etherpad.wikimedia.org/p/312 for (anonymous) questions/comments!

Maximum Likelihood Estimators (MLE) *(cont.)*

CSE 312 24Su

Lecture 21

Logistics

> Concept check 19 will be due with today’s concept check 20

> Concept check 18 late due date tonight

> Keep an eye out for final exam post, HW6 post, and video walkthrough of the elevator problem tonight/tomorrow morning!

Up till now…

So far, the probability questions we've asked have followed a pattern:

You're given a model with the probabilities you need to make predictions.

> $X \sim \mathrm{Bin}(n, p)$, compute some probabilities about X, compute $E[X]$

> We have a distribution that takes on these outcomes with these probabilities. Compute the probability of some event or set of outcomes.

> *<u>Before we run the entire experiment</u>, let's make some predictions*

In real world, we usually don't know all the rules of a random experiment
hence tail bounds, CLT, etc. to estimate probabilities in these situations
But, can we estimate those missing rules/parameters to a distribution?

Maximum Likelihood Estimation

We derive an estimate $\widehat{\boldsymbol{\theta}}$ for the parameter $\boldsymbol{\theta}$ based on observed data

1. We're going to run the random experiment a bunch of times (i.e., collect a bunch of samples from the distribution) -> this gives *data*

e.g., we flip a coin that follows $\mathrm{Ber}(p)$ *10 times and write down the results – HTTTH…*

2. Estimate the missing rules (unknown parameter(s)) *based on the data*

How do we do this? Well, we got some data - High probability events happen more often than low probability events. So, ***guess the rules that maximize the probability of the events we saw*** (relative to other choices of the rules).

e.g., what is the value of p *that makes the probability of seeing HTTTH… the highest?*

To do this, we will define a function that will tell us the probability of seeing particular data (a particular set of samples from the distribution) based on a particular value of the unknown parameter(s) θ

Likelihood function

Remember, our goal:
1. Collect some data/samples from the distribution
2. Find an estimate for θ

$\boldsymbol{\mathcal{L}(E;\theta)}$ is $\mathbb{P}(E)$ when the experiment is run with θ

"what is probability of seeing the event E (in our case, the set of data), if the experiment is run with the parameter θ?"

We can't use probability notation because likelihood doesn't follow the same rules

Coin example

We ran the experiment 10 times independently. The result was HTTTHHTHHH

$\mathcal{L}(\textbf{HTTTHHTHHH};\boldsymbol{\theta}) = \theta^6(1-\theta)^4$ (multiply because independent)

"Probability of observing HTTTHHTHHH ***if θ** is probability of heads on a single flip"*

Likelihood Function: Likelihood of n observations (from discrete distribution)

$$\mathcal{L}(x_1, x_2, \dots, x_n;\theta) = \Pi_{i=1}^{n}\mathbb{P}(x_i;\theta)$$

Maximum Likelihood Estimation

We will choose the estimator $\hat{\theta} = \mathrm{argmax}_{\theta}\ \mathcal{L}(E;\theta)$

“the value of θ that makes the likelihood of seeing the observed data the highest”

Maximum Likelihood Estimator

The maximum likelihood estimator of the parameter θ is:

$$\hat{\theta} = \mathrm{argmax}_{\theta}\ \mathcal{L}(E;\theta)$$

θ is a variable, $\hat{\theta}$ is a number (or formula given the event).

Use $\hat{\theta}_{\mathrm{MLE}}$ if we want to emphasize how we found the estimator.

Remember, our goal:
1. Collect some data/samples from the distribution
2. Find an estimate for θ

Maximum Likelihood Estimator

The maximum likelihood estimator of the parameter θ is: $\hat{\theta} = \operatorname{argmax}_{\theta}\ \mathcal{L}(E;\theta)$

Coin example (goal: estimate $\boldsymbol{\theta} = \boldsymbol{p}$, the probability of heads on a flip)

We ran the experiment 10 times independently. The result was HTTTHHTHHH

$\mathcal{L}(\mathbf{HTTTHHTHHH};\boldsymbol{\theta}) = \theta^6(1-\theta)^4$ (multiply because independent)

Now, find the value of θ that maximizes the likelihood.

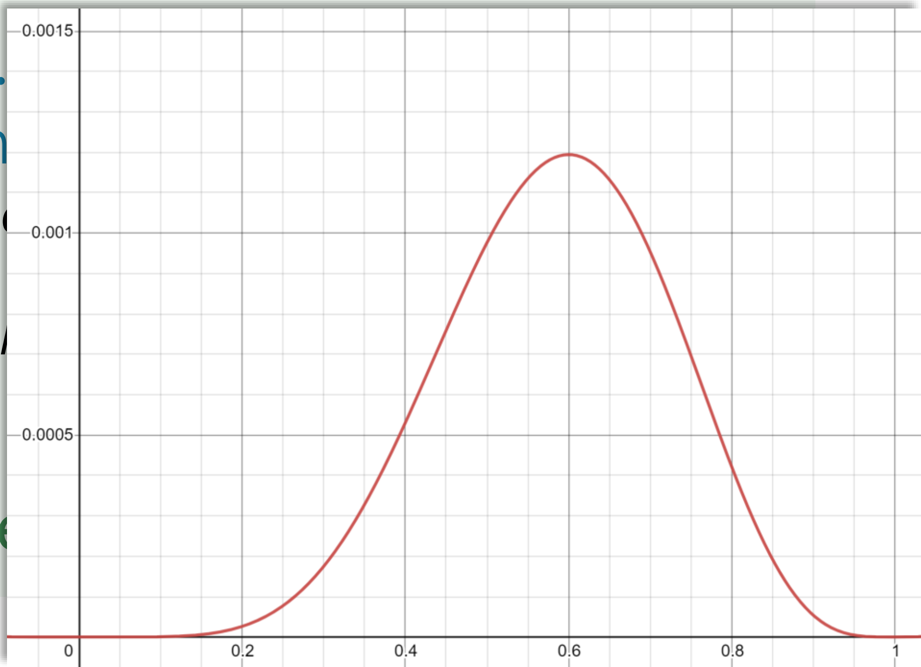
Calculus!! 🥳 Take the derivative of $\mathcal{L}(\mathrm{E};\theta)$, set to 0, an

Take the derivative: $\frac{d}{d\theta}\theta^6(1-\theta)^4 = 6\theta^5(1-\theta)^4 - 4\theta$

Set to 0 and solve: (now, we're solving for the *maximu*

$6\hat{\theta}^5\left(1-\hat{\theta}\right)^4 - 4\hat{\theta}^6\left(1-\hat{\theta}\right)^3 = 0 \Rightarrow 6\left(1-\hat{\theta}\right) - 4\hat{\theta} = 0$

The MLE $\hat{\theta}$ estimating the true $\boldsymbol{\theta} = \boldsymbol{p}$ is $\mathbf{3/5}$ just like we



Is that really the maximum?

What we really did was find the critical point (which could either be the maximum **or** the minimum), so ideally do **second derivative test** to check

1. Take the *second* derivative (the derivative of the derivative)

2. If negative everywhere around the critical point, it is the maximum

In this class, we won't ask you to do the second derivative test, you can assume the solution you find is a maximum ☺

> to sanity check your answer, at least make sure that the estimator you find is valid for what you are trying to estimate

Our MLE process so far…

We're given that there's a distribution with some unknown parameter(s) θ.
There are independent observations $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}$ from this distribution.

To find the MLE $\hat{\theta}$ for this unknown parameter(s) θ….

1. Write the likelihood function - $\mathcal{L}(x_1, x_2, \ldots, x_n; \theta) = \Pi_{i=1}^{n} \mathbb{P}(x_i; \theta)$
   multiply (not add) probabilities of seeing each of the observations based on θ

2. Take the derivative of the log-likelihood function - $\frac{d}{d\theta}(\mathcal{L}(x_1, x_2, \ldots, x_n; \theta)$

3. Set the derivative to 0, and solve for the MLE $\widehat{\boldsymbol{\theta}}$
   remember to switch from θ to $\hat{\theta}$ in this step because we're now solving for the MLE

4. Verify it is a maximum with second derivative test (not required for 312)

Half a step backwards…

Since the likelihood function is a product of probabilities of seeing each of the samples, we’re going to be taking the derivative of products a lot

The product rule is not fun!! There has to be a better way!

Half a step backwards…

Since the likelihood function is a product of probabilities of seeing each of the samples, we're going to be taking the derivative of products a lot

The product rule is not fun!! There has to be a better way!

Take the log of the likelihood function before taking the derivative!

Recall: $\ln(a \cdot b) = \ln(a) + \ln(b)$

And, we don't need the product rule if our expression is a sum!

Can we still take the max? Yes! $\ln()$ is an increasing function, so

$\text{argmax}_\theta \ln(\mathcal{L}(E;\theta)) = \text{argmax}_\theta \mathcal{L}(E;\theta)$

"the log of the likelihood will increase as the likelihood increases and vice verse, so, the value of θ that maximizes the log likelihood also maximized the likelihood"

Half a step backwards…

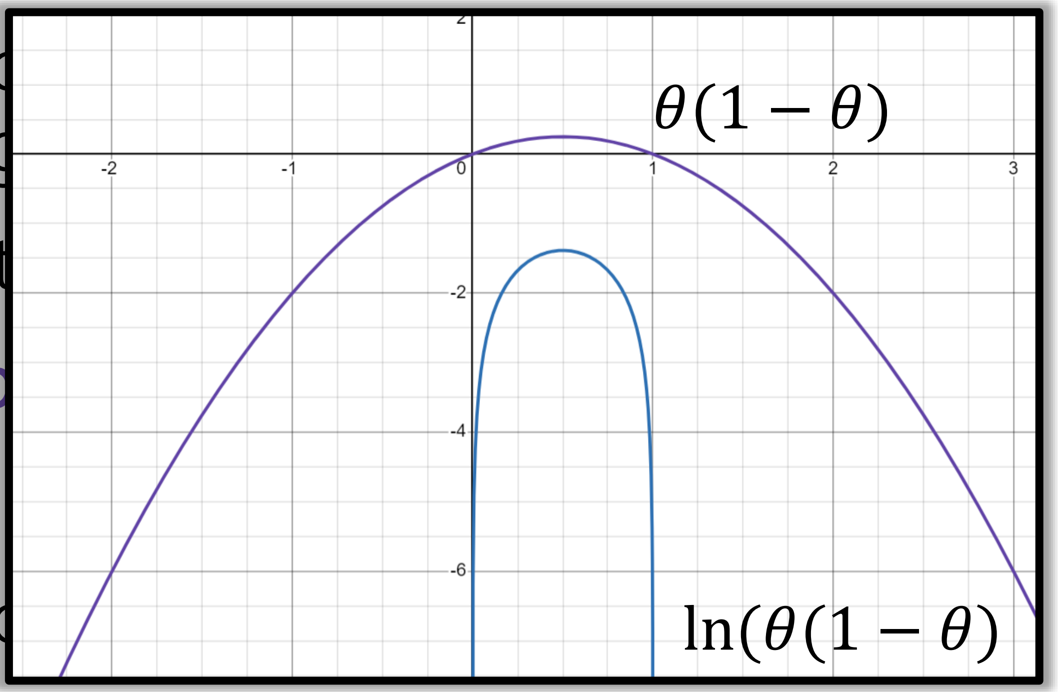
Since the likelihood function is a produc
of the samples, we're going to be takin

The product rule is not fun!! There has t

Take the log of the likelihood function b

Recall: $\ln(a \cdot b) = \ln(a) + \ln(b)$

And, we don't need the product rule if c



Can we still take the max? Yes! $\ln()$ is an increasing function, so

$$\text{argmax}_\theta \ln(\mathcal{L}(E;\theta)) = \text{argmax}_\theta \mathcal{L}(E;\theta)$$

"the log of the likelihood will increase as the likelihood increases and vice verse, so, the value of θ that maximizes the log likelihood also maximized the likelihood"

Coin flips is easier

1. Likelihood function: $\mathcal{L}(\text{HTTTHHTHHH};\theta) = \theta^6(1-\theta)^4$

2. Take the log: $\ln(\mathcal{L}(\text{HTTTHHTHHH};\theta) = 6\ln(\theta) + 4\ln(1-\theta)$

3. Take the derivative: $\frac{d}{d\theta}\ln\big(\mathcal{L}(\cdot)\big) = \frac{6}{\theta} - \frac{4}{1-\theta}$

Derivative is much easier!!

4. Set to $\mathbf{0}$ and solve:

$$\frac{6}{\widehat{\theta}} - \frac{4}{1-\widehat{\theta}} = 0 \Rightarrow \frac{6}{\widehat{\theta}} = \frac{4}{1-\widehat{\theta}} \Rightarrow 6 - 6\hat{\theta} = 4\hat{\theta} \Rightarrow \hat{\theta} = \frac{3}{5}$$

5. Check it's a maximum (can skip in 312)

$\frac{d^2}{d\theta^2} = \frac{-6}{\theta^2} - \frac{4}{(1-\theta)^2} < 0$ everywhere, so any critical point is a maximum.

Solving MLE (the process)

We're given that there's a distribution with some unknown parameter(s) θ. There are independent observations $x_1, x_2, \ldots, x_n$ from this distribution.

To find the MLE $\hat{\theta}$ for this unknown parameter(s) θ....

1. Write the likelihood function

   multiply (not add) probabilities of seeing each of the observations based on θ

2. Take the log $\boldsymbol{ln}(..)$ of the likelihood function (makes the math easier)

   use log rules and simplify fully, as much as you can, to make the math easier later

3. Take the derivative of the log-likelihood function

4. Set the derivative to 0, and solve for the MLE $\boldsymbol{\hat{\theta}}$

   remember to switch from θ to $\hat{\theta}$ in this step because we're now solving for the MLE

5. Verify it is a maximum with second derivative test (not required for 312)

Important Log Rules

- $\ln(a \cdot b) = \ln(a) + \ln(b)$
- $\ln\left(a^{b}\right) = \mathrm{b} \cdot \ln(a)$
- $\ln(a - b) = \frac{\ln(a)}{\ln(b)}$
- $\ln(e^{a}) = a$
- $\frac{d}{d\theta}\ln(m) = \frac{1}{\mathrm{m}} \cdot \frac{d}{d\theta}(m)$

MLEs with *Continuous* Random Variables

What about continuous random variables?

Can’t use probability, since the probability is going to be 0.

Likelihood Function: Likelihood of n observations (from *continuous* distribution)

$$\mathcal{L}(x_1, x_2, \ldots, x_n; \theta) = \Pi_i^n$$

What about continuous random variables?

Can’t use probability, since the probability is going to be 0.

Can use the density!

It’s supposed to show relative chances, that’s all we’re trying to find anyway.

Likelihood Function: Likelihood of n observations (from *continuous* distribution)

$$\mathcal{L}(x_1, x_2, \dots, x_n; \theta) = \Pi_i^n f_X(x_i; \theta)$$

All other steps are exactly the same!

Example: MLE on Continuous Normal

Suppose you get values $x_1, x_2, \ldots x_n$ from independent draws of a normal random variable $\mathcal{N}(\mu, 1)$ (for an unknown μ). Find the MLE for μ.

We'll also call these "realizations" of the random variable.

Example: MLE on Continuous Normal

Suppose you get values $x_1, x_2, \ldots x_n$ from independent draws of a normal random variable $\mathcal{N}(\mu, 1)$ (for an unknown μ)

We'll also call these "realizations" of the random variable.

1. Write the likelihood function: $\mathcal{L}(x_i; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{1}{2}(x_i-\mu)^2\right)}$

Example: MLE on Continuous Normal

Suppose you get values $x_1, x_2, \ldots x_n$ from independent draws of a normal random variable $\mathcal{N}(\mu, 1)$ (for an unknown μ)

We'll also call these "realizations" of the random variable.

1. Write the likelihood function: $\mathcal{L}(x_i; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{1}{2}(x_i-\mu)^2\right)}$

2. Take the log of the likelihood $\ln(\mathcal{L}(x_i; \mu)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2$

$$\begin{aligned}
\ln\big(\mathcal{L}(x_i; \mu)\big) &= \textstyle\sum_{i=1}^{n} \ln(\frac{1}{\sqrt{2\pi}} e^{\left(-\frac{1}{2}(x_i-\mu)^2\right)}) \\
&= \textstyle\sum_{i=1}^{n} \ln(\frac{1}{\sqrt{2\pi}}) + \ln(e^{\left(-\frac{1}{2}(x_i-\mu)^2\right)}) \\
&= \textstyle\sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2
\end{aligned}$$

Example: MLE on Continuous Normal

Suppose you get values $x_1, x_2, \dots x_n$ from independent draws of a normal random variable $\mathcal{N}(\mu, 1)$ (for an unknown μ)

We'll also call these "realizations" of the random variable.

1. Write the likelihood function: $\mathcal{L}(x_i; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{1}{2}(x_i-\mu)^2\right)}$

2. Take the log of the likelihood $\ln(\mathcal{L}(x_i; \mu)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2$

3. Take the derivative: $\frac{d}{d\mu}\ln(\mathcal{L}(x_i; \mu)) = \sum_{i=1}^{n} x_i - \mu$

$$\frac{d}{d\mu}\ln(\mathcal{L}(x_i; \mu)) = \frac{d}{d\mu}\sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2 = \sum_{i=1}^{n}\left(\frac{d}{d\mu}\left(\ln\left(\frac{1}{\sqrt{2\pi}}\right)\right) - \frac{d}{d\mu}\left(\frac{1}{2}d(x_i - \mu)^2\right)\right)$$
$$= \sum_{i=1}^{n} -\left(\frac{1}{2}\cdot 2 \cdot -1(x_i - \mu)\right) = \sum_{i=1}^{n} x_i - \mu$$

Example: MLE on Continuous Normal

Suppose you get values $x_1, x_2, \ldots x_n$ from independent draws of a normal random variable $\mathcal{N}(\mu, 1)$ (for an unknown μ)

We'll also call these "realizations" of the random variable.

1. Write the likelihood function: $\mathcal{L}(x_i;\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{1}{2}(x_i-\mu)^2\right)}$

2. Take the log of the likelihood $\ln(\mathcal{L}(x_i;\mu)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2$

3. Take the derivative: $\frac{d}{d\mu}\ln(\mathcal{L}(x_i;\mu)) = \sum_{i=1}^{n} x_i - \mu$

4. Set to 0 and solve: $\sum_{i=1}^{n} x_i - \hat{\mu} = 0 \Rightarrow \sum_{i=1}^{n} x_i = \hat{\mu} \cdot n \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$

Example: MLE on Continuous Normal

Suppose you get values $x_1, x_2, \ldots x_n$ from independent draws of a normal random variable $\mathcal{N}(\mu, 1)$ (for an unknown μ)

We'll also call these "realizations" of the random variable.

1. Write the likelihood function: $\mathcal{L}(x_i; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{1}{2}(x_i - \mu)^2\right)}$

2. Take the log of the likelihood $\ln(\mathcal{L}(x_i; \mu)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2$

3. Take the derivative: $\frac{d}{d\mu} \ln(\mathcal{L}(x_i; \mu)) = \sum_{i=1}^{n} x_i - \mu$

4. Set to 0 and solve: $\sum_{i=1}^{n} x_i - \hat{\mu} = 0 \Rightarrow \sum_{i=1}^{n} x_i = \hat{\mu} \cdot n \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$

5. Second derivative test: $\frac{d^2}{d\mu^2} \ln(\mathcal{L}) = -n$. Second derivative is negative everywhere, so log-likelihood is concave down and this is a maximizer.

MLEs with *Multiple* Parameters

Solving MLE with 2 parameters

There's a distribution with some unknown parameters θ_1, θ_2. There are independent observations $x_1, x_2, \dots, x_n$ from this distribution.

To find the MLEs $\widehat{\theta_1}$ and $\widehat{\theta_2}$

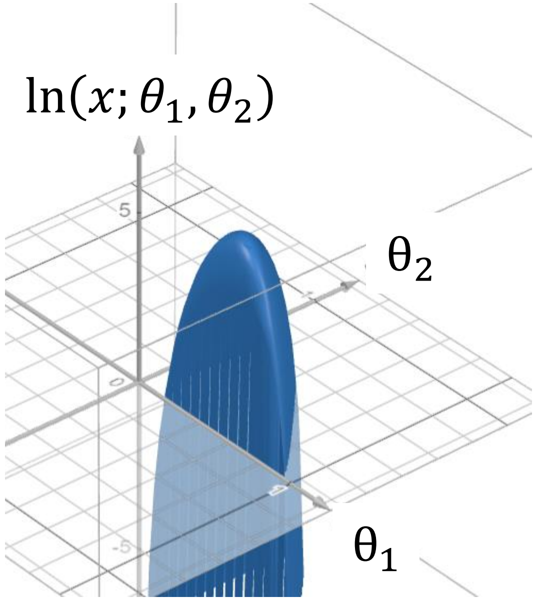
1. Write the likelihood function: $\mathcal{L}(x_1, \dots, x_n; \theta_1, \theta_2)$

   the likelihood function will now be in terms of θ_1 and θ_2

2. Take the log $\boldsymbol{ln}(..)$ of the likelihood function

3. Take the derivatives of the log-likelihood function

   There are 2 parameters, so take the partial derivative with respect to θ_1 and the partial derivative with respect to θ_2



4. Set the derivatives to 0, and solve for the MLE $\widehat{\boldsymbol{\theta}_1}$, $\widehat{\boldsymbol{\theta}_2}$

   remember to switch from θ to $\hat{\theta}$ in this step because we're now solving for the MLE

5. Verify it is a maximum with second derivative test (not required for 312)

Example: Generalizing Normals

Let θ_μ and θ_{σ^2} be the unknown mean and variance of a normal distribution. We get independent draws $x_1, x_2, \ldots, x_n$ from the distribution. Find the MLEs for $\theta_\mu$and θ_{σ^2}.

1. Likelihood function: $\mathcal{L}\left(x_1, \ldots, x_n; \theta_\mu, \theta_{\sigma^2}\right) =$

Example: Generalizing Normals

Let θ_μ and θ_{σ^2} be the unknown mean and variance of a normal distribution. We get independent draws $x_1, x_2, \ldots, x_n$ from the distribution.

1. Likelihood function: $\mathcal{L}\left(x_1, \ldots, x_n; \theta_\mu, \theta_{\sigma^2}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{\theta_{\sigma^2} 2\pi}} \mathrm{e}^{\left(-\frac{1}{2} \cdot \frac{\left(x_i - \theta_\mu\right)^2}{\theta_{\sigma 2}}\right)}$

Example: Generalizing Normals

Let θ_μ and θ_{σ^2} be the unknown mean and variance of a normal distribution. We get independent draws $x_1, x_2, \ldots, x_n$ from the distribution.

1. Likelihood function: $\mathcal{L}\left(x_1, \ldots, x_n; \theta_\mu, \theta_{\sigma^2}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{\theta_{\sigma 2} 2\pi}} \mathrm{e}^{\left(-\frac{1}{2} \cdot \frac{\left(x_i - \theta_\mu\right)^2}{\theta_{\sigma 2}}\right)}$

2. Log-likelihood: $\ln\left(\mathcal{L}\left(x_i; \theta_\mu, \theta_{\sigma^2}\right)\right) = -\frac{n}{2}\ln(\theta_{\sigma^2}) - \frac{n \cdot \ln(2\pi)}{2} - \frac{1}{2\theta_{\sigma 2}} \sum_{i=1}^{n} \left(x_i - \theta_\mu\right)^2$

$$\begin{aligned}
\ln\left(\mathcal{L}\left(x_i; \theta_\mu, \theta_{\sigma^2}\right)\right) &= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{\theta_{\sigma 2} 2\pi}}\right) - \frac{1}{2} \cdot \frac{\left(x_i - \theta_\mu\right)^2}{\theta_{\sigma 2}} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\ln(\theta_{\sigma^2}) - \frac{1}{2}\ln(2\pi) - \frac{1}{2} \cdot \frac{\left(x_i - \theta_\mu\right)^2}{\theta_{\sigma 2}} \\
&= -\frac{n}{2}\ln(\theta_{\sigma^2}) - \frac{n \cdot \ln(2\pi)}{2} - \frac{1}{2\theta_{\sigma 2}} \sum_{i=1}^{n} \left(x_i - \theta_\mu\right)^2
\end{aligned}$$

Example: Generalizing Normals

Let θ_μ and θ_{σ^2} be the unknown mean and variance of a normal distribution. We get independent draws $x_1, x_2, \dots, x_n$ from the distribution.

1. Likelihood function: $\mathcal{L}(x_1, \dots, x_n; \theta_\mu, \theta_{\sigma^2}) = \prod_{i=1}^{n} \frac{1}{\sqrt{\theta_{\sigma 2} 2\pi}} \mathrm{e}^{\left(-\frac{1}{2} \cdot \frac{(x_i - \theta_\mu)^2}{\theta_{\sigma 2}}\right)}$

2. Log-likelihood: $\ln\left(\mathcal{L}(x_i; \theta_\mu, \theta_{\sigma^2})\right) = -\frac{n}{2}\ln(\theta_{\sigma^2}) - \frac{n \cdot \ln(2\pi)}{2} - \frac{1}{2\theta_{\sigma 2}} \sum_{i=1}^{n} (x_i - \theta_\mu)^2$

3. Take the derivatives:

Partial derivative w.r.t θ_μ: $\frac{\partial}{\partial \theta_\mu} \ln(\mathcal{L}(..)) =$

see next slide…

Example: Generalizing Normals

Let θ_μ and θ_{σ^2} be the unknown mean and variance of a normal distribution. We get independent draws $x_1, x_2, \dots, x_n$ from the distribution.

1. Likelihood function: $\mathcal{L}(x_1, \dots, x_n; \theta_\mu, \theta_{\sigma^2}) = \prod_{i=1}^{n} \frac{1}{\sqrt{\theta_{\sigma 2} 2\pi}} \mathrm{e}^{\left(-\frac{1}{2} \cdot \frac{(x_i - \theta_\mu)^2}{\theta_{\sigma 2}}\right)}$

2. Log-likelihood: $\ln\left(\mathcal{L}(x_i; \theta_\mu, \theta_{\sigma^2})\right) = -\frac{n}{2}\ln(\theta_{\sigma^2}) - \frac{n \cdot \ln(2\pi)}{2} - \frac{1}{2\theta_{\sigma 2}} \sum_{i=1}^{n} (x_i - \theta_\mu)^2$

3. Take the derivatives:

   Partial derivative w.r.t θ_{σ^2}: $\frac{\partial}{\partial \theta_{\sigma 2}} \ln(\mathcal{L}(..)) =$

see next slide…

Example: Generalizing Normals

Let θ_μ and θ_{σ^2} be the unknown mean and variance of a normal distribution. We get independent draws $x_1, x_2, \dots, x_n$ from the distribution.

1. Likelihood function: $\mathcal{L}(x_1, \dots, x_n; \theta_\mu, \theta_{\sigma^2}) = \prod_{i=1}^{n} \frac{1}{\sqrt{\theta_{\sigma^2} 2\pi}} \mathrm{e}^{\left(-\frac{1}{2}\cdot\frac{(x_i-\theta_\mu)^2}{\theta_{\sigma^2}}\right)}$

2. Log-likelihood: $\ln\left(\mathcal{L}(x_i; \theta_\mu, \theta_{\sigma^2})\right) = -\frac{n}{2}\ln(\theta_{\sigma^2}) - \frac{n\cdot\ln(2\pi)}{2} - \frac{1}{2\theta_{\sigma^2}}\sum_{i=1}^{n}(x_i - \theta_\mu)^2$

3. Take the derivatives:

   Partial derivative w.r.t θ_μ: $\frac{\partial}{\partial\theta_\mu}\ln(\mathcal{L}(..)) = \sum_{i=1}^{n}\frac{(x_i-\theta_\mu)}{\theta_{\sigma^2}}$

   Partial derivative w.r.t θ_{σ^2}: $\frac{\partial}{\partial\theta_{\sigma^2}}\ln(\mathcal{L}(..)) = -\frac{n}{2\theta_{\sigma^2}} + \frac{1}{2(\theta_{\sigma^2})^2}\sum_{i=1}^{n}(x_i - \theta_\mu)^2$

see next slide…

Example: Generalizing Normals

Let θ_μ and θ_{σ^2} be the unknown mean and variance of a normal distribution. We get independent draws $x_1, x_2, \ldots, x_n$ from the distribution.

…

3. Take the derivatives:

Partial derivative w.r.t θ_μ: $\frac{\partial}{\partial\theta_\mu}\ln\big(\mathcal{L}(..)\big) = \sum_{i=1}^{n}\frac{(x_i-\theta_\mu)}{\theta_{\sigma^2}}$

Partial derivative w.r.t θ_{σ^2}: $\frac{\partial}{\partial\theta_{\sigma^2}}\ln(\mathcal{L}(..) = -\frac{n}{2\theta_{\sigma^2}} + \frac{1}{2\left(\theta_{\sigma^2}\right)^2}\sum_{i=1}^{n}\left(x_i-\theta_\mu\right)^2$

4. Set to 0 and solve the system of equations for $\widehat{\theta_\mu}$ and $\widehat{\theta_{\sigma^2}}$:

$\sum_{i=1}^{n}\frac{(x_i-\widehat{\theta_\mu})}{\widehat{\theta_{\sigma^2}}} = 0$ *and* $-\frac{n}{2\widehat{\theta_{\sigma^2}}} + \frac{1}{2\left(\widehat{\theta_{\sigma^2}}\right)^2}\sum_{i=1}^{n}\left(x_i-\widehat{\theta_\mu}\right)^2 = 0$

after a bunch of algebra… $\theta_\mu = \frac{\sum_{i=1}^{n}x_i}{n}$ and $\widehat{\theta_{\sigma^2}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{\sum_{i=1}^{n}x_i}{n}\right)^2$

Example: Generalizing Normals - Summary

If you get independent samples $x_1, x_2, \dots, x_n$ from a $\mathcal{N}(\mu, \sigma^2)$ where μ and $\sigma^2$are unknown, the maximum likelihood estimates of the normal is:

$\widehat{\theta_\mu} = \frac{\sum_{i=1}^n x_i}{n}$ and $\widehat{\theta_{\sigma^2}} = \frac{1}{n}\sum_{i=1}^n \left(x_i - \frac{\sum_{i=1}^n x_i}{n}\right)^2$

The MLE of the mean ($\widehat{\theta_\mu}$) is the **sample mean** that is the estimate of μ is the average value of all the data points.

The MLE for the variance is **population variance**: compute that average squared distance the *observed samples* are away from the sample mean

General MLE Process

There's a distribution with some unknown parameter(s) θ. There are independent observations $x_1, x_2, \dots, x_n$ from this distribution.

1. Write the likelihood function: $\mathcal{L}(x_1, \dots, x_n; \theta)$

   multiply the probability/density of seeing each of those observations, based on the value of θ(s)

2. Take the log $\boldsymbol{ln}(..)$ of the likelihood function

3. Take the derivative(s) of the log-likelihood function

   If 2 parameters, take the partial derivative w.r.t θ_1 and the partial derivative w.r.t θ_2

4. Set the derivatives to 0, and solve for the MLE(s) $\boldsymbol{\hat{\theta}}$

   remember to switch from θ to $\hat{\theta}$ in this step because we're now solving for the MLE

5. Verify it is a maximum with second derivative test (not required for 312)

Properties of Estimators

How to we tell whether an estimator is “good”?

Biased

When we created our estimator, we based it off of a particular set of observed data. But was this data biased? Is the estimator generalizable?

So, one property we want from an estimator is for it to be **unbiased**.

An estimator $\hat{\theta}$ is "**unbiased**" if

$$\mathbb{E}\left[\hat{\theta}\right] = \theta$$

The expectation is taken over the randomness in the samples we drew.

The formula is fixed, the data we draw to evaluate the formula becomes the source of the randomness. We redefine each observed sample as *random variables*.

We're checking, "on average over *many* samples, is our estimator correct?"

Biased

If an estimator isn't unbiased then it's **biased**.

The "**bias**" of an estimator $\hat{\theta}$ is

$$\mathrm{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}] - \theta$$

$\mathrm{Bias}(\hat{\theta}, \theta) = 0$ --> *unbiased* ☺

$\mathrm{Bias}(\hat{\theta}, \theta) > 0$ --> *overestimate* ☹

$\mathrm{Bias}(\hat{\theta}, \theta) < 0$ --> *underestimate* ☹

Biased

If an estimator isn't unbiased then it's **biased**.

The "**bias**" of an estimator $\hat{\theta}$ is
$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}] - \theta$$

$\text{Bias}(\hat{\theta}, \theta) = 0$ --> *unbiased* ☺

$\text{Bias}(\hat{\theta}, \theta) > 0$ --> *overestimate* ☹

$\text{Bias}(\hat{\theta}, \theta) < 0$ --> *underestimate* ☹

Sometimes we can "fix" an estimator to be unbiased by adding/multiplying a constant to $\hat{\theta}$ to make $\text{Bias}(\hat{\theta}, \theta) = 0$

Are our MLEs biased?

We aim to evaluate the performance of our **estimator on average**. Instead of focusing on our specific data set, we *define random variables representing the distribution from which each sample is drawn*.

Example: MLE for mean of normal distribution

Each observed x_i is a samples from $\mathcal{N}(\mu, \sigma^2)$. Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

In general, our MLE is computed as: $\widehat{\theta_\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$

$$\mathbb{E}\left[\widehat{\theta_\mu}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] =$$

Are our MLEs biased?

We aim to evaluate the performance of our **estimator on average**. Instead of focusing on our specific data set, we *define random variables representing the distribution from which each sample is drawn.*

Example: MLE for mean of normal distribution

Each observed x_i is a samples from $\mathcal{N}(\mu, \sigma^2)$. Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

In general, our MLE is computed as: $\widehat{\theta_\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$

$$\mathbb{E}[\widehat{\theta_\mu}] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n}\mathbb{E}[\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \cdot n \cdot \mu = \mu$$

Unbiased! ✅

Are our MLEs biased?

We aim to evaluate the performance of our **estimator on average**. Instead of focusing on our specific data set, we *define random variables representing the distribution from which each sample is drawn.*

Example: MLE for probability of heads on a coin flip ($\boldsymbol{\theta}$)

Each observed coin toss x_i is a sample from $\mathrm{Ber}(\theta)$. Let $X_i \sim \mathrm{Ber}(\theta)$.

In general, our MLE is computed as: $\hat{\theta} = \frac{\text{num heads}}{\text{total flips}} = \underline{\hspace{3cm}}$

$\mathbb{E}[\hat{\theta}] =$

pollev.com/cse312

Are our MLEs biased?

We aim to evaluate the performance of our **estimator on average**. Instead of focusing on our specific data set, we *define random variables representing the distribution from which each sample is drawn*.

Example: MLE for probability of heads on a coin flip ($\boldsymbol{\theta}$)

Each observed coin toss x_i is a sample from $\mathrm{Ber}(\theta)$. Let $X_i \sim \mathrm{Ber}(\theta)$.

In general, our MLE is computed as: $\hat{\theta} = \frac{\text{num heads}}{\text{total flips}} = \frac{\sum_i^n X_i}{n}$

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}\left[\frac{\sum_i^n X_i}{n}\right] = \frac{1}{n}\mathbb{E}[\sum_i^n X_i] = \frac{1}{n}\sum_i^n \mathbb{E}[X_i] = \frac{1}{n}\sum_i^n \theta = \frac{1}{n} \cdot n \cdot \theta = \theta$$

Unbiased! ✅

Are our MLEs biased?

We aim to evaluate the performance of our **estimator on average**. Instead of focusing on our specific data set, we *define random variables representing the distribution from which each sample is drawn*.

Example: MLE for *variance* of normal distribution

Each observed x_i is a samples from $\mathcal{N}(\mu, \sigma^2)$. Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

In general, our MLE is computed as: $\widehat{\theta_{\sigma^2}} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\theta_\mu}\right)^2]$

$\mathbb{E}\left[\widehat{\theta_{\sigma^2}}\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\left(X_i - \widehat{\theta_\mu}\right)^2\right] =$… an ✨ algebraic miracle ✨ … $= \frac{n-1}{n} \cdot \sigma^2$

Unbiased! ❌

Intuition: $\widehat{\theta_\mu} = \sum X_i / n$. So when that gets squared, there are terms that have $X_i X_j$ terms and $X_i \cdot X_i$ terms. The $1/n$ fraction of terms that are $X_i X_i$ decrease the variance because you can't deviate from yourself.

That Algebraic Miracle

$$= \frac{1}{n}\mathbb{E}\left[\sum x_i^2 - 2x_i\widehat{\theta_\mu} + \widehat{\theta_\mu}^2\right]$$

$$= \frac{1}{n}\mathbb{E}\left[\sum x_i^2\right] - \frac{1}{n}\mathbb{E}\left[\sum 2x_i\widehat{\theta_\mu} - \sum\widehat{\theta_\mu}^2\right]$$

$$= \frac{1}{n}n\mathbb{E}[x_1^2] - \frac{1}{n}\mathbb{E}\left[2\widehat{\theta_\mu}\sum x_i - \sum\widehat{\theta_\mu}^2\right]$$

$$= \mathbb{E}[x_1^2] - \frac{1}{n}\mathbb{E}\left[2n\widehat{\theta_\mu}^2 - n\widehat{\theta_\mu}^2\right]$$

$$\boxed{\widehat{\theta_\mu} = \sum x_i / n}$$

$$= \mathbb{E}[x_1^2] - \frac{1}{n}\mathbb{E}\left[n\widehat{\theta_\mu}^2\right]$$

$$= \mathbb{E}[x_1^2] - \mathbb{E}\left[\widehat{\theta_\mu}^2\right]$$

More of That Algebraic Miracle

$$\mathbb{E}\left[\widehat{\theta_\mu}^2\right] = \mathbb{E}\left[\left(\frac{\Sigma x_i}{n}\right)\left(\frac{\Sigma x_i}{n}\right)\right]$$

$$= \frac{1}{n^2}\mathbb{E}\left[\sum_{i\neq j} x_i \cdot x_j + \sum_i x_i^2\right]$$

These are the $x_i x_i$ terms.

$$= \frac{1}{n^2}\mathbb{E}\left[\sum_{i\neq j} x_i \cdot x_j\right] + \frac{1}{n^2}\mathbb{E}\left[\sum_i x_i^2\right]$$

$$= \frac{1}{n^2} \cdot n(n-1)\mathbb{E}[x_1 \cdot x_2] + \frac{1}{n^2} n\mathbb{E}[x_1^2]$$

$$= \frac{n-1}{n}\mathbb{E}[x_1]\mathbb{E}[x_1] + \frac{1}{n}\mathbb{E}[x_1^2]$$

This is where the $x_i x_i$ terms end up

Wrapping Up the Algebraic Miracle

$\mathbb{E}[\theta_{\sigma^2}] = \mathbb{E}[x_1^2] - \mathbb{E}\left[\widehat{\theta_\mu}^2\right]$

Plugging in $\mathbb{E}\left[\widehat{\theta_\mu}^2\right] = \frac{n-1}{n}\mathbb{E}[x_1]\mathbb{E}[x_1] + \frac{1}{n}\mathbb{E}[x_1^2]$ we get:

$\mathbb{E}[\theta_{\sigma^2}] = \mathbb{E}[x_1^2] - \left(\frac{n-1}{n}\mathbb{E}[x_1]\mathbb{E}[x_1] + \frac{1}{n}\mathbb{E}[x_1^2]\right)$

$= \mathbb{E}[x_1^2] - \frac{n-1}{n}\mathbb{E}[x_1]^2 - \frac{1}{n}\mathbb{E}[x_1^2]$

$= \frac{n-1}{n}\mathbb{E}[x_1^2] - \frac{n-1}{n}\mathbb{E}[x_1]^2$

$= \frac{n-1}{n}\mathrm{Var}(x_1)$

Not Unbiased

$\mathbb{E}\left[\widehat{\theta_{\sigma^2}}\right] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \widehat{\theta_\mu}\right)^2]$

$= \frac{n-1}{n}\sigma^2$

Which is not what we wanted. This is biased. But it's not *too* biased…

An estimator $\hat{\theta}$ is "**consistent**" if

$$\lim_{n\to\infty} \mathbb{E}\left[\hat{\theta}\right] = \theta$$

The MLE is always consistent, but can be biased or unbiased.

Correction

The MLE slightly underestimates the true variance.

You could correct for this! Just multiply by $\boldsymbol{\frac{n}{n-1}}$.

This would give you a formula of:

$$\frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \widehat{\theta_\mu}\right)^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \widehat{\theta_\mu}\right)^2$$

$\widehat{\theta_\mu}$ is the sample mean.

Called the "**sample variance**" because it's the variance you estimate if you want an (unbiased) estimate of the variance given only a sample.

If you took a statistics course, you might have learned the square root of this as the definition of standard deviation.

Fun Facts

What’s with the $n - 1$?

Sooooooooooo, why is the MLE off?

Intuition 1: when we’re comparing to the real mean, x_1 doesn’t affect the real mean (the mean is what the mean is regardless of what you draw).

But when you compare to the sample mean, x_1 pulls the sample mean toward it, decreasing the variance a tiny bit.

Intuition 2: We only have $n - 1$ “degrees of freedom” with the mean and $n - 1$ of the data points, you know the final data point. Only $n - 1$ of the data points have “information” the last is fixed by the sample mean.

Why does it matter?

When statisticians are estimating a variance from a sample, they usually divide by $n-1$ instead of n.

They also (with unknown variance) generally don’t use the CLT to estimate probabilities.

A “t-test” is used when scientists/statisticians think their data is approximately normal, but they don’t know the variance.

They aren’t using the $\Phi()$ table, they’re using a different table based on the altered variance estimates.

Why use MLEs? Are there other estimators?

If you have a prior distribution over what values of θ are likely, combining the idea of Bayes rule with the idea of an MLE will give you

Maximum a posteriori probability estimation (MAP)

You pick the maximum value of $\mathbb{P}(\theta|E)$ starting from a known prior over possible values of θ.

$$\mathbf{argmax}_{\theta} \frac{\mathbb{P}(E|\theta)\cdot\mathbb{P}(\theta)}{\mathbb{P}(E)} = \mathbf{argmax}_{\theta}\mathbb{P}(E|\theta)\cdot\mathbb{P}(\theta)$$

$\mathbb{P}(E)$ is a constant, so the argmax is unchanged if you ignore it.

Note when prior is constant, you get MLE!