm{ML}} = n/(x_1 + \ldots + x_n)}$$

Example I

X is said to have exponential distribution with param. θ if

$$f_X(x;\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

$X_1, \ldots, X_n$: i.i.d., exponential(θ)

What is the ML Estimate of θ?

$$\max_{\theta} \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

$$\max_{\theta} \left(n \log \theta - \theta \sum_{i=1}^{n} x_i \right)$$

$$\hat{\theta}_{\mathsf{ML}} = n/(x_1 + \ldots + x_n)$$

$$\boxed{\hat{\Theta}_n = \frac{n}{X_1 + \cdots + X_n}}$$

Desirable Properties of Estimators I

- **Unbiased:** $\mathbf{E}[\hat{\Theta}_n] = \theta$
  - exponential example, with $n = 1$:
    $\mathbf{E}[1/X_1] = \infty \neq \theta$
    (biased)

- **Bias**: $\mathbf{E}[\hat{\Theta}_n] - \theta$
  - Unbiased: bias equals zero

Example: Sample Mean

- $X_1, \ldots, X_n$: i.i.d., mean θ, variance σ^2

Example: Sample Mean

- $X_1, \ldots, X_n$: i.i.d., mean θ, variance σ^2

$X_i = \theta + W_i$

W_i: i.i.d., mean, 0, variance σ^2

Example: Sample Mean

- $X_1, \ldots, X_n$: i.i.d., mean θ, variance σ^2

$X_i = \theta + W_i$

W_i: i.i.d., mean, 0, variance σ^2

$$\hat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

Example: Sample Mean

- $X_1, \ldots, X_n$: i.i.d., mean θ, variance σ^2

$X_i = \theta + W_i$

W_i: i.i.d., mean, 0, variance σ^2

$\hat{\Theta}_n = \text{sample mean} = M_n = \dfrac{X_1 + \cdots + X_n}{n}$

Properties:

- $\mathbf{E}[\hat{\Theta}_n] = \theta$ (unbiased)

Example: Sample Mean

- $X_1, \ldots, X_n$: i.i.d., mean θ, variance σ^2

$X_i = \theta + W_i$

W_i: i.i.d., mean, 0, variance σ^2

$$\hat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

Properties:

- $\mathbf{E}[\hat{\Theta}_n] = \theta$ (unbiased)
- WLLN: $\hat{\Theta}_n \to \theta$ (consistency)

Example

- Recall the Romeo and Juliet example, where Juliet was late on any date by a random amount X, uniformly distributed over the interval [0,θ] where θ is unknown. Unlike in the previous case, let us assume here that θ is deterministic (i.e. not random). What is the ML estimate of θ, if Juliet is late by an amount x on their first date?

Example

- Recall the Romeo and Juliet example, where Juliet was late on any date by a random amount X, uniformly distributed over the interval [0,θ] where θ is unknown. Unlike in the previous case, let us assume here that θ is deterministic (i.e. not random). What is the ML estimate of θ, if Juliet is late by an amount x on their first date?

Data PDF

$$f_X(x;\theta) = \begin{cases} 1/\theta, & \text{if } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Example

- Recall the Romeo and Juliet example, where Juliet was late on any date by a random amount X, uniformly distributed over the interval [0,θ] where θ is unknown. Unlike in the previous case, let us assume here that θ is deterministic (i.e. not random). What is the ML estimate of θ, if Juliet is late by an amount x on their first date?

Data PDF

$$f_X(x;\theta) = \begin{cases} 1/\theta, & \text{if } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

ML estimate of θ is x