STA261: Week 4

Likelihood Inference II

Alex Stringer

Jan 29th - Feb 2nd, 2018

Disclaimer

The materials in these slides are intended to be a companion to the course textbook, *Mathematical Statistics and Data Analysis, Third Edition*, by John A Rice. Material in the slides may or may not be taken directly from this source. These slides were organized and typeset by Alex Stringer.

A big thanks to Jerry Brunner as well for providing inspiration for assignment questions.

License

Last week

Last week, we talked about

- Sufficiency
- The likelihood function
- Maximum likelihood estimators

We got what seems to be a satisfying recipe for finding parameter estimates for data from known families of distributions.

So we're done I guess?

This week

. . . not quite. We still need to argue that in general, the MLE procedure provides reasonable estimators.

This week we will study the asymptotic (large-sample) distribution of the MLE

Recall

We have found some examples of MLE's. For example, with $X_i \sim N(\mu, \sigma)$ we found that

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}) = \left(\bar{X}, s\right)$$

where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \mu\right)^2}$$

Note we divide by n and not $n-1$.

Plot the log-likelihood

We can plot the log-likelihood, which is the function that these quantities alledgedly maximize, as a function of μ, and of σ, for a fixed dataset.

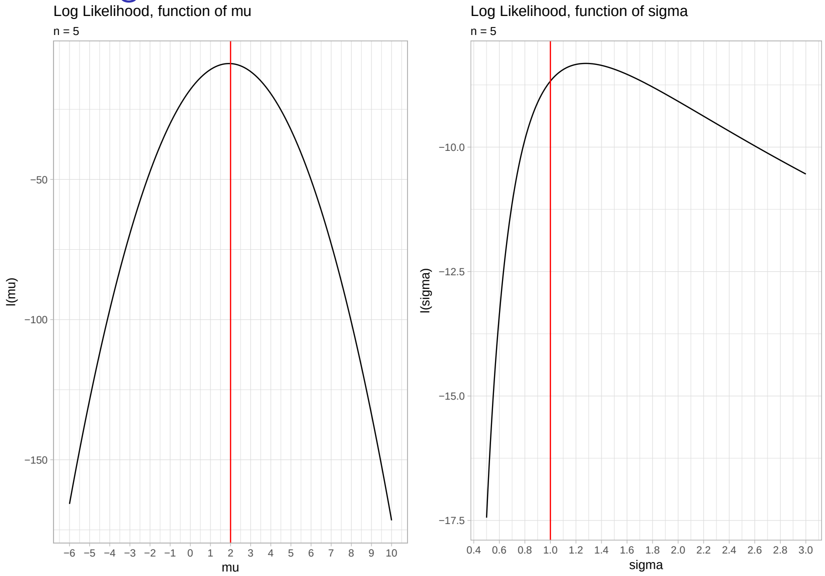
We actually plot the *contours* of μ for fixed σ, and of σ for fixed μ.

Consider $n = 5$ and $\mathbf{x} = (2.89, 3.63, 1.33, 1.81, -0.05)$; 5 values sampled independently from a $N(2, 1)$ distribution.

The log-likelihood is

$$\ell(\mu) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Plot the log-likelihood



Plot the log-likelihood

The red lines show the true parameter values. The MLE is the point at which each curve is a maximum. Why aren't they the same?

We generated the data randomly, and a sample size of 5 is small, so the variability is large.

Actually, when you think about it, we did very well with 5 datapoints.

Recall: Curvature

The *curvature* of a function refers to its (absolute) second derivative,

$$\left|\frac{\partial^2 f(x)}{\partial x^2}\right|$$

It's called "curvature" because the second derivative defines how "peaked" or "flat" the function is around a point. If the curvature is high at x, then slopes tangent to $f(x)$ at x are changing very rapidly in the vicinity of x, and the function is very peaked. If the curvature is low, then slopes tangent to $f(x)$ are changing slowly in the vicinity of x, and the function is flat.

Curvature

The plots illustrate the importance of the *curvature* of the (log) likelihood.

Remember that the likelihood function defines which values of θ are *plausible* given the observed data.

Likelihoods that are very *peaked* around their maximums define a very narrow range of plausible values for θ. Likelihoods that are very *flat* around their maximums define a very wide range of plausible values for θ.

All of this is for a given set of observed data.

Sampling Distribution

The plot for $\ell(\mu)$ tells us that higher and lower values for μ are equally plausible, given the observed data.

Knowing that the MLE is $\hat{\mu} = \bar{X}$, we know its sampling distribution is $N(\mu, \sigma/\sqrt{n})$, so this makes a bit of sense- in repeated samples, we expect that the distribution of the $\hat{\mu}$ we calculate to be symmetric and centered on the true value μ.

Sampling Distribution

The plot for $\ell(\sigma)$ tells a different story. Lower values are less plausible (given the observed data) than higher values (why?).

Later in the course, we will derive the exact sampling distribution of $\hat{\sigma}^2$. It has a long right tail, which tells the same story as the log-likelihood function here.

Example

Let's check out a more difficult example. Let $X_i \sim Gamma(\alpha, \beta)$, with density

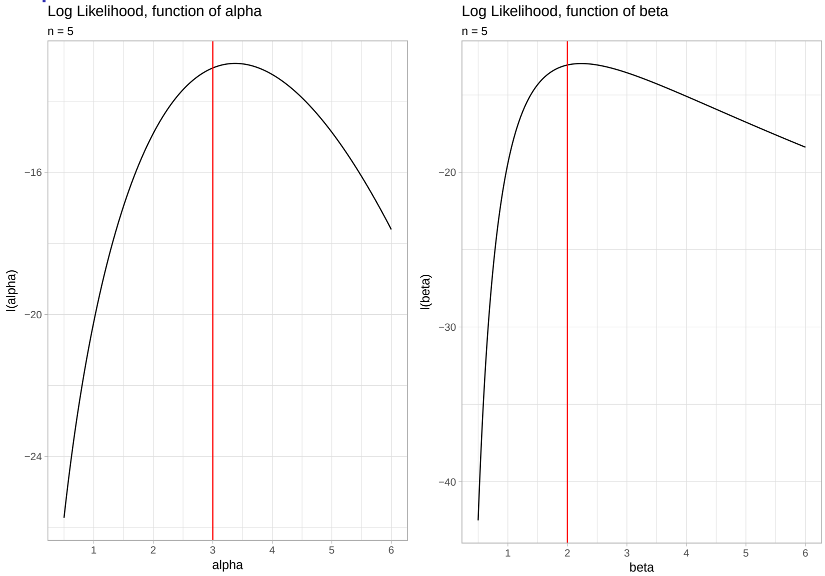
$$f_{X_i}(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right)$$

and log-likeilhood (homework: verify this)

$$\ell(\alpha, \beta) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{1}{\beta} \sum_{i=1}^{n} x_i$$

Let's look at the likelihood for a dataset with $n = 5$ as a function of α, β, for true values $(\alpha, \beta) = (3, 2)$.

Example


Log Likelihood, function of alpha
n = 5
l(alpha)
−16
−20
−24
1
2
3
4
5
6
alpha
Log Likelihood, function of beta
n = 5
l(beta)
−20
−30
−40
1
2
3
4
5
6
beta

Example

We can try to maximize the log-likelihood analytically. We get:

$$\frac{\partial \ell}{\partial \alpha} = -n\psi(\alpha) - n\log\beta + \sum_{i=1}^{n}\log x_i$$

$$\frac{\partial \ell}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i$$

Setting to 0 gives

$$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}}$$

$$0 = -n\psi(\alpha) - n\log\frac{\bar{x}}{\hat{\alpha}} + \sum_{i=1}^{n}\log x_i$$

Note $\psi(x)$ is the *digamma function*, which is just defined as $\psi(x) = \partial\log\Gamma(x)/\partial x$ and has no simple formula.

Example

Even though the likelihood in the α-dimension looked simple when we plotted it, the MLE for α is defined as the solution to a complicated non-linear equation with no closed-form answer.

In general, we can obtain the MLE by employing some sort of root-finding method, e.g. Newton's method, on the partial derivatives of the likelihood function.

But I relied heavily on the closed-form formulae for the MLE in the normal example to describe properties of the sampling distribution. What do we do here?

We need a few more theoretical results.

The Score Vector

We mostly talk about the log-likelihood existing for a fixed dataset, and treat it as a function of the unknown parameters θ, which are themselves fixed constants.

But if instead of plugging observed $\mathbf{x}$ into $\ell(\theta)$, what if we plugged in the random variable $\mathbf{X}$? The result is a function of a random variable, so is itself a random variable.

So is its derivative, $S(\theta) = \frac{\partial \ell}{\partial \theta}$.

Random Functions of θ

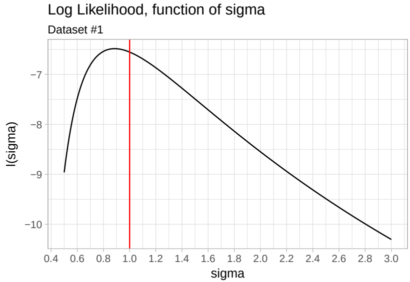
What this means is that

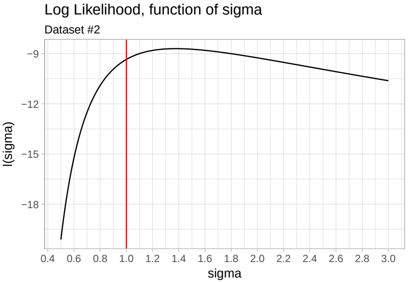
- For any fixed dataset $\mathbf{x}$, the log-likelihood and functions thereof are functions of θ, $\ell(\theta|\mathbf{x})$
- Every observed dataset gives a *different* function of theta
- So, the function of theta we get in a given sample is itself a random variable

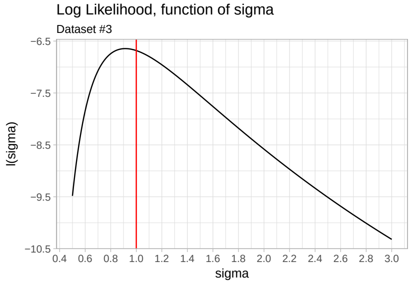
Example

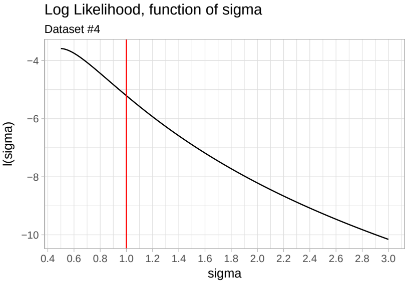
To illustrate this, let me generate a few datasets of size $n = 5$ from that same $N(2, 1)$ distribution from earlier. We'll plot the likelihood for σ for each.

Random Functions of σ









The Score Vector

Definition: the **Score Function**, **Score Statistic**, or **Score Vector** is the vector of partial derivatives of the log-likelihood with respect to the parameter θ,

$$S(\theta) = \frac{\partial \ell}{\partial \theta}$$

When treated as dependent on the observed data, this is just a regular old function, and we have been using it up until now to find the MLE.

When treated as dependent on the random variable $\mathbf{X}$, it is a random variable. Every random sample we generate gives us a new $S(\theta)$.

The Score Vector

Don't over-think this. Consider the whole procedure:

- Observe a dataset
- Plug in the values to the derivative of the log-likelihood
- Get a "value" of $S(\theta)$

Because we'd get a different value for each new dataset, $S(\theta)$ is a random variable.

For example, for the normal distribution, $S(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)$. Each new dataset would give a different function of μ, because the X_i would be different.

The Expected Score Vector

Because $S(\theta)$ is a random variable, it has a sampling distribution. It has a mean, and it has a standard deviation. These are *also* functions of θ.

The mean of the score vector is the average value of the score vector across all possible samples. It's still a function of θ, although not of $\mathbf{X}$ (why?).

The mean/variance of the score vector evaluated at any particular θ_0 are just numbers (well, vectors).

Parameter Space

Define the *parameter space* Ω to be the set of all values that θ can take.

For the normal distribution, $\theta = (\mu, \sigma)$ and $\Omega = (-\infty, \infty) \times (0, \infty)$.

For the binomal, $\theta = p$ and $\Omega = (0, 1)$.

The Expected Score Vector

Let θ_0 be the *true* value of θ, which is the unknown value we're trying to estimate. We do know that $\theta_0 \in \Omega$.

We have been talking about $\ell(\theta)$ and $S(\theta)$ as functions of θ, which they are. But of particular interest are the quantities $\ell(\theta_0)$ and $S(\theta_0)$, the *values* of these functions at the *true parameter value* θ_0.

We know that by definition of the MLE, $\ell(\hat{\theta}) \geq \ell(\theta)$ for all $\theta \in \Omega$.

It is of interest to study the behaviour of these functions at values of θ close to θ_0.

The Expected Score Vector

We'll do our math on the univariate case, $d = 1$, so $\Omega \subset \mathbb{R}$ and θ is just a number.

Proposition: $E(S(\theta_0)) = 0$.

Regularity Conditions

There are some assumptions required for this to be true:

- The true parameter value $\theta_0 \in \Omega_0$, the *interior* of the parameter space. For practical purposes, this just means that Ω is an open subset of $\mathbb{R}^d$. For σ in the normal example, think about the difference between $\Omega = (0, \infty)$ vs $\Omega = [0, \infty)$.
- The support of the distribution of $\mathbf{X}$ doesn't depend on θ. Think about the continuous uniform MLE example, and why we couldn't use calculus there.
- The log-likelihood is *thrice continuously differentiable*, that is, $\ell'''(\theta)$ exists and is continuous. Note that $\ell'''(\theta) = 0$ counts, as is the case for the normal distribution.

There are a few other technical assumptions required, but these are the important ones.

The Expected Score Vector

Proof: The log-likelihood is a sum over the whole dataset,

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta)$$

where $\ell_i(\theta) \equiv \log f(x_i|\theta)$. Because differentiation is a linear operation, we can write

$$S(\theta) = \sum_{i=1}^{n} s_i(\theta)$$

as well, and show that $E(s_i(\theta_0)) = 0$ for each $i = 1 \ldots n$.

The Expected Score Vector

$$
\begin{aligned}
E(s_i) &= \int_x \frac{\partial \ell_i(\theta)}{\partial \theta} f(x_i|\theta_0)dx \\
&= \int_x \frac{\partial \log f(x_i|\theta)}{\partial \theta} f(x_i|\theta_0)dx \\
&= \int_x \frac{1}{f(x_i|\theta)} \frac{\partial f(x_i|\theta)}{\partial \theta} f(x_i|\theta_0)dx
\end{aligned}
$$

The Expected Score Vector

At the point $\theta = \theta_0$,

$$\begin{aligned}
&= \frac{\partial}{\partial\theta}\int_x \frac{1}{f(x_i|\theta_0)} \times f(x_i|\theta_0) \times f(x_i|\theta_0)dx \\
&= \frac{\partial}{\partial\theta}\int_x f(x_i|\theta_0)dx \\
&= \frac{\partial}{\partial\theta}(1) \\
&= 0
\end{aligned}$$

What just happened?

We just showed that, across all possible datasets, the score vector *for each datapoint* is, on average, equal to zero at $\theta = \theta_0$.

But the θ at which the score vector equals zero is the θ at which $\ell(\theta)$ is maximized.

This is suggestive of a nice property of $\hat{\theta}$.

How close to 0 is $S(\theta_0)$ likely to be in any given sample?

Variance of the Score

Consider the variance of an individual score element, $Var(s_i(\theta_0)) = E(s_i(\theta_0)^2)$ (because $E(s_i(\theta_0)) = 0$):

$$E(s_i(\theta)^2) = \int_x \left(\frac{\partial \log f(x_i|\theta)}{\partial \theta} \right)^2 f(x_i|\theta_0) dx$$

We showed previously that

$$0 = \int_x \frac{\partial \log f(x_i|\theta_0)}{\partial \theta} f(x_i|\theta_0) dx$$

Differentiate both sides of that identity to obtain

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int_x \frac{\partial \log f(x_i|\theta_0)}{\partial \theta} f(x_i|\theta_0) dx \\ &= \int_x \frac{\partial^2 \log f(x_i|\theta_0)}{\partial \theta^2} f(x_i|\theta_0) dx + \int_x \left(\frac{\partial \log f(x_i|\theta_0)}{\partial \theta} \right)^2 f(x_i|\theta_0) dx \end{aligned}$$

Variance of the Score

The first term is

$$\int_x \frac{\partial^2 \log f(x_i|\theta_0)}{\partial \theta^2} f(x_i|\theta_0) dx = E\left(\frac{\partial^2 \log f(x_i|\theta_0)}{\partial \theta^2}\right)$$

which is the expected curvature of the log likelihood at the true parameter value.

Variance of the Score

The second term is

$$\int_x \left(\frac{\partial \log f(x_i|\theta_0)}{\partial \theta} \right)^2 f(x_i|\theta_0)dx = E(s_i(\theta_0)^2)$$

We just showed that

$$Var(s_i(\theta_0)) = -E\left(\frac{\partial^2 \log f(x_i|\theta_0)}{\partial \theta^2} \right)$$

Remember before, when I said the range of plausible values for θ defined by the likelihood $\ell(\theta)$ depended on its curvature?

"Expected Curvature"

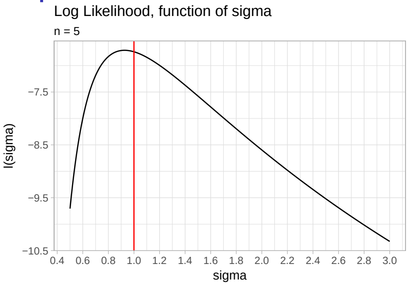
If you were just wrapping your head around the idea of the derivative of the log-likelihood being a random variable with a sampling distribution, then the idea of the "expected curvature" probably sounds pretty far-out.

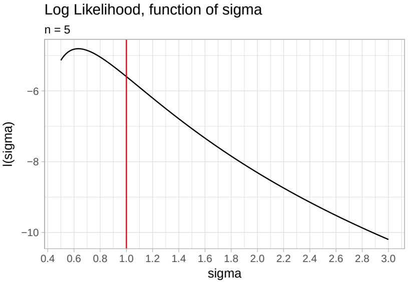
Remember: those curves I plotted earlier were defined by the observed data that I used to generate them. Different data leads to different curves.

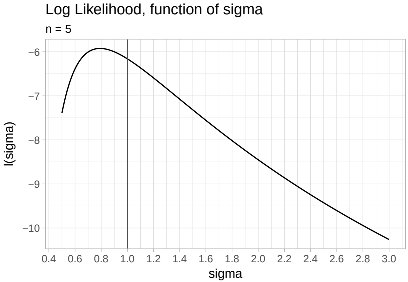
And, different curvature.

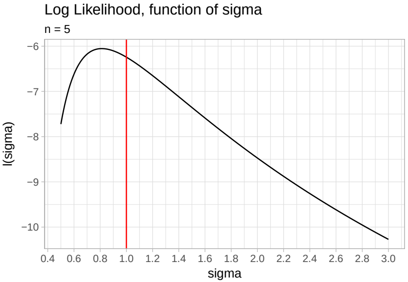
Let's look again at a few different log-likelihoods for σ, and this time pay special attention to the peaked/flatness around the **true value** $\theta = \theta_0$.

Example


Log Likelihood, function of sigma
n = 5
l(sigma)
−7.5
−8.5
−9.5
−10.5
0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0
sigma


Log Likelihood, function of sigma
n = 5
l(sigma)
−6
−8
−10
0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0
sigma


Log Likelihood, function of sigma
n = 5
l(sigma)
−6
−7
−8
−9
−10
0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0
sigma


Log Likelihood, function of sigma
n = 5
l(sigma)
−6
−7
−8
−9
−10
0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0
sigma

Information

The curvature of the log-likelihood is fundamentally related to our ability to estimate θ using the observed data.

We define the **Fisher Information** for a datapoint as

$$\begin{aligned} I_i(\theta) &= E\left(\left(\frac{\partial \ell(\theta|x_i)}{\partial \theta}\right)^2\right) \\ &= Var(s_i(\theta)) \end{aligned}$$

We have showed that

$$I_i(\theta_0) = -E\left(\frac{\partial^2 \ell(\theta|x_i)}{\partial \theta^2}\right)\Bigg|_{\theta=\theta_0}$$

Information in the Sample

Because differentiation and expectation are both linear operations, the information in the sample is the sum of the information in each datapoint (for IID data):

$$\begin{aligned} I(\theta|\mathbf{x}) &= -E\left(\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial \theta^2}\right) \\ &= -E\left(\frac{\partial^2 \sum_{i=1}^n \ell(\theta|x_i)}{\partial \theta^2}\right) \\ &= \sum_{i=1}^n I_i(\theta) \end{aligned}$$

Information in the Sample

But this expectation is taken across x, so under the IID assumption, $I_i(\theta) \equiv I_0(\theta)$, and we have

$$I(\theta) = nI_0(\theta)$$

That is, the Fisher Information for the sample is n times the information for a single datapoint.

Observed Information

In general, the expectation involved in calculating $I_i(\theta)$ may or may not be tractable.

In finite samples, we can *estimate* the Fisher Information using the data,

$$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2 \ell(\theta|x_i)}{\partial \theta^2} = -\frac{\partial^2 \sum_{i=1}^{n} \ell(\theta|x_i)}{\partial \theta^2}$$

Observed Information

The observed information for each datapoint won't generally be equal.

Consider the *average* observed information for a datapoint,

$$\frac{1}{n}J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ell(\theta|x_i)}{\partial \theta^2}$$

The LLN implies that this is a consistent estimator of $I_0(\theta)$, which motivates the use of $J(\theta)$ to estimate $I(\theta)$ in finite samples.

Remember though, this is only a trick to use if you can't evaluate the expectation required to get the real Fisher Information.

Example

That was a lot of theory, so let's revisit the normal example. We have for μ,

$$\ell(\mu) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$
$$S(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$
$$J(\mu) = \frac{n}{\sigma^2}$$
$$I(\mu) = E(J(\mu)) = \frac{n}{\sigma^2}$$

Because the data cancels out of the second derivative, the observed and Fisher information are the same for this example.

Example

For σ^2,

$$\ell(\sigma^2) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$S(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$J(\sigma^2) = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\begin{aligned} I(\sigma^2) = E(J(\sigma^2)) &= -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^{n}E\,(x_i - \mu)^2 \\ &= -\frac{n}{2\sigma^4} + \frac{n}{\sigma^4} \\ &= \frac{n}{2\sigma^4} \end{aligned}$$

Consistency

Notice how

$$\frac{1}{n}J(\sigma^2) = -\frac{1}{2\sigma^4} + \frac{1}{n\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2$$

provides a consistent estimator of

$$\frac{1}{n}I(\sigma^2) = \frac{1}{2\sigma^4}$$

We spoke of this in the general case, but it helps to take note of it in specific examples like this.

Multiparameter case

All of these results hold in the case where the dimension of θ, $d > 1$. The score vector is a vector having mean equal to a vector of zeroes.

The Fisher Information is now a matrix:

$$I(\theta) = -E\left(\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\right)$$

which is a matrix with i, j element equal to

$$I(\theta)_{ij} = -E\left(\frac{\partial^2 \ell(\theta)}{\partial\theta_i\partial\theta_j}\right)$$

It's the negative expected *Hessian* of $\ell(\theta)$.

The observed information is just the negative Hessian, i.e. the Fisher Information without the E.

Multiparameter case

While in general, matrix calculus is used to find these, that's mostly done in statistical modelling, when the number of parameters is large (or at least greater than 2 or 3).

For our purposes, it's more practical to just deal with each parameter separately and stack the results in a vector/matrix.

Example

For the normal example, we found previously

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

$$J(\mu) = \frac{n}{\sigma^2}$$

$$I(\mu) = E(J(\mu)) = \frac{n}{\sigma^2}$$

$$S(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$J(\sigma^2) = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$I(\sigma^2) = E(J(\sigma^2)) = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^{n}E\,(x_i - \mu)^2$$

Example

The only missing piece is the off-diagonal element of $I(\theta)$,

$$\begin{aligned} I(\mu,\sigma)_{1,2} = I(\mu,\sigma)_{2,1} &= -E\left(\frac{\partial^2 \ell(\mu,\sigma)}{\partial\mu\partial\sigma^2}\right) \\ &= -E\left(\frac{\partial^2 \ell(\mu,\sigma)}{\partial\sigma^2\partial\mu}\right) \end{aligned}$$

It doesn't matter whether you differentiate $S(\mu)$ by σ^2 or the other way around, you'll get the same answer if the function is twice continuously differentiable. We assumed it was *thrice* continuously differentiable, so you're good to just pick the one that looks easier.

Example

$$
\begin{aligned}
-E\left(\frac{\partial^2 \ell(\mu,\sigma)}{\partial\sigma^2 \partial\mu}\right) &= -E\left(\frac{\partial}{\partial\sigma^2}\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\right) \\
&= -E\left(-\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)\right) \\
&= \frac{1}{\sigma^4}\sum_{i=1}^{n}E\left(x_i-\mu\right) \\
&= 0
\end{aligned}
$$

This doesn't always happen. It is one of the special properties of the normal distribution

Summary

To summarize, when asked to get the score vector and fisher information for a multiparameter problem:

- Compute the derivatives of the log likelihood with respect to each parameter; stack these in a vector, and that's your score vector
- Compute the second derivatives of the log likelihood, including all mixed partials; stack these in a matrix, and that's your *observed* information
- Compute the expectation of each element in this matrix (with respect to x). That's your Fisher Information
- If you can't compute the expectations because the expressions are too messy, just stick with the observed information

Summary so far

That was a lot of work. But now we have:

$$E(S(\theta_0)) = 0$$
$$Var(S(\theta_0)) = I(\theta_0)$$

Also, we can express the score function as a sum of independent contributions from each datapoint:

$$S(\theta_0) = \sum_{i=1}^{n} s_i(\theta_0)$$

each of which has mean 0 and variance $I_0(\theta_0)$ at the true parameter value θ_0.

What should we do with this information?

A Central Limit Theorem

Under all the conditions necessary for the facts on the previous slide to be true,

$$\frac{S(\theta_0)}{\sqrt{I(\theta_0)}} \xrightarrow{d} N(0,1)$$

Because of that super confusing example from lecture 2 regarding replacement of the standard deviation of the sum in the denominator with a consistent estimate, we also have

$$\frac{S(\theta_0)}{\sqrt{J(\theta_0)}} \xrightarrow{d} N(0,1)$$

A Central Limit Theorem

This isn't that useful in its own right. But remember, the value θ at which $S(\theta)$ equals 0 is, by definition, the MLE.

So on average, the score function generated by our sample is maximized *at the true value of* θ, θ_0.

We can do even better.

Consistency of the MLE (textbook, 275 - 276)

But first, let's revist a question I asked at the end of last lecture. Is the MLE consistent?

That is, does $\hat{\theta} \xrightarrow{p} \theta_0$?

Proposition: the MLE is consistent.

Consistency of the MLE (textbook, 275 - 276)

Proof: This is only a sketch of the fully rigorous proof, which we don't have time for. Consider the quantity

$$\frac{1}{n}\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ell_i(\theta)$$

We can express the log-likelihood as a sum of independent contributions from each datapoint under the assumption of IID sampling. This is a sample mean of independent quantities, so by the LLN,

$$\frac{1}{n}\ell(\theta) \xrightarrow{p} E\log f(X|\theta)$$

The best we can do for a proof that $\hat{\theta} \xrightarrow{p} \theta_0$ at this point is to show that θ_0 maximizes $E\log f(X|\theta)$, and then argue that since $\hat{\theta}$ maximizes $\frac{1}{n}\ell(\theta)$ and $\frac{1}{n}\ell(\theta) \xrightarrow{p} E\log f(X|\theta)$, $\hat{\theta} \xrightarrow{p} \theta_0$.

Consistency of the MLE (textbook, 275 - 276)

To maximize $E \log f(X|\theta)$, take a derivative

$$\begin{aligned}
\frac{\partial}{\partial \theta} E \log f(X|\theta) &= \frac{\partial}{\partial \theta} \int_x \log f(X|\theta) f(X|\theta_0) dx \\
&= \int_x \frac{\partial}{\partial \theta} \log f(X|\theta) f(X|\theta_0) dx \\
&= \int_x \frac{1}{f(X|\theta)} \frac{\partial}{\partial \theta} f(X|\theta) f(X|\theta_0) dx
\end{aligned}$$

Consistency of the MLE (textbook, 275 - 276)

If $\theta = \theta_0$ then this becomes

$$
\begin{aligned}
&= \int_x \frac{1}{f(X|\theta_0)} \frac{\partial}{\partial\theta} f(X|\theta_0) f(X|\theta_0) dx \\
&= \int_x \frac{\partial}{\partial\theta} f(X|\theta_0) dx \\
&= \frac{\partial}{\partial\theta} \int_x f(X|\theta_0) dx \\
&= \frac{\partial}{\partial\theta} (1) \\
&= 0
\end{aligned}
$$

Consistency of the MLE (textbook, 275 - 276)

From this, we argue that since $\hat{\theta}$ maximizes $\frac{1}{n}\sum_{i=1}^{n}\ell(\theta)$, θ_0 maximizes $E\log f(X|\theta)$, and $\frac{1}{n}\sum_{i=1}^{n}\ell(\theta) \xrightarrow{p} E\log f(X|\theta)$, $\hat{\theta} \xrightarrow{p} \theta_0$.

Thus, the MLE is a consistent estimator for θ.

This is the level of rigour we will use at this point. More rigorous arguments for this can be covered in upper year/graduate courses on the theory of likelihood inference.

Another Central Limit Theorem (Textbook, 277 - 278)

Now we state and prove one of the fundamental results of statistical inference.

Theorem: Asymptotic Distribution of the MLE: Under all of the same conditions as before,

$$\sqrt{I(\theta_0)}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N(0,1)$$

Another Central Limit Theorem (Textbook, 277 - 278)

Proof: Approximate the score function at the MLE using a Taylor expansion about the true value θ_0. Remember that the score function at the MLE is 0 by definition:

$$
\begin{aligned}
0 = S(\hat{\theta}) &\approx S(\theta_0) + (\hat{\theta} - \theta_0)S'(\theta_0) \\
\implies (\hat{\theta} - \theta_0) &\approx \frac{S(\theta_0)}{J(\theta_0)}
\end{aligned}
$$

where we defined $J(\theta) = -S'(\theta)$ previously, though we didn't use that exact notation.

Another Central Limit Theorem (Textbook, 277 - 278)

At this point, the proof in the textbook gets really confusing, at least to me. I think the following is clearer.

We have

$$\begin{aligned}\sqrt{I(\theta_0)}(\hat{\theta} - \theta_0) &\approx \sqrt{I(\theta_0)}\frac{S(\theta_0)}{J(\theta_0)} \\ &= \frac{I(\theta_0)}{J(\theta_0)} \times \frac{S(\theta_0)}{\sqrt{I(\theta_0)}} \\ &\xrightarrow{p} (1) \times Z\end{aligned}$$

where $Z \sim N(0, 1)$. The term on the left works because we showed earlier that $J(\theta_0) \xrightarrow{p} I(\theta_0)$, and the term on the right is the central limit theorem for the score vector from a few slides back. The result then follows by Slutsky's lemma for multiplication.

Major Result!

This is a major result for two reasons:

- The theorem itself will be used later in the course to develop an extremely general theory of *hypothesis testing*
- In finite samples, we say that:

The Maximum Likelihood Estimator is approximately normallly distributed with mean equal to the true value θ_0 *and variance equal to the inverse Fisher Information,* $1/I(\theta_0)$.

In practice, we don't know θ_0, so to evaluate the variance of the MLE we plug the MLE itself into the Fisher information, which is justified because $\hat{\theta}$ is consistent for θ_0.

And if we can't get at the Fisher Information, we just use the observed information.

Examples

Let's do some examples to illustrate why this is so good.

Let $X_i \sim N(\mu, \sigma)$. Find the MLE for $\theta = (\mu, \sigma)$, and its asymptotic variance.

We saw that $\hat{\mu} = \bar{X}$, and $I(\mu) = n/\sigma^2$.

The CLT for the MLE says that $E(\hat{\mu}) \to \mu_0$ and $Var(\hat{\mu}) \to 1/I(\mu_0)$

But when the data is already normal, we see that these approximations are actually exact. This is a special property of the Normal distribution.

Examples

For σ^2 we had $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$. The CLT for the MLE says that $E(\hat{\sigma}^2) \to \sigma^2$ and $Var(\hat{\sigma}^2) \to 1/I(\sigma_0^2)$.

Again, we can directly evaluate:

$$\begin{aligned} E(\hat{\sigma}^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i - \mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 \\ &= \sigma^2 \end{aligned}$$

The normal approximation for the MLE is exact when the data is already normal.

Example

While this seems trivial in theory, it *does* solve a problem we had before: what to do when estimating both quantities at once? In the above, I fixed σ^2 to estimate μ, and vice-versa.

When there is variability from both sources, we have

$$\hat{\sigma}^2 = \left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 \right) = \left(\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \right)$$

We will derive its exact mean later- but the CLT for the MLE still guarantees that $E(\hat{\sigma}^2) \to \sigma^2$, even when we use $\hat{\mu}$ to compute $\hat{\sigma}^2$.

In general, the MLE for one parameter may be a function of the other parameters, and you'll have to plug in *their* MLEs.

Multidimensional Case

The same central limit theorems hold when θ is a vector.

However, since the information is a matrix,

$$I(\theta)_{ij} = -E\left(\frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial_{\theta_j}}\right)$$

The asymptotic *covariance matrix* is given by the *inverse* of the information matrix.

So to get the variance/covariance of MLEs when there is more than one parameter,

- Compute the Fisher information (or the observed information)
- Invert it
- The i, i element of this is the variance of $\hat{\theta}_i$, and the i, j element is the covariance between $\hat{\theta}_i$ and $\hat{\theta}_j$.

Multidimensional Case

In the normal example, find $Cov\left(\hat{\mu}, \hat{\sigma^2}\right)$.

We already showed that

$$I(\mu, \sigma^2)_{12} = I(\mu, \sigma^2)_{21} = 0$$

so the information matrix is diagonal, and $\hat{\mu}$ and $\hat{\sigma^2}$ are *asymptotically uncorrelated*.

Example

Because the information matrix is diagonal, it is the case in this example that

$$Var(\hat{\mu}) = \frac{1}{I(\hat{\mu})}$$
$$Var(\hat{\sigma}^2) = \frac{1}{I(\hat{\sigma}^2)}$$

We see that the asymptotic variance is $Var(\hat{\mu}) = \sigma^2/n$, which is actually equal to the exact variance.

To use this in practice, plug in $\hat{\sigma}^2$ for σ^2.

Example

The variance of $\hat{\sigma}^2$ would be more annoying to derive directly. We saw earlier that

$$I(\sigma^2) = \frac{n}{2\sigma^4}$$

So in practice for finite n, we can approximate the variance of $\hat{\sigma}^2$ by

$$Var(\hat{\sigma}^2) \approx \frac{2\sigma^4}{n}$$

To use this in practice, plug in $\hat{\sigma}^2$ for σ^2. It's okay that the variance of $\hat{\sigma}^2$ is a function of $(\hat{\sigma}^2)$.

Example

Remember the Gamma example from the beginning, where we couldn't find an expression for $\hat{\alpha}$?

I told you we could obtain estimates of α numerically. Well, our theory still applies, and now we can get at the approximate sampling distribution of this estimator that we can't even find a closed-form expression for.

The true parameter values in this example were $(\alpha, \beta) = (3, 2)$, and the log likelihood is

$$\ell(\alpha, \beta) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{1}{\beta} \sum_{i=1}^{n} x_i$$

Example

The score vector is

$$S(\alpha) = -n\psi(\alpha) - n\log\beta + \sum_{i=1}^{n}\log x_i$$

$$S(\beta) = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i$$

Example

The Observed Information is (don't forget the negative):

$$J(\alpha, \beta)_{\alpha,\alpha} = n\psi_1(\alpha)$$

$$J(\alpha, \beta)_{\beta,\beta} = -\frac{n\alpha}{\beta^2} + \frac{2}{\beta^3}\sum_{i=1}^{n} x_i$$

$$J(\alpha, \beta)_{\alpha,\beta} = \frac{n}{\beta}$$

Note: $\psi_1(\alpha)$ is the *trigamma* function, defined as $\psi_1(x) = \frac{\partial^2 \log \Gamma(x)}{\partial x^2}$. I am not making this up.

Example

The Fisher Information is

$$I(\alpha,\beta)_{\alpha,\alpha} = E\left(J(\alpha,\beta)_{\alpha,\alpha}\right) = n\psi_1(\alpha)$$

$$I(\alpha,\beta)_{\beta,\beta} = E\left(J(\alpha,\beta)_{\beta,\beta}\right) = -\frac{n\alpha}{\beta^2} + \frac{2}{\beta^3}\sum_{i=1}^{n} E(x_i)$$

$$I(\alpha,\beta)_{\alpha,\beta} = E\left(J(\alpha,\beta)_{\alpha,\beta}\right) = \frac{n}{\beta}$$

For the Gamma distribution, $E(X) = \alpha\beta$, so we can write

$$I(\alpha,\beta)_{\beta,\beta} = \frac{n\alpha}{\beta^2}$$

Example

The Fisher Information is therefore

$$n \begin{pmatrix} \psi_1(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}$$

We can apply the formula for the inverse of a 2×2 matrix to get an explicit expression for the asymptotic variance matrix of $(\hat{\alpha}, \hat{\beta})$, but in practice it is more common to calculate the Fisher Information directly (for a given sample and value of the MLE), then invert it numerically.

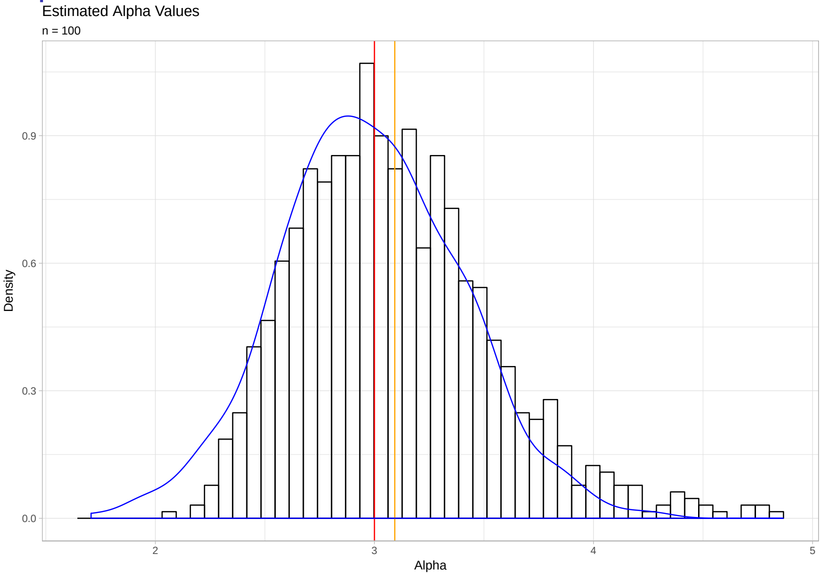
This is because in general, $d > 2$.

Example

Let's take a look at the sampling distribution of $\hat{\alpha}$ for this example. I am going to

- Sample some random datasets of size $n = 100$ from a $Gamma(3, 2)$ distribution
- For each sample,
  - Find $\hat{\alpha}, \hat{\beta}$ numerically, and plot a normalized histogram of $\hat{\alpha}$ values
  - Calculate the Fisher Information matrix at $\hat{\alpha}$ and $\hat{\beta}$
- Average the resulting estimates of the variance of $\hat{\alpha}$, and compare this to the emprical variance of $\hat{\alpha}$ from the values that I calculate
- Overlay a normal curve with the true mean and standard deviation

Example

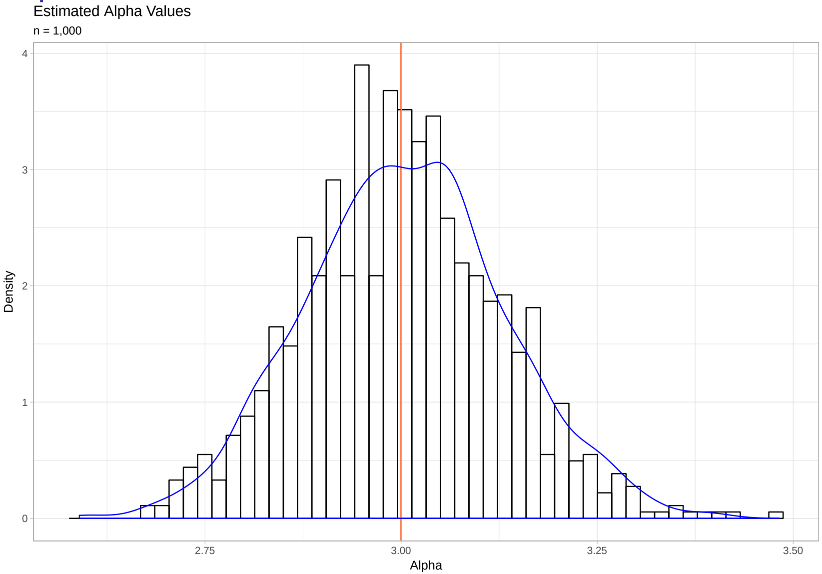


Example

Looks pretty close. What about as we increase the dataset size?

Let's look again for $n = 1000$.

Example


Estimated Alpha Values
n = 1,000
Density
Alpha
0
1
2
3
4
2.75
3.00
3.25
3.50