Maximum Likelihood Estimation

CSE 312 26Su

Lecture 21

Announcements

- Final information on website
- Homework 6 is due today
- Homework 7 (last homework!) will be released today
- Quiz 7 on Friday (Tail Bounds + MLE)

Where Are We?

Today:

- Tail Bounds
  - Chernoff Bound
  - not a tail bound, but Union Bound
- Maximum Likelihood Estimation

Today:

- Maximum Likelihood Estimation
- Biased / Unbiased Estimators

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.

> $X \sim \mathbf{Bin}(n, p)$, 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*

Suppose you flip a coin independently 10 times, and you see

HTTTHHTHHH

(6 heads, 4 tails)

What is your estimate of p, the probability the coin comes up heads?

Maybe $p = \frac{6}{10}$ $\left(X \sim \mathrm{Ber}\left(\frac{6}{10}\right)\right)$ *But how to argue "objectively" this is the best estimate?*

Likelihood function

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

$\mathcal{L}(E;\boldsymbol{\theta})$ 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**

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

"Probability of observing HTTTHHTHHH if $\boldsymbol{\theta}$ is probability of heads on a single flip"

We want to pick the value of $\boldsymbol{\theta}$ that make this likelihood function the largest. So…

Maximum Likelihood Estimation

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

"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:

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

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

Use $\hat{\theta}_{\text{MLE}}$ 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

Coin flips is easier

1. Likelihood function: $\mathcal{L}(\text{HTTTHHTHHH}; \theta) = \theta^6(1-\theta)^4$
2. Take the log: $\ln\big(\mathcal{L}(\text{HTTTHHTHHH}; \theta)\big) = 6\ln(\theta) + 4\ln(1-\theta)$
3. Take the derivative: $\frac{d}{d\theta}\ln\big(\mathcal{L}(\cdot)\big) = \frac{6}{\theta} - \frac{4}{1-\theta}$
4. Set to $\mathbf{0}$ and solve:

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

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 must be 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 $\mathcal{L}(E;\theta)$
   Usually $\prod \mathbb{P}(x_i;\theta)$ for discrete and $\prod f(x_i;\theta)$ for continuous
2. Take the log $\ln(\ldots)$ of the likelihood function (makes the math easier)
3. Take the derivative of the log-likelihood function
4. Set the derivative to 0 and solve for the MLE $\hat{\theta}$
   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

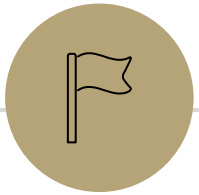
$$\ln(a \cdot b) = \ln(a) + \ln(b)$$

$$\ln\left(a^{b}\right) = b \cdot \ln(a)$$

$$\ln(a - b) = \frac{\ln(a)}{\ln(b)}$$

$$\ln(e^{a}) = a$$

$$\frac{d}{d\theta}\ln(m) = \frac{1}{m} \cdot \frac{d}{d\theta}(m)$$



MLEs with *Continuous* RVs

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)

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

Continuous Example

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

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

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

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

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

Finding $\hat{\mu}$

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

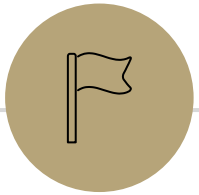
3. Take the derivative: $\frac{d}{d\mu}\ln(\mathcal{L}(x_i;\mu)) = \sum_{i=1}^{n} x_i - \mu$

4. Set the derivative to 0 and solving:

$$\sum_{i=1}^{n} x_i - \hat{\mu} = 0 \Rightarrow \sum_{i=1}^{n} x_i = \hat{\mu} \cdot n \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

5. Verify it is a maximum with 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 average of the x_i is a maximizer.



MLEs with *Multiple* Parameters

Generalizing Normals

We just saw to estimate μ for $\mathcal{N}(\mu, 1)$ we get:

$\hat{\mu} = \sum x_i / n$

Now what happens if we know our data is normally distributed but both the mean and the variance are unknown?

Generalizing Normals

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

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

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

With multiple parameters, take partial derivatives to find maxima.

Generalizing Normals

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

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

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

3. Take the derivative:

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

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

Partial derivative w.r.t. θ_{σ^2}

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

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

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

$$\frac{\partial}{\partial\theta_{\sigma^2}}\ln(\mathcal{L})=-\frac{n}{2\theta_{\sigma^2}}+\frac{1}{2\left(\theta_{\sigma^2}\right)^2}\sum_{i=1}^{n}\left(x_i-\theta_\mu\right)^2$$

Generalizing Normals

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

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

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

3. Take the derivative:

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

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

Generalizing Normals

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

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

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

3. Take the derivative:

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}: $\frac{\partial}{\partial \theta_\mu} \ln(\mathcal{L}(..)) = -\frac{n}{2\theta_{\sigma^2}} + \frac{1}{2(\theta_{\sigma^2})^2} \sum_{i=1}^{n} (x_i - \theta_\mu)^2$

4. Set the derivative to 0 and solve for the MLEs $\widehat{\theta_\mu}$, $\widehat{\theta_{\sigma^2}}$

Solving for $\widehat{\theta_\mu}$, $\widehat{\theta_{\sigma^2}}$

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

$$\frac{\partial}{\partial \theta_\mu} \ln(\mathcal{L}) = \sum_{i=1}^{n} \frac{(x_i - \theta_\mu)}{\theta_{\sigma^2}}$$

Setting equal to 0 and solving

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

5. Verify it is a maximum with second derivative test:

$$\frac{\partial^2}{\partial \theta_\mu^2} = -\frac{n}{\theta_{\sigma^2}}$$

θ_{σ^2} is an estimate of variance ≥ 0 and when draws aren't identical, it won't be 0. So second derivative is negative and we have a maximizer.

Solving for $\widehat{\theta_\mu}$, $\widehat{\theta_{\sigma^2}}$

$$\frac{\partial}{\partial \theta_{\sigma^2}} \ln(\mathcal{L}) = -\frac{n}{2\theta_{\sigma^2}} + \frac{1}{2(\theta_{\sigma^2})^2} \sum_{i=1}^{n} (x_i - \theta_\mu)^2$$

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

$$\Rightarrow -\frac{n}{2}\widehat{\theta_{\sigma^2}} + \frac{1}{2} \sum_{i=1}^{n} (x_i - \theta_\mu)^2 = 0 \text{ (multiply by } (\widehat{\theta_{\sigma^2}})^2)$$

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

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

To get the overall max
We'll plug in $\widehat{\theta_\mu}$

Generalizing Normals Summary

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

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

The MLE for the mean is the **sample mean** that is the estimate of μ is the average value of all the data points.

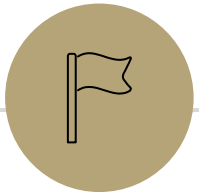
The MLE for the variance is the **population variance**: compute average squared distance the *observed samples* are away from the sample mean

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 unknown parameter(s) θ…

1. Write the likelihood function $\mathcal{L}(E;\theta)$
   Usually $\prod \mathbb{P}(x_i;\theta)$ for discrete and $\prod f(x_i;\theta)$ for continuous
2. Take the log $\ln(\ldots)$ of the likelihood function (makes the math easier)
3. Take the (partial) derivative(s) of the log-likelihood function
4. Set the derivatives to 0 and solve for the MLEs $\hat{\theta}$
   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 do we tell whether an estimator is “good”?

Biased

One property we might want from an estimator is for it to be **unbiased**.

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

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

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*.

"On average over *many* samples, is our estimator correct?"

Biased

If the bias isn't unbiased then it is **biased**.

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

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

$\text{Bias}(\theta, \hat{\theta}) = 0$ → *unbiased*

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

$\text{Bias}(\theta, \hat{\theta}) < 0$ → *underestimate*

Biased

If the bias isn't unbiased then it is **biased**.

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

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

$\text{Bias}(\theta, \hat{\theta}) = 0$ → *unbiased*

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

$\text{Bias}(\theta, \hat{\theta}) < 0$ → *underestimate*

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

Are our MLEs unbiased?

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

Example: MLE for mean of normal distribution

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

$$\widehat{\theta_\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$$

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

Are our MLEs unbiased?

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

Example: MLE for mean of normal distribution

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

$$\widehat{\theta_\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$$

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

Unbiased!

Are our MLEs unbiased?

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

Example: MLE for probability of heads on a coin flip

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

In general, our MLE is: $\hat{\theta} = \frac{\textbf{num heads}}{\textbf{total flips}} = \_\_\_\_\_$

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

Are our MLEs unbiased?

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

Example: MLE for probability of heads on a coin flip

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

In general, our MLE is: $\hat{\theta} = \frac{\text{num heads}}{\text{total flips}} = \frac{\sum_{i=1}^{n} X_i}{n}$

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

Unbiased!

Are our MLEs unbiased?

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

Example: MLE for variance of a normal distribution

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

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

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

Are our MLEs unbiased? (variance)

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

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

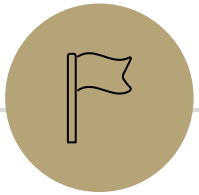
$$= \frac{1}{n}\mathbb{E}\left[\sum X_i^2 - 2X_i\widehat{\theta_\mu} + \widehat{\theta_\mu}^2\right]$$

…

Then an algebraic miracle occurs…

$= \frac{n-1}{n} \cdot \sigma^2$ where $\sigma^2 = \mathbb{E}[(X_i - \mathbb{E}[X_i])\text{^}2]$

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



Optional: Algebra

Showing MLE of Variance is biased

That Algebraic Miracle

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

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

$$= \frac{1}{n}n\mathbb{E}[X_1^2] - \frac{1}{n}\mathbb{E}\left[2\widehat{\theta_\mu}\sum X_i - \sum\widehat{\theta_\mu}^2\right]$$

$$= \mathbb{E}[X_1^2] - \frac{1}{n}\mathbb{E}\left[2n\widehat{\theta_\mu}^2 - n\widehat{\theta_\mu}^2\right]$$

$$= \mathbb{E}[X_1^2] - \frac{1}{n}\mathbb{E}\left[n\widehat{\theta_\mu}^2\right]$$

$$= \mathbb{E}[X_1^2] - \mathbb{E}\left[\widehat{\theta_\mu}^2\right]$$

$$\widehat{\theta_\mu} = \sum X_i/n$$

More of That Algebraic Miracle

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

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

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

$$= \frac{1}{n^2}\cdot n(n-1)\mathbb{E}[X_1 \cdot X_2] + \frac{1}{n^2} n\mathbb{E}[X_1^2]$$

$$= \frac{n-1}{n}\mathbb{E}[X_1]\mathbb{E}[X_1] + \frac{1}{n}\mathbb{E}[X_1^2]$$

These are the $x_i x_i$ terms.

This is where the $x_i x_i$ terms end up

Wrapping Up the Algebraic Miracle

$$\mathbb{E}[\theta_{\sigma^2}] = \mathbb{E}[X_1^2] - \mathbb{E}\left[\widehat{\theta_\mu}^2\right]$$

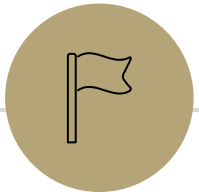
Plugging in $\mathbb{E}\left[\widehat{\theta_\mu}^2\right] = \frac{n-1}{n}\mathbb{E}[X_1]\mathbb{E}[X_1] + \frac{1}{n}\mathbb{E}[X_1^2]$ we get:

$$\mathbb{E}[\theta_{\sigma^2}] = \mathbb{E}[X_1^2] - \left(\frac{n-1}{n}\mathbb{E}[X_1]\mathbb{E}[X_1] + \frac{1}{n}\mathbb{E}[X_1^2]\right)$$

$$= \mathbb{E}[X_1^2] - \frac{n-1}{n}\mathbb{E}[X_1]^2 - \frac{1}{n}\mathbb{E}[X_1^2]$$

$$= \frac{n-1}{n}\mathbb{E}[X_1^2] - \frac{n-1}{n}\mathbb{E}[X_1]^2$$

$$= \frac{n-1}{n}\mathrm{Var}(X_1)$$



Non-Optional Takeaways

Not Unbiased

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

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

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 consistent (under some very mild assumptions),
but it can be biased or unbiased.

Correction

The MLE slightly underestimates the true variance.

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

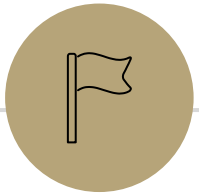
This would give you a formula of:

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

$= \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \widehat{\theta_\mu}\right)^2$ where $\widehat{\theta_\mu}$ 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 probably learned the square root of this as the definition of standard deviation.



Fun Facts

What's with the $n-1$?

Soooooooooo, 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 Φ() 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!