Announcements

HW8, we made an Ed post about leaf
↳ if you find another way, we'll give it credit.

HW9 out tonight 2 find an MLE
1 take a practice final

Don't forget Real World 3.

$\hat{\theta}$ $\hat{\mu}$

More MLE

CSE 312 Spring 21
Lecture 26

Outline

Last time: Trying to estimate an unknown parameter θ of a distribution.

We chose the "maximum likelihood estimator"

$\text{argmax}_\theta \mathcal{L}(x;\theta)$

Usually: write likelihood, take log, take derivative, confirm maximum

Today: What happens when you have two parameters, MLEs that aren't what you expect.

Question for today

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

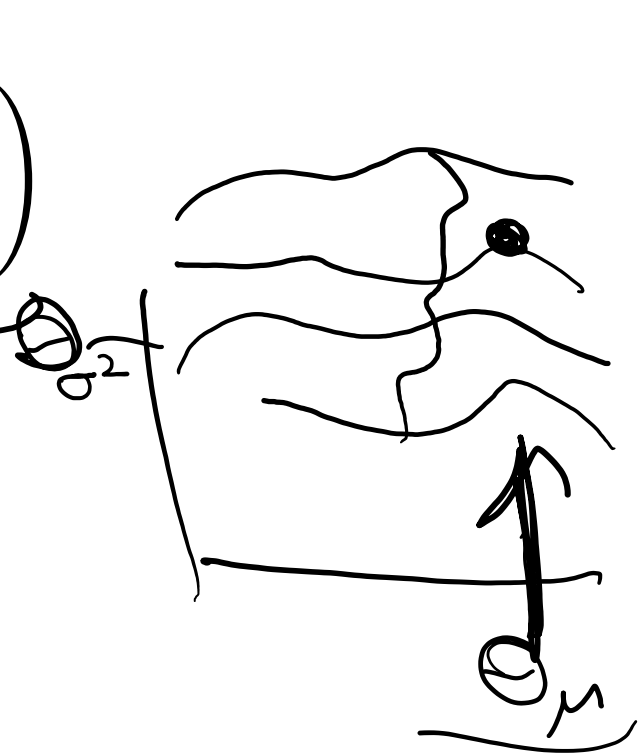
Now what happens if we know our data is $\mathcal{N}()$ but nothing else. Both the mean and the variance are unknown.

Log-likelihood

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

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

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



Expectation

Arithmetic is nearly identical to known variance case.

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

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

Setting equal to 0 and solving

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

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

θ_{σ^2} is an estimate of a variance. It'll never be negative (and as long as the draws aren't identical it won't be 0). So the second derivative is negative and we really have a maximizer.

Variance

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

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

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

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

Variance part 2

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

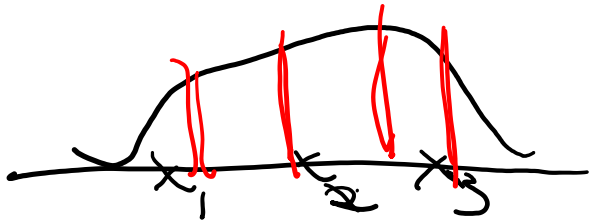
$$-\frac{n}{2\theta_{\sigma^2}} + \frac{1}{2(\theta_{\sigma^2})^2} \sum_{i=1}^{n} (x_i - \theta_\mu)^2 = 0$$

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

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

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

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:

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

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"

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.

(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 coin-flips (if we generalized a bit) would be

$$\frac{\text{num heads}}{\text{total flips}}$$

p = prob heads single flip

X_i = ind. for i^{th} flip is heads

n = total # of flips

$$\hat{\theta} = \frac{\sum X_i}{n}$$

$$E[\hat{\theta}] = E\left[\frac{\sum X_i}{n}\right] = \frac{1}{n} E[\sum X_i] = \frac{1}{n} \sum E[X_i]$$

$$\frac{1}{n} \sum p = \frac{1}{n} \cdot n \cdot p = p$$

Fill out the poll everywhere so Robbie knows how long to explain
Go to pollev.com/cse312

Are our MLEs biased?

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

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

Unbiased!

Fill out the poll everywhere so Robbie knows how long to explain
Go to pollev.com/cse312

Unbiased?

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

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

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

…

Then an algebraic miracle occurs…

$$= \frac{n-1}{n}\cdot\sigma^2 \text{ where } \sigma^2 = \mathbb{E}\left[x_i - \mathbb{E}[x_i]\right]$$

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

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

MLEs are always consistent

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:

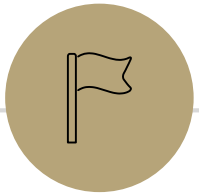
$$\frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \widehat{\theta_\mu}\right)^2$$

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

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 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.

A "t-test" is used when scientists/statisticians think their

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!