MLE continued

CSE 312 Summer 21
Lecture 22

Finding $\hat{\mu}$ for Normals

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

$$\frac{d}{d\mu}\ln(\mathcal{L}) = \sum_{i=1}^{n} x_i - \mu$$

Setting $\mu = 0$ and solving:

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

Check using the 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.

Two Parameter Estimation for Normals

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 maximum likelihood estimator of 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 variance of the experiment "choose one of the x_i at random"

Summary

Given: an event E (usually n i.i.d. samples from a distribution with unknown parameter θ).

1. Find likelihood $\mathcal{L}(E;\theta)$

Usually $\prod \mathbb{P}(x_i;\theta)$ for discrete and $\prod f(x_i;\theta)$ for continuous

2. Maximize the likelihood. Usually:

A. Take the log (if it will make the math easier)
B. Take the derivative
C. Set the derivative to 0 and solve

3. Use the second derivative test to confirm you have a maximizer

Biased

One property we might 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.

(those samples are random variables).

So, we're not consistently overestimating or underestimating.

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

Are our MLEs unbiased?

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

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

Unbiased!

Are our MLEs biased?

Our estimate for the probability coin-flips being heads (if we generalized a bit) would be $\frac{\textbf{num heads}}{\textbf{total flips}}$

Fill out the poll everywhere so Kushal knows how long to explain
Go to pollev.com/cse312su21

Are our MLEs biased?

Our estimate for the probability coin-flips being heads (if we generalized a bit) would be $\frac{\textbf{num heads}}{\textbf{total flips}}$

What is $\mathbb{E}\left[\frac{\textbf{num heads}}{\textbf{total flips}}\right] = \mathbb{E}\left[\frac{\Sigma x_i}{n}\right] = \frac{\theta \cdot n}{n} = \theta$

Unbiased!

Fill out the poll everywhere so Kushal knows how long to explain
Go to pollev.com/cse312su21

Is MLE for Variance of Normal unbiased?

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

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

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

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

$$= \mathbb{E}\left[x_i^2\right] - \mathbb{E}\left[\widehat{\theta_\mu}^2\right] = Var(x_i) + \mathbb{E}[x_i]^2 - \left(Var\left(\widehat{\theta_\mu}\right) + \mathbb{E}\left[\widehat{\theta_\mu}\right]^2\right) = Var(x_i) - Var\left(\widehat{\theta_\mu}\right)$$

$$= Var(x_i) - Var\left(\frac{\sum x_i}{n}\right) = Var(x_i) - \frac{1}{n^2}Var(\sum x_i) = Var(x_i) - \frac{1}{n^2}\cdot n\cdot Var(x_i)$$

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

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

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 a biased estimator. But it's not too biased…

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

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

Correction

The MLE slightly underestimates the true variance.

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

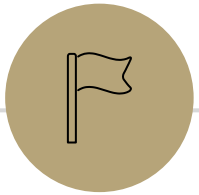
This would give you a formula of:

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

$= \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \widehat{\theta_\mu}\right)^2$ 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 this as the definition of variance.



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.

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

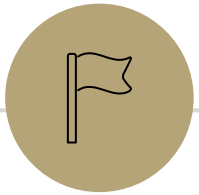
Maximum a posteriori probability estimation (MAP)

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!



More Practice

MLE of Continuous Uniform Distribution

Let $x_1, x_2, \ldots, x_n$ be independent samples from the continuous uniform distribution, $X \sim Unif(0, \theta)$. What is the MLE $\hat{\theta}$ for θ?

Fill out the poll everywhere so Kushal knows how long to explain
Go to pollev.com/cse312su21

MLE of Continuous Uniform Distribution

Let $x_1, x_2, \ldots, x_n$ be independent samples from the continuous uniform distribution, $X \sim Unif(0, \theta)$. What is the MLE $\hat{\theta}$ for θ?

$$\mathcal{L}(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} \frac{1}{\theta - 0} = \theta^{-n}$$

$$\frac{d}{d\theta} \mathcal{L}(\cdot\,; \theta) = \frac{d}{d\theta}(\theta^{-n}) = -n\theta^{-n-1}$$

The derivative is 0 only when $\theta = \infty$

When $\theta = \infty$, $\mathcal{L}(\cdot\,; \theta) = 0$. This is a minimum and not a maximum.

Additionally, we know that all $x_i \in [0, \theta]$.

Hence, if $\theta < \max(x_1, x_2, \ldots, x_n)$, $\mathcal{L}(\cdot\,; \theta) = 0$

MLE of Continuous Uniform Distribution

When $\theta = \infty$, $\mathcal{L}(\cdot\,;\theta) = 0$. This is a minimum and not a maximum.

Additionally, we know that all $x_i \in [0, \theta]$.

Hence, if $\theta < \max(x_1, x_2, \ldots, x_n)$, $\mathcal{L}(\cdot\,;\theta) = 0$

This point of discontinuity, $\max(x_1, x_2, \ldots, x_n)$ is where the likelihood is maximized. Hence the MLE is $\hat{\theta} = \max(x_1, x_2, \ldots, x_n)$