# More MLE

# Outline

Last time: Trying to estimate an unknown parameter $\theta$ of a distribution.

We chose the "maximum likelihood estimator"

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

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

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

# 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.

$$\mathcal{L}(x_1, x_2, \ldots, x_n; \boldsymbol{\theta}) = \prod f_X(x_i; \boldsymbol{\theta})$$

# Continuous Example

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

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

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

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

# Finding $\hat{\mu}$

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

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

Setting derivative $= 0$ and solving:

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

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.

# Summary

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

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. Set the derivative to 0 and solve

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

# Generalizing Normals

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

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

Now what happens if we know our data is $\mathcal{N}()$ but both the mean and the variance are unknown.?

# Log-likelihood

Let $\theta_\mu$ and $\theta_{\sigma^2}$ 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.

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

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

With multiple parameters, take partial derivatives to find maxima.

# Expectation

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

$$\frac{\partial}{\partial \theta_\mu} \ln(\mathcal{L}) =$$

Setting equal to $0$ and solving

# Expectation

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

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

Setting equal to $0$ and solving

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

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

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

Take the partial derivative with respect to $\theta_{\sigma^2}$. It'll be easier if you apply some log and exponent rules first.

$$\log(x^y) = y \cdot \log(x).$$

$$\log(ab) = \log(a) + \log(b).$$

$$\frac{1}{\sqrt{a}} = a^{-1/2}$$

# Variance

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

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

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

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

# Variance part 2

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

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

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

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

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

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

# Summary

If you get independent samples $x_1, x_2, \ldots, x_n$ from a $\mathcal{N}(\mu, \sigma^2)$ where $\mu$ and $\sigma^2$ 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 $\mu$ 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**
> $$\mathbb{E}\left[\hat{\theta}\right] = \theta$$

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.

So we're not consistently overestimating or underestimating.

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

# Are our MLEs unbiased?

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

$$\mathbb{E}\big[\widehat{\theta_\mu}\big] = \frac{1}{n}\,\mathbb{E}[\textstyle\sum x_i] = \frac{1}{n}\textstyle\sum\mathbb{E}[x_i] = \frac{1}{n}\cdot n \cdot \mu = \mu$$

Unbiased!

# Are our MLEs biased?

Our estimate for the coin-flips (if we generalized a bit) would be

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

pollev.com/robbie

# Are our MLEs biased?

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

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

Unbiased!

# Unbiased?

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

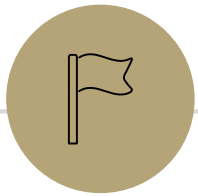$$= \frac{1}{n}\mathbb{E}\left[\sum\left(x_i - \widehat{\theta_\mu}\right)^2\right]$$

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

…

Then an algebraic miracle occurs…

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

Intuition for the algebra miracle:
$\widehat{\theta_\mu} = \sum x_i/n$. So when that gets squared, there are terms that have $x_i x_j$ 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.

# Optional: Algebra

Showing MLE of Variance is biased

# That Algebraic Miracle

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

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

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

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

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

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

# More of That Algebraic Miracle

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

These are the $x_i x_i$ terms.

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

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

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

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

This is where the $x_i x_i$ terms end up

# Wrapping Up the Algebraic Miracle

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

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

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

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

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

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

# Not Unbiased

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

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

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

**An estimator $\hat{\theta}$ is "consistent" if**
$$\lim_{n \to \infty} \mathbb{E}\left[\hat{\theta}\right] = \theta$$

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:

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

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left( x_i - \widehat{\theta_\mu} \right)^2 \text{ where } \widehat{\theta_\mu} \text{ is the sample mean.}$$

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.

# What about $X$ and $Y$ from last Friday?

We polled $n$ people and said, Flip a coin:
If coin is heads OR you have cheated on a partner, tell me "heads"
If coin is tails AND you have never cheated on a partner, tell me "tails"

$X$ was the number of people polled who said "heads"

$Y$ was the number of people polled who cheated on a partner.

We're trying to find an estimator for $Y$.

# What about $X$ and $Y$ from last Friday?

We polled $n$ people and said, Flip a coin:

If coin is heads OR you have cheated on a partner, tell me "heads"

If coin is tails AND you have never cheated on a partner, tell me "tails"

$X$ was the number of people polled who said "heads"

$Y$ was the number of people polled who cheated on a partner.

We're trying to find an estimator for $Y$.

We picked the estimator "$\hat{Y} = 2\left(X - \frac{n}{2}\right)$"

$\mathbb{E}[\hat{Y}] = 2\left(\mathbb{E}[X] - \frac{n}{2}\right) = 2\left(\left[\frac{n}{2} + \frac{np}{2}\right] - \frac{n}{2}\right) = np$ where $p$ is fraction of people who cheated. I.e., $Y$ This was an unbiased estimator!

# What about $X$ and $Y$ from last Friday?

$X$ was the number of people polled who said "heads"

$Y$ was the number of people polled who cheated on a spouse.

We're trying to find an estimator for.

$$\mathcal{L}(X = k; Y) = \binom{n-Y}{k-Y} \cdot 5^{k-Y} \cdot 5^{n-k} = \binom{n-Y}{k-Y} \cdot 5^{n-Y}$$
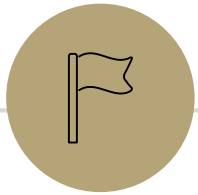
The binomial coefficient is maximized when it's $\binom{m}{m/2}$

Analysis is more complicated because we can't use calculus (defined only on integers)

$$k - Y = \frac{n-Y}{2} \Rightarrow k - \frac{n}{2} = \frac{Y}{2} \Rightarrow Y = 2\left(k - \frac{n}{2}\right)$$

So this is also an MLE!

This estimator is only handling the randomness in the coin flips, not the randomness in who was selected. You get the same answer if you back up that far.

# Fun Facts

# What's with the $n - 1$?

Soooooooooo, why is the MLE for variance 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 $\theta$ 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 $\theta$.

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

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

Note when prior is constant, you get MLE!