STA261: Week 3

Sufficiency and Likelihood Inference I

Alex Stringer

Jan 22nd - 26th, 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

Recap

Last week, we learned about **consistency**, which is Property 1 of an estimator: as we get more and more data, estimates from this estimator should get closer and closer to the parameter they are estimating.

We also learned the Method of Moments, which gives consistent estimators algorithmically.

This Week

This week we are going to talk about Property **2**: We should base our estimator off of all the information in the sample. Our estimator should be a *summary* of the full sample. Whether we know the entire dataset or just our estimate $\hat{\theta}$, we should make the same conclusions regarding θ.

Let's look at an example.

Use all of the data

Suppose we have $X_1, X_2 \sim N(\mu, 1)$ independently, and we wish to estimate μ. I propose the following three estimators:

1. $\hat{\mu}_1 : X_1$
2. $\hat{\mu}_2 : X_2$
3. $\hat{\mu}_3 : \frac{X_1+X_2}{2}$

Which do you prefer?

Use all of the data

It may seem like a silly question; of course we want to use $\bar{X}$ instead of a single data point X_1. But answering *why* is the hard part.

Let's formulate the question a different way: given that we have observed a particular value of X_1, does observing X_2 change the amount of information in the sample about μ?

Conditioning on an estimator

Let's look at what happens if we take $\hat{\mu} = X_1$. I told you X_1 and X_2 were independent, so the conditional distribution of X_2 given X_1 is just

$$f_{X_2|X_1}(x_2|x_1) = f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_2 - \mu)^2\right)$$

This conditional distribution is a function of μ. So even knowing the value of our estimator $\hat{\mu} = X_1$, the values that X_2 are likely to take on will depend on μ.

Therefore, knowing X_2 gives us *information* about μ.

Reminder

Reminder from intro stats or 257: if $X \sim N(\mu, \sigma)$, then

$$P(-2\sigma < X - \mu < 2\sigma) \approx 95\%$$

About 95% of the mass of a normal distribution lies within 2 standard deviations of the mean.

We think anything farther away than 2σ from μ is pretty implausible.

Example

Suppose we observe $(X_1, X_2) = (4, 6)$. We know $\sigma = 1$ (it was given in the problem statement).

If $X_1 = 4$, what is the range of plausible values for the parameter μ?

- $\mu = 3$ wouldn't be weird; X_1 would be 1 SD from the mean
- $\mu = 2$ wouldn't be weird; X_1 would be 2 SD from the mean
- $\mu = 5$ wouldn't be weird; X_1 would be 1 SD from the mean
- $\mu = 6$ wouldn't be weird; X_1 would be 2 SD from the mean

Obviously $\mu = 4$ is the least weird value- our "best guess" with the information we have.

Example

If $X_2 = 6$, what are the range of plausible values for the parameter μ?

- $\mu = 5$ wouldn't be weird; X_2 would be 1 SD from the mean
- $\mu = 4$ wouldn't be weird; X_2 would be 2 SD from the mean
- $\mu = 7$ wouldn't be weird; X_2 would be 1 SD from the mean
- $\mu = 8$ wouldn't be weird; X_2 would be 2 SD from the mean

Obviously $\mu = 6$ is the least weird value- our "best guess" with the information we have.

Example

But now the question: having observed $X_1 = 4$, does observing $X_2 = 6$ *change* what we think the plausible values of μ could be?

- $\mu = 2$
- $\mu = 3$
- $\mu = 4$
- $\mu = 5$
- $\mu = 6$
- $\mu = 7$
- $\mu = 8$

Conditioning on an estimator

Now consider $\hat{\mu} = \bar{X}$, so in our example $\hat{\mu} = \frac{X_1+X_2}{2} = 5$. We may consider the conditonal distribution of $X_2|\bar{X} = t$.

It's a little tricky to derive without using some multivariate stats, which you are not responsible for in this course. So for this example I'll just tell you that

$$X_2|\bar{X} = t \sim N(t, 1/2)$$

This does not depend on μ.

Conditioning on an estimator

In particular, once we have observed $\bar{X} = t$, knowing the specific value of X_2 does not change what values of μ we might think are plausible. That is, $\bar{X}$ contains all the *information* in the sample about μ, such that *knowing* $\bar{X}$ *renders the rest of the sample uninformative about* μ.

Keep in mind that when we say "information", we are talking about information with respect to a particular parameter. Knowing the rest of the sample is still necessary if we want to estimate other things, or check model assumptions.

Conditioning on an estimator

For example, the following two samples give the same information about μ:

$$\mathbf{x}_1 = (1, 1.1, 0.9)$$
$$\mathbf{x}_2 = (-9999, 3, 9999)$$

but we wouldn't think those data came from the same distribution. Specifically, they contain different information about σ.

Check out Anscombe's Quartet:
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Definition: Sufficiency (Textbook, section 8.7)

We can formalize this notion as follows.

Defintion: An estimator $\hat{\theta}(X_1, \ldots, X_n)$ is said to be **sufficient** for the parameter θ if the conditional distribution of the sample $X_1, \ldots, X_n$ given $\hat{\theta} = t$ does not depend on θ, for any t.

Notes

The textbook uses the notation $T(X_1, \ldots, X_n)$ instead of $\hat{\theta}(X_1, \ldots, X_n)$- sorry!

Technically, any function that takes in data and returns a (possibly lower-dimensional) summary is called a **statistic**, so you may hear the term *sufficient statistic* in the literature, or from me.

Remember how we defined an estimator last week: as a *function* $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^m$ that takes in realizations of a set of random variables ("data") and returns an estimate of a parameter.

Sufficiency says that knowing the sufficient statistic and knowing the whole sample gives us the same information about which values of θ were likely to have generated the observed data.

Notation

By the way, I am also going to use two notations for vectors interchangably:

$$\mathbf{X} = (X_1, \ldots, X_n)$$
$$\mathbf{x} = (x_1, \ldots, x_n)$$

It should always be clear from the context when I mean X to be a vector, and you should get used to seeing either notation.

Upper case $\mathbf{X}$ will still mean random variable, and lower case $\mathbf{x}$ will still mean realization of random variable, i.e. a datapoint.

Example (textbook, page 306)

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter θ. Show $\hat{\theta} = \sum_{i=1}^{n} X_i$ is sufficient for θ.

We need to evaluate

$$P(X_1 = x_1, \ldots, X_n = x_n | \hat{\theta} = t) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, \hat{\theta} = t)}{P(\hat{\theta} = t)}$$

and then we can just look at whether or not the result is a function of θ.

Example (textbook, page 306)

In this case we can just work out all the probabilities we need from scratch.

Notice that $\hat{\theta} \sim Bin(n, \theta)$ so $P(\hat{\theta} = t) = \binom{n}{t}\theta^t(1-\theta)^{n-t}$.

Example (textbook, page 306)

For the numerator, first notice that the value of $\hat{\theta}$ is completely determined by the values of $X_1, \ldots, X_n$, so really the joint probability of the sample and the estimator is just the probability of the sample itself. This is the probability of seeing any particular combination of t 1's and $n - t$ 0's, which is equal to the binomial probability function *without* the first factor (why?)

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n, \hat{\theta} = t) &= P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \theta^t (1 - \theta)^{n-t} \end{aligned}$$

Example (textbook, page 306)

We can evaluate directly:

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n | \hat{\theta} = t) &= \frac{P(X_1 = x_1, \ldots, X_n = x_n, \hat{\theta} = t)}{P(\hat{\theta} = t)} \\ &= \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} \\ &= \frac{1}{\binom{n}{t}} \end{aligned}$$

which does not depend on θ. Hence $\hat{\theta} = \sum_{i=1}^{n} X_i$ is sufficient for θ.

Factorization

You saw that computing this conditional distribution is annoying, and in some cases, not possible analytically. Even in the simple case of the normal conditional on $\overline{X}$, I had to use specifc theorems about the normal distribution that we don't have time to get in to. And even that would only have worked for that example, not in general.

We need a better way to find sufficient estimators, and test if an estimator is sufficient for a parameter.

Factorization Theorem

This is sometimes called "Neyman Factorization", or just "The Factorization Theorem".

Let $\hat{\theta}$ be an estimator of θ. Then $\hat{\theta}$ is sufficient for θ **if and only if** the joint density of $X_1, \ldots, X_n$ can be factored as follows:

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = g(\hat{\theta}, \theta) \times h(\mathbf{x})$$

Example

This looks still to be pretty abstract, but it happens remarkably often.

E.g. if $X_1, \ldots, X_n \sim N(\mu, 1)$ independently then the joint density is

$$
\begin{aligned}
f(\mathbf{x}) &= \prod_{i=1}^{n} f_{x_i}(x_i) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \times \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \times \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \times \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \times \exp\left(-\frac{1}{2}\left(-2n\bar{x}\mu + n\mu^2\right)\right) \\
&= h(\mathbf{x}) \times g(\bar{x}, \mu)
\end{aligned}
$$

Example

In the Bernoulli example from before, the joint density of the sample is

$$P(X_1 = x_1, \ldots, X_n = x_n|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$
$$= h(\mathbf{x}) \times g\left(\sum_{i=1}^n X_i, \theta\right)$$

It's not cheating to take $h(\mathbf{x}) = 1$.

Proof (Textbook, page 307)

We'll prove the factorization theorem. The textbook says that the continuous case is too hard (actually, it says "subtle"). This is only because accurately modifying the expression $P(\hat{\theta} = t)$ so that we don't divide by 0 requires too much detail to bother with here.

Let's look at the discrete case.

Proof (Textbook, page 307)

($\Leftarrow$): Suppose $P(\mathbf{X} = \mathbf{x}) = g(\hat{\theta}, \theta) \times h(\mathbf{x})$. Then

$$\begin{aligned} P(\hat{\theta} = t) &= \sum_{\mathbf{x}|\hat{\theta}(\mathbf{x})=t} P(\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x}|\hat{\theta}(\mathbf{x})=t} g(\hat{\theta}, \theta) \times h(\mathbf{x}) \\ &= g(t, \theta) \times c(\mathbf{x}) \end{aligned}$$

where $c(\mathbf{x})$ is a constant that does not depend on θ.

Proof (Textbook, page 307)

The joint distribution $P(\mathbf{X} = \mathbf{x}, \hat{\theta} = t)$ is actually equal to the marginal distribution of the sample, $P(\mathbf{X} = \mathbf{x})$. This is because the value of $\mathbf{X}$ (which remember, we are denoting $\mathbf{x}$) completely determines the value of $\hat{\theta}(\mathbf{x})$.

Another way of putting this is to say that $P(\hat{\theta} = t | \mathbf{X} = \mathbf{x}) = 1$, which implies

$$P(\mathbf{X} = \mathbf{x}, \hat{\theta} = t) = P(\mathbf{X} = \mathbf{x})$$

Proof (Textbook, page 307)

We can then evaluate the conditional distribution

$$\begin{aligned} P(\mathbf{X}=\mathbf{x}|\hat{\theta}=t) &= \frac{P(\mathbf{X}=\mathbf{x},\hat{\theta}=t)}{P(\hat{\theta}=t)} \\ &= \frac{P(\mathbf{X}=\mathbf{x})}{P(\hat{\theta}=t)} \\ &= \frac{g(t,\theta)\times h(\mathbf{x})}{g(t,\theta)\times c(\mathbf{x})} \\ &= \frac{h(\mathbf{x})}{c(\mathbf{x})} \end{aligned}$$

which doesn't depend on θ. Hence $\hat{\theta}$ is sufficient for θ.

Proof (Textbook, page 307)

($\Rightarrow$): now suppose $\hat{\theta}$ is sufficient for θ. Then $P(\mathbf{X} = \mathbf{x}|\hat{\theta})$ doesn't depend on θ. We can write

$$P(\mathbf{X} = \mathbf{x}, \hat{\theta} = t) = P(\mathbf{X} = \mathbf{x}) = P(\hat{\theta} = t) \times P(\mathbf{X} = \mathbf{x}|\hat{\theta} = t)$$

But in this case, $P(\mathbf{X} = \mathbf{x}|\hat{\theta})$ doesn't depend on θ, so we have

$$P(\mathbf{X} = \mathbf{x}) = g(\hat{\theta}, \theta) \times h(\mathbf{x})$$

as was to be shown.

Example

We can look at more examples.

Let $X_i \sim Gamma(\alpha, 1), i = 1 \ldots n$, so that each X_i has density

$$f_{X_i}(x_i) = \frac{1}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-x_i}$$

Show $\hat{\alpha} = \prod_{i=1}^{n} x_i$ is sufficient for α.

Example

Compute the joint density

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

$$= \left(\frac{1}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp\left(-\sum_{i=1}^{n} x_i\right)$$

which factors with

$$g\left(\prod_{i=1}^{n} x_i, \alpha\right) = \left(\frac{1}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1}$$

$$h(\mathbf{x}) = \exp\left(-\sum_{i=1}^{n} x_i\right)$$

Note

Like with consistency, just because an estimator happens to be sufficient for a parameter doesn't mean that it's a good estimator overall.

For example, $\hat{\alpha} = \frac{\prod_{i=1}^{n} x_i}{1{,}000{,}000}$ is also sufficient for α in the previous example.

One-to-one Functions of Sufficient Statistics are Sufficient

In fact, any one-to-one function of a sufficient statistic is sufficient.

Prove this on assignment 3.

Motivation

There's some further motivation, though, for basing estimators off of sufficient statistics.

Recall the fourth property mentioned last week, which we haven't discussed in detail yet: we want our estimator to have *low variance*.

Rao-Blackwell Theorem (textbook, pg 310)

Let $\hat{\theta}$ be any estimator of θ, having finite variance ($E(\hat{\theta}^2) < \infty$). Let T be any sufficient statistic (for θ), and let $\tilde{\theta} = E(\hat{\theta}|T)$. Then

$$Var(\tilde{\theta}) \leq Var(\hat{\theta})$$

with equality if and only if $\hat{\theta} = \tilde{\theta}$.

I've stated the theorem slightly differently than the textbook, but the proof in the book shows that this version is equivalent.

Proof: Assignment 3. Again, the proof is in the textbook; I am asking you to go over it in detail on your own.

Rao-Blackwell Theorem (textbook, pg 310)

The Rao-Blackwell theorem says that if we choose to base our estimator off of something other than a sufficient statistic, then we can always find another estimator with the same mean and lower variance.

It also gives a recipe for *finding* such an estimator, which helps with our intuition. In practice though, the Rao-Blackwell theorem is not often used to find estimators, because

1. Computing $E(\hat{\theta}|T)$ is usually harder than just going back and finding a better estimator from scratch
2. We don't usually manage to mess up and find an estimator based on a sufficient statistic that doesn't already have minimum variance

Rao-Blackwell Theorem (textbook, pg 310)

Where it *is* used is in a more detailed discussion of sufficiency. There is a distinction between statistics that provide good estimators of parameters, and statistics that provide "minimal" or "optimal" summaries of data.

For example, consider $\sum_{i=1}^{n} X_i$ vs $\bar{X}$. The former is slightly simpler to compute (don't divide by n), so it provides a more basic summary of the data while still being sufficient.

The latter is definitely a better estimate of $\mu = E(X)$. In particular, it has much lower variance.

Remember: our goal in this course is focussed around parameter estimation.

Example

Recall the example where $X_1, X_2 \sim N(0,1)$ independently, and I told you that

$$X_2|\bar{X} = t \sim N(t, 1/2)$$

Then with $\hat{\theta} = X_2$, say, we get $\tilde{\theta} = E(X_2|\bar{X}) = \bar{X}$, and we recover our usual estimator fo μ.

Towards finding estimators

We've talked about how to tell whether an estimator is sufficient, and about how to improve one estimator if we know about a sufficient statistic. Are these tools enough to actually find good estimators?

We have also talked about factoring the joint density of the sample given the parameters, and analyzing how the result depends on the parameters.

Let's take this idea one step further.

Joint Distribution of the Sample

So far we have talked about how we should use all the data we have when making inferences about θ. There is a converse to this statement as well.

One of the main principles behind the "Frequentist" approach to parameter estimation that we adopt in this course is that inferences about a parameter should be based *only* on the observed data.

So we should use all the data, and nothing but.

Of course, this isn't really true, since we make a *ton* of assumptions about the family of distributions from which the data came- but as I said in this course, we are going to take the family as fixed and known, and talk only of parameter estimation.

Joint Distribution of the Sample

How can we achieve this goal?

Consider the joint distribution of the sample, $f(\mathbf{x}|\theta)$. For a fixed, observed dataset $\mathbf{x}$ generated from this distribution, what is the simplest sufficient statistic we can think of, given the above philosophy of using only the observed data to estimate θ?

The Data is Sufficient

Really what we are saying is simply that the data $\mathbf{x}$ itself is sufficient for θ.

Trivially, we can write

$$f(\mathbf{x}|\theta) = g(\mathbf{x}, \theta) \times h(\mathbf{x})$$

and just take $h(\mathbf{x}) = 1$.

What this means is that we are justified in using only $f(\mathbf{x}|\theta)$ to estimate θ.

Likelihood

But I told you before that since we have already *observed* the dataset, there is no randomness left in $\mathbf{x}$. We're conditioning all inference based on the observed data only. So what is the point of looking at its distribution?

Definition: The **likelihood** function is the joint distribution of the data, treated as a function of the parameters:

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

It's exactly the same formula, just treated as a function of θ for fixed $\mathbf{x}$ rather than the other way around.

It's one of the most important definitions in all of Statistics.

Example

Suppose $X_i \sim Bern(p)$ is a random sample of coin flips. Find the likelihood function for p.

Example

Solution: the likelihood function is equal to the probability of observing any particular sequence of 0's and 1's. Since each trial is independent, the probability of observing a sequence is just equal to the product of the probabilities of observing each result:

$$\begin{aligned} L(p|\mathbf{x}) =& p \times p \times p \times \ldots \times (1-p) \times (1-p) \ldots \times (1-p) \\ =& p^{\sum_{i=1}^{n} x_i} \times (1-p)^{\sum_{i=1}^{n} (1-x_i)} \end{aligned}$$

We multiply by p for each 1 in the sequence, and by (1-p) for each 0 in the sequence. There are $\sum_{i=1}^{n} x_i$ 1's, and $\sum_{i=1}^{n}(1-x_i) = n - \sum_{i=1}^{n} x_i$ 0's.

Note any sequence with the same number of 0's and 1's in it gets the same likelihood for p- the order doesn't matter (can you connect this to the concept of sufficiency?).

Example: IID Data

The case where we have IID (Independent, Identically Distributed) data is the most common, and gets special attention.

Suppose $X_i \sim F_\theta$ independently. Denote the corresponding density of each x_i as $f_x(x|\theta)$ Then

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f_{x_i}(x_i|\theta)$$

Likelihood

Treating this as a function of θ, there is a lot we can do with $L(\theta)$ (I'm dropping the dependence on $\mathbf{x}$ in my notation).

Consider this: values of θ that give a higher $L(\theta)$ are more likely to have generated the observed data. (Why?)

Maximum Likelihood

This idea gives us a way to find estimators.

Definition: for an observed sample $\mathbf{x}$ with joint density $f(\mathbf{x}|\theta) = L(\theta)$, the **maximum likelihood estimator** of θ is the value of θ that maximizes $L(\theta)$.

This gives us *the value of the parameter that was most likely to have generated the data we observed*.

Practically, we just turned our estimation problem into an optimization problem.

Log-likelihood

Now that we've introduced the likelihood $L(\theta)$, let's not actually use it. Instead let's use the *log-likelihood*

$$\ell(\theta) = \log L(\theta)$$

(base e log, which some may know as ln).

This gives the same MLE (Maximum Likelihood Estimators) as working with $L(\theta)$ because log is a monotone function, so in general $f(x)$ and $\log f(x)$ have the same optima.

Why?

1. The likelihood of an independent sample is a product over the density of each point. The log-likelihood is a sum.
2. Densities themselves are often a product of several factors anyways, and a log of these gives a sum. Sums are way easier to work with than products.
3. The log-likelihood has a bunch of awesome theoretical properties that we will discuss in the coming weeks
4. Numerically more stable. Likelihood is a product of density values, most of which are small, so can get really small really fast. Log likelihood is a sum of logged values, which gets sort of small much less fast.

Example

Simple example: $X_i \sim N(\mu, 1)$. The likelihood is

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

The log likelihood, on the other hand, is

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

Example

To optimize this, take two derivatives with respect to μ. Set the first to zero and solve; check that the second one is negative $\Rightarrow$ local maximum.

$$\partial\ell/\partial\mu = \sum_{i=1}^{n}(x_i - \mu) = 0 \implies \hat{\mu} = \bar{x}$$

$$\partial^2\ell/\partial\mu^2 = -n < 0 \forall \mu \implies \hat{\mu} = \bar{x} \text{ maximum of } \ell(\mu)$$

You will work out the details for yourself on Assignment 3, and re-do when the standard deviation is also a parameter to be estimated.

Multivariable Optimization

I'm expecting you're familiar with basic univariate optimization from calculus like we just did. If you're not though, the previous example contains all you need to know.

I'm not expecting you're all familiar with multivariable optimization, so let's briefly discuss the procedure we'll use in this course.

Multivariable Optimization

To optimize a multivariable function $f : \mathbb{R}^d \to \mathbb{R}$:

1. Take derivatives with respect to each variable, holding all others fixed (these are called *partial* derivatives, and is what the ∂ symbols refer to), and set each of these derivatives equal to zero.
2. This gives you a system of equations. Solve the system to get estimators for each parameter.
3. The part we're going to skip is the multivariable generalization of the second derivative test: check whether the negative of the *Hessian* (matrix of second-order partial derivatives) is *positive definite*.

So in this course, just find the estimators using step 1 and 2, because I don't want to spend any more time on this, even though it's super important.

Example

We can revisit the normal example when both parameters are unknown. The log-likelihood function is now

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

with derivatives

$$\partial\ell/\partial\mu = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \implies \hat{\mu} = \bar{x}$$

$$\partial\ell/\partial\sigma^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Notes

Notice I did two tricky things:

- Assumed that $\sigma > 0$, which I did twice (where?). You should always first recognize what your *parameter space* - the space of all possible values of θ - is, and only optimize within it.
- Maximized with respect to σ^2, rather than σ. This works because the MLE any one-to-one function of the parameter, $\psi = g(\theta)$, is $\hat{\psi} = g(\hat{\theta})$. This is not a trivial fact, and it is extremely useful.

Example

Find a MLE for the *precision* in the previous example, which is a fancy word sometimes used to refer to the inverse of the variance. That is, find the MLE for $\psi = 1/\sigma^2$.

Example

The log-likelihood can be re-parametrized in terms of ψ:

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

with derivatives

$$\partial\ell/\partial\mu = \psi\sum_{i=1}^{n}(x_i - \mu) = 0 \implies \hat{\mu} = \bar{x}$$

$$\partial\ell/\partial\psi = \frac{n}{2\psi} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \implies \hat{\psi} = \frac{n}{\sum_{i=1}^{n}(x_i - \mu)^2}$$

Example

..or, we could have used the fact that since the function $g(x) = 1/x$ is one-to-one ($x \neq 0$), the MLE for $\psi = 1/\sigma^2$ is

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

In MLE problems, reparametrizing the distribution to make it easier to differentiate is a common useful technique.

You can't always use calculus

You can't maximize the likelihood using calculus techniques in three major common cases:

- The likelihood is not continuous on the whole parameter space, so you can't take derivatives
- The parameter space is not an *open* subset of $\mathbb{R}^d$. This occurs when it includes its boundary. For example, what if we had allowed $\sigma = 0$ in the previous example?
- The support of the distribution depends on the parameters. This one is more subtle, so let's take a look at an example

Example

Let $X_i \sim Unif(0, b)$, the continuous uniform distribution on the open interval $(0, b)$. Find the MLE of b.

The likelihood function is

$$L(b) = \prod_{i=1}^{n} \frac{1}{b}$$

which is a strictly decreasing function of b. In particular, it is unbounded as $b \to 0$.

Or is it?

Example

We made a mistake: we didn't express all of the dependency on b explicitly in $L(b)$. We actually should have written

$$L(b) = \prod_{i=1}^{n} \frac{1}{b} \times I(x_i \leq b)$$

I'll leave it to you to

- Convince yourself that this is true and
- Show that the MLE is $\hat{b} = max(x_i)$

Connection to Sufficiency

The MLE is sufficient. Why?

Exercise (Assignment 3): show that the MLE can depend on the data only through the value of a sufficient statistic.

Because the MLE is a function of a sufficient statistic, it is itself sufficient.

Is it consistent?