Chapter 12

Mathematical Statistics

Contrast (among the parts of this course)

Math Stat: The methods discussed in this chapter all assume you KNOW the mathematical form of the distribution of your data, with some unknown parameters.

Stat: using bootstrapping and randomization tests: In the statistics with simulation part of the course, we did not make assumptions about the mathematical form of the distribution of the data.

Traditional Applied Stat methods: In the part of this course where we discuss results from typical applied statistics courses, the methods are using a certain theoretical model so that you know the mathematical form of the distribution of your data, but there are many "footnotes" and conditions about this, because many of the techniques are robust against certain deviations from those usual assumptions and so they are used more widely than the list of assumptions indicates.

Estimation

How do we derive an estimator of a parameter?

What properties do we want a good estimator to have?

Methods to Derive an Estimator

- Maximum Likelihood Estimator (MLE)

- Method of Moments Estimator
- Least Squares Estimator
- Bayesian Estimator

Desirable Properties of an Estimator

- Unbiased
  (The expected value of the estimator is the parameter being estimated)
- Small Variance

You're only going to use one of them to make your estimates. So you'd like to know that it was part of a cloud that was altogether.

Properties of MLE

- "Asymptotically" Unbiased
(As we take larger and larger samples, any bias there is in the MLE gets smaller and smaller.)
- Asymptotically, the variance gets smaller and smaller.
- Asymptotically, the dist'n of the MLE goes to a normal dist'n
- The MLE of a function of a parameter is simply that same function applied to the MLE of the parameter. (Invariance property.)

Other estimators

- Estimators derived by other methods are useful in various situations.
- The first two properties in the list for the MLE, together, characterize a "consistent" estimator. It would be unusual to use an estimator that is not consistent.
- However, one might easily choose to use an estimator that doesn't meet the last two conditions listed.
- Generally speaking, if we are less certain of the theoretical dist'n of our data, we might be more interested in a "robust" estimator than one that achieves these other conditions, which are based on our assumptions about the distribution, particularly when the sample sizes aren't large.

If an estimator itself is a biased estimator, if it meets the conditions that it gets large, it gets closer and closer to unbiased, and the variance closer and closer, it's called consistent.

Finding a MLE

- What is the Likelihood Function?
- What is the meaning of it?
- Why is maximizing it a good idea?
- Is there one method that will always work to maximize the Likelihood function?

Example 1 for MLE

Suppose that we want to estimate the recidivism rate for juvenile offenders in a large city.

We randomly sample 25 juveniles from the juveniles who were released from jail and determine whether each did or did not return to jail within a two-year period.

The model

We do this and, in our random sample of 25,
12 had actually returned to jail.

What is our estimator of *p*, the population proportion who return to jail within two years?

We think of this as 25 instances of a Bernoulli r.v.
each of which has probability *p* of success.
So the probability of getting the sample we obtained is

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$$
$$= p \cdot p \cdot \ldots \cdot p \cdot (1-p) \cdot \ldots \cdot (1-p)$$
$$= p^{12}(1-p)^{13}$$

Strategy to Maximize this probability

What strategy is useful here?

We might say that we want
the answer for the value of p
that makes this joint probability
the largest it can be,
given the data we have.

Other strategies are possible,
but this is the strategy that leads to what we call the "maximum likelihood estimator."

Let's do some calculus

Let's call this function that
we want to maximize L.

$$L(p) = p^{12}(1-p)^{13} \quad \text{for} \quad 0 \le p \le 1$$

Stop the video now and think about how to
do this as a calculus problem.

- How will you find the critical point(s)?
- How will you decide if each is a
  local maximum or local minimum or neither?
- This is maximization of a continuous
  function on a closed interval.
  How do you find the absolute maximum?

Did you try it?

- This calculus problem wasn't too messy.
- But it is true that taking the derivative was somewhat tedious and taking the second derivative even more tedious.
- So maybe you didn't get further than noticing that.
- If so, that's OK.

What's a shortcut here?

- Let's think about functions.
- If I were to think of a monotone increasing function to apply to $L(p)$, and then find the value(s) of p that maximize this new function,
- they will be the same value(s) that maximize the original function!

Is there a good choice for such a function?

- Did you go far enough into the calculus to notice that the difficulty here (if you found it difficult!) was that the derivative of this product is “messier” than the original function?
- But if we had a sum instead of a product, it would not get “messier.”
- What monotone function do we know that would turn products into sums?

Yes! There is such a function!

- If we take the natural logarithm of our pdf here, then that makes the problem much easier.

Here's the general notation

We need to change from our pdf (a function of the data given the parameter) to something we identify as a function of the parameter given the data. The name for that in statistics is the Likelihood Function, *L*

$$L(p \mid X_1, X_2, \ldots, X_n) = f(X_1, X_2, \ldots, X_n \mid p)$$

$$L(p) = f(X_1, X_2, \ldots, X_n \mid p)$$

Further, we just said that we want to take the log of it, so

Standard notation is "lower-case letter l" for that.

$$l(p) = \log(L(p))$$

Did you find the maximum?

Turn off your video and do some work here.

Using this, find the critical point(s) and then find a critical point that gives an absolute maximum on the appropriate interval.

Here are some hints:

1. If you're struggling, do it first with the numerical values 25 and 12 and 13.
2. Then go back and do it again with symbols for some or all of those given numbers.

Solve for the critical value

$$\text{Let } S = \sum_{i=1}^{n} X$$

$$L(p) = p^{S}(1-p)^{(n-S)} \qquad 0 \le p \le 1$$

$$l(p) = \log(L(p)) = S \cdot \log p + (n-S)\log(1-p)$$

$$\frac{d}{dp} l(p) = \frac{S}{p} + (n-S)\frac{1}{1-p}(-1) = 0$$

$$S(1-p) - (n-S)p = 0$$

$$S - Sp - np + Sp = 0$$

$$np = S$$

$$p = \frac{S}{n}$$

$$\hat{p} = \frac{S}{n}$$

Notation

Notice that, as we went through all the algebra and calculus, we were using the symbol p for the parameter of the distribution. And, to find the critical point, we solved for p.

It is inappropriate to say (or think) that this statistic IS the value of the parameter. We are deriving an estimator for the parameter, which is different from the parameter itself. The standard notation for an estimator is to use the symbol for the parameter and put a "hat" over it to make a new symbol $\hat{p}$.

That's what we see in the last line.

Other symbols besides "hat" such as a "tilde", $\tilde{p}$, can be used when we are looking at two different estimators in order to distinguish them from each other.

Example 1 for MLE: Conclusion

- $\hat{p} = \frac{S}{n}$ is the only critical point on a function that is continuous on the interval.
- Check the value of the function at the endpoints and this critical point to see which gives the maximum.

$$L(\hat{p}) = L\left(\frac{S}{n}\right) > 0 \quad \text{and} \quad L(0) = L(1) = 0\,.$$

- Thus our critical value is gives the maximum of the likelihood function and is our "maximum likelihood estimator of p."

Note that it was easier to revert back to the Likelihood Function to check for the maximum, since the log likelihood function here is not defined on the endpoints of the domain of the parameter.

Example 2 for MLE

- Suppose our data have a distribution for which the domain of the random variable depends on the parameter.

- Of course, we want to use the data to estimate the parameter.

Example 2 for MLE

- Suppose our data have a distribution for which the domain of the random variable is defined by one of the parameters.
- We want to use the data to estimate the parameter.
- In this situation, our usual calculus methods (with or without using the log likelihood function) are not useful. We revert to simply graphing the Likelihood Function. The shape will give us information about the maximum.

X_{max}

MLE in a Uniform Dist'n – writing the function

We have *n* independent variables:

$$X_1, X_2, \ldots, X_n \sim \text{Uniform}(0, \theta) \quad \text{for } \theta > 0$$

So the likelihood function is

$$L(\theta) = f(X_1, X_2, \ldots, X_n \mid \theta)$$

$$= \left(\frac{1}{\theta}\right)^n \quad \text{for each } i,\ 0 < X_i < \theta$$

$$= \left(\frac{1}{\theta}\right)^n \quad \text{for } 0 < X_{\min} \text{ and } X_{\max} < \theta$$

MLE in a Uniform Dist'n – Likelihood function

More thoroughly,

$$L(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{for } 0 < X_{\min} \text{ and } X_{\max} < \theta \\ 0 & \text{elsewhere} \end{cases}$$

Let's graph this.

Put θ on the horizontal axis and $L(\theta)$ on the vertical axis.

MLE in a Uniform Dist'n – making graph

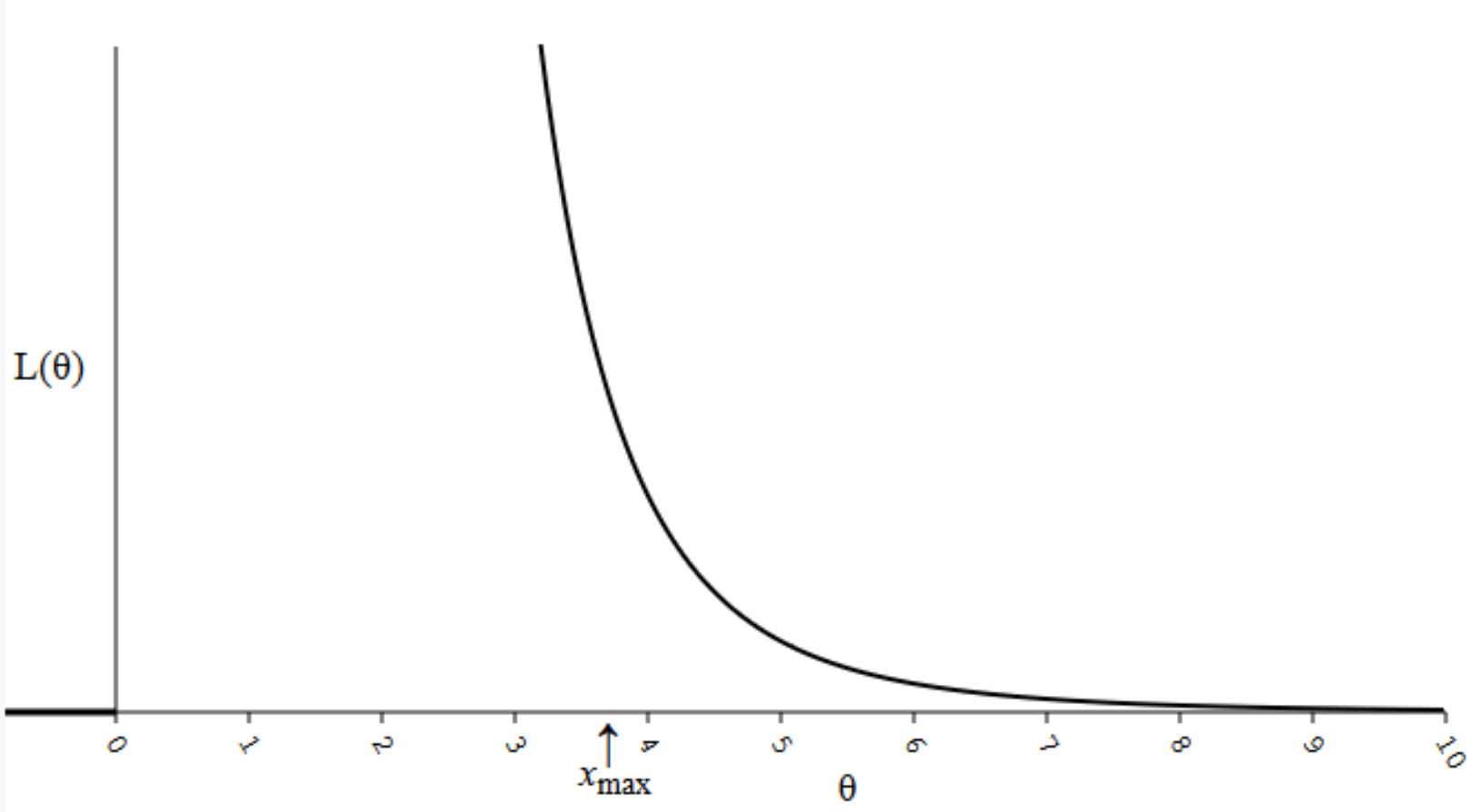
To obtain numbers to plot points,
choose a value for *n*.

But I notice that, for $n \geq 1$
the shape of a rough sketch is always the same

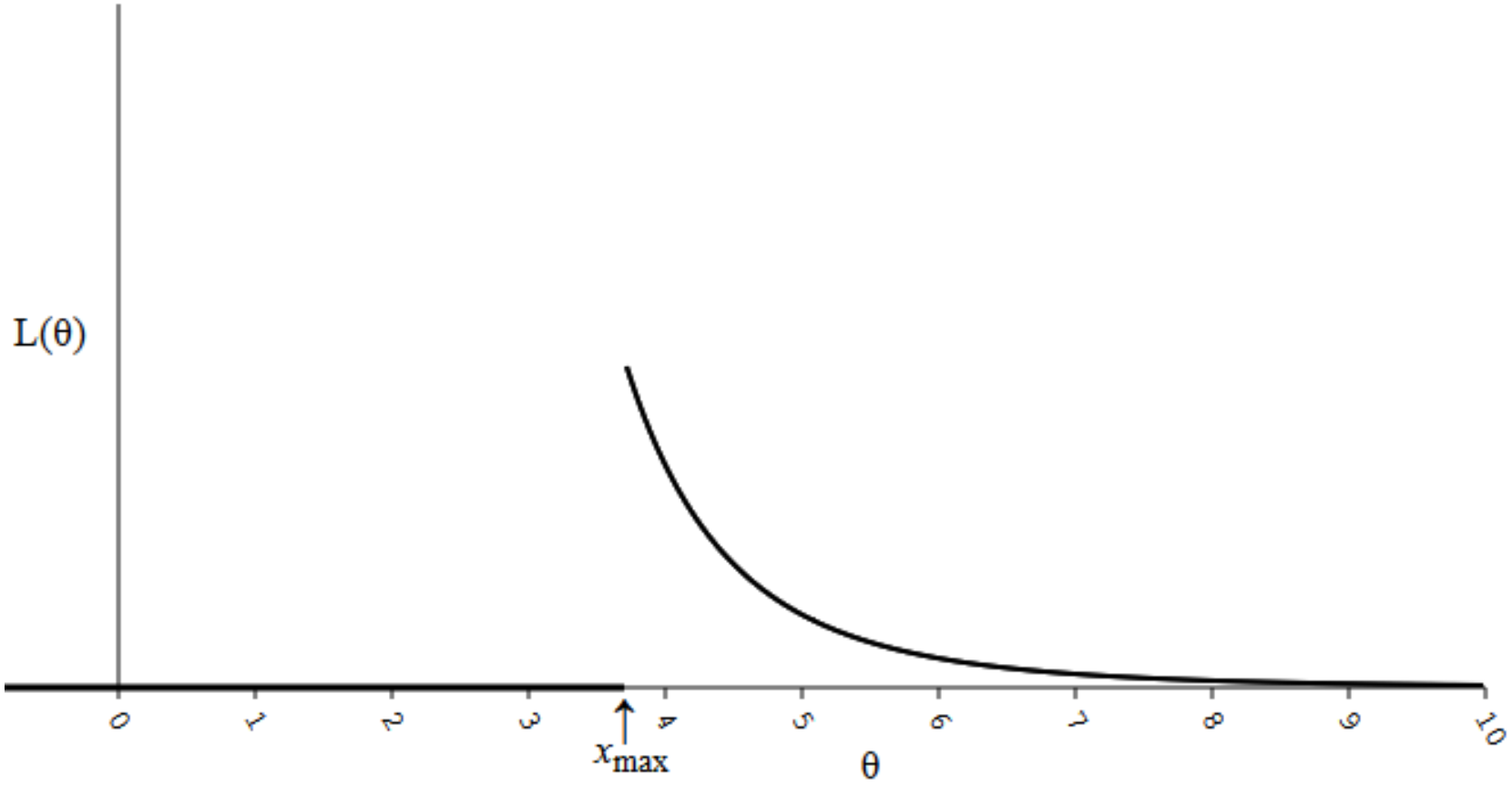
The next slide shows that shape but does not take
into account the restriction on θ.

The following slide is our complete graph of $L(\theta)$, showing
that, because $\theta > X_{\max}$, then the non-zero part of $L(\theta)$
"starts" at the numerical point $\left(X_{\max}, L(X_{\max})\right)$

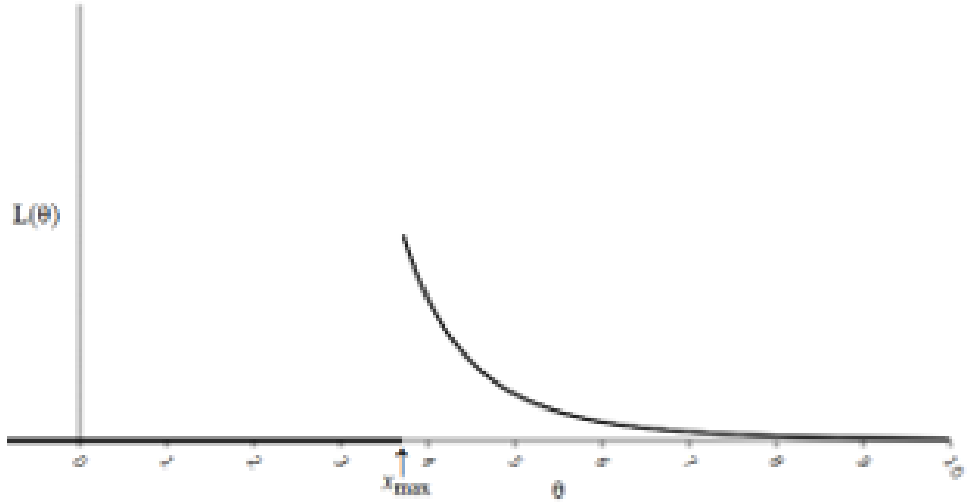
MLE in a Uniform Dist'n – first step of graph



MLE in a Uniform Dist'n: final graph



MLE in a Uniform Dist'n - conclusion



It is clear that the maximum value of $L(\theta)$ is achieved at the numerical value of $\theta = X_{\max}$

That tells us that the maximum likelihood estimator (MLE)of this parameter θ is $X_{\max}$

Using the standard notation for an estimator

$$\hat{\theta} = X_{\max}$$

Example 2 for MLE

- In this situation, our usual calculus methods
(with or without using the log likelihood function)
are not useful
because the function isn't differentiable as needed.
- We revert to simply graphing the Likelihood Function. The shape will give us information about the maximum.

X_{max}

MLE in a Uniform Dist'n – writing the function

We have independent variables:

$$X_1, X_2, \ldots, X_n \sim \text{Uniform}(0, \theta) \quad \text{for } \theta > 0$$

So the likelihood function is

$$L(\theta) = f(X_1, X_2, \ldots, X_n \mid \theta)$$

$$= \left(\frac{1}{\theta}\right)^n \quad \text{for each } i,\ 0 < X_i < \theta$$

$$= \left(\frac{1}{\theta}\right)^n \quad \text{for } 0 < X_{\min} \text{ and } X_{\max} < \theta$$

MLE in a Uniform Dist'n – Likelihood function

More thoroughly,

$$L(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{for } 0 < X_{\min} \text{ and } X_{\max} < \theta \\ 0 & \text{elsewhere} \end{cases}$$

Let's graph this.

Put θ on the horizontal axis and $L(\theta)$ on the vertical axis.

MLE in a Uniform Dist'n – making graph

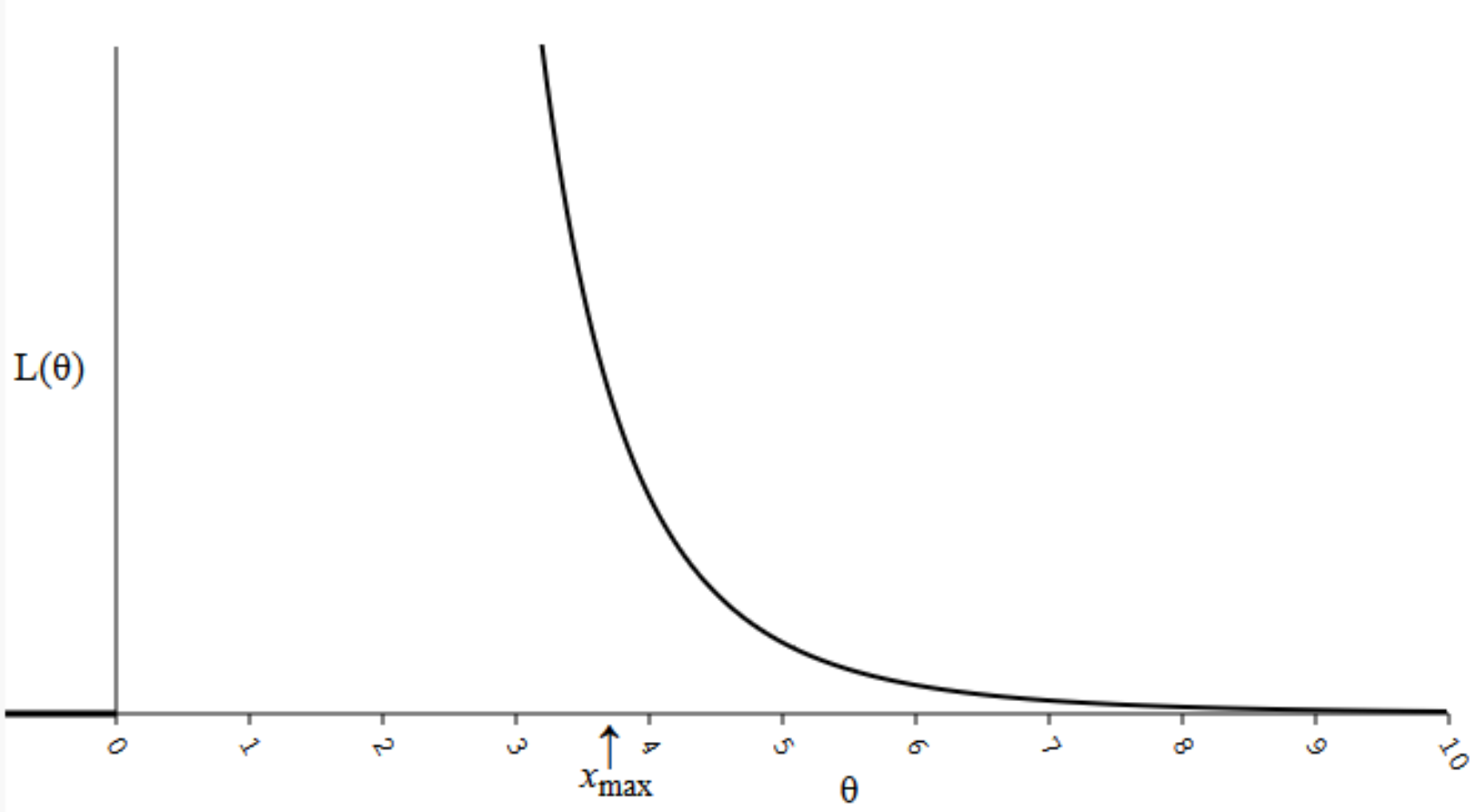
To obtain numbers to plot points,
choose a value for *n*.

But I notice that, for $n \geq 1$
the shape of a rough sketch is always the same

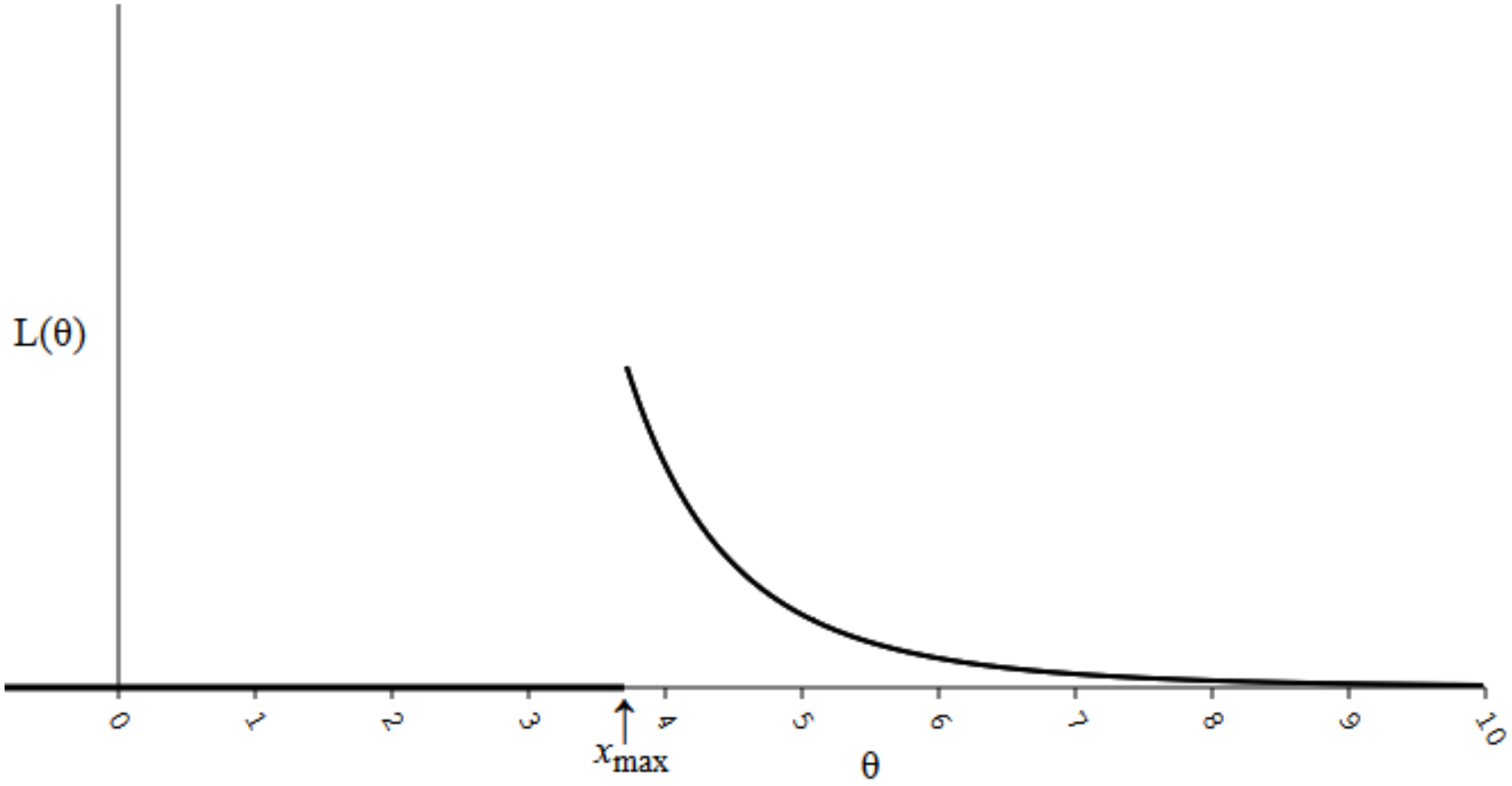
The next slide shows that shape but does not take into account the restriction on θ.

The following slide is our complete graph of $L(\theta)$, showing that, because $\theta > X_{\max}$, then the non-zero part of $L(\theta)$ "starts" at the numerical point $\left(X_{\max}, L(X_{\max})\right)$

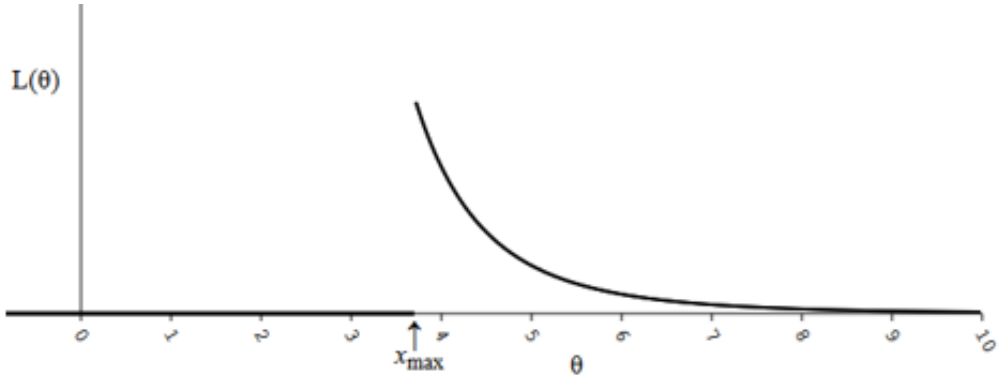
MLE in a Uniform Dist'n – first step of graph



MLE in a Uniform Dist'n: final graph



MLE in a Uniform Dist'n - conclusion



It is clear that the maximum value of $L(\theta)$ is achieved at the numerical value of $\theta = X_{\max}$

That tells us that the maximum likelihood estimator (MLE) of this parameter θ is $X_{\max}$

Using the standard notation for an estimator

$$\hat{\theta} = X_{\max}$$

Transition to Comparing Estimators

- We have only discussed one method of deriving estimators here, although some other methods were mentioned.
- Of course, people can just “make up” plausible estimators as well.
- Thus, it is important to mention methods of comparing estimators.

Comparing Estimators

- Earlier we mentioned that the most prominent attributes we want in an estimator are:
  - Unbiased (or close to that)
  - Small Variance
- Methods to compare estimators
  - Mean Squared Error
  - Relative Efficiency

Comparing Estimators 2

- To think about these, we need to talk more about Unbiased estimators and the Bias of estimators.
- You already know about computing the variance of estimators. (Estimators are also known as statistics or random variables.)
- Two methods
  - Mean Squared Error.
    Takes more explanation and discussion.
  - Relative Efficiency: Ratio of the variances.
    Simple to compute. We'll discuss interpretation of it.

Definition of Bias

Suppose W is a statistic made up from the sample values from a common distribution with a parameter θ.

$W(X_1, X_2, \ldots, X_n)$

We can compute the expected value of W: $E(W)$

We're quite interested in: $\text{Bias}(W) = E(W) - \theta$

Sidelight: Sometimes W is a function of multiple parameters. Or we just want to emphasize the parameter of interest. So you'll see this notation instead:

$E_\theta(W)$ and $\text{Bias}_\theta(W) = E_\theta(W) - \theta$

Mean Squared Error

We are also interested in how close the values of W come to the parameter θ as we evaluate W for various sets of sample values.

We might measure this by the average squared distance between the value of the estimator and the parameter value. That is called the Mean Squared Error (MSE.)

$$\text{MSE} = E(W-\theta)^2$$

It can be shown that

$$\text{MSE} = E(W-\theta)^2 = \text{Var}(W) + \left(\text{Bias}(W)\right)^2$$

Unbiased Estimators and MSE

An unbiased estimator (quite desirable!) is an estimator whose Bias is 0.

If W is an unbiased estimator, then its mean squared error is the same as its variance.

Do we ever use a biased estimator?

Yes.

- For example, maximum likelihood estimators are often (slightly) biased.
- But, at the beginning of this chapter, we learned some good reasons to use maximum likelihood estimators.

Example: A Biased MLE

- We might choose to start with a MLE and try to modify it make a slightly different, unbiased, estimator.
- We will illustrate that in the next few minutes.
- First, we will think about how we already modified an estimator in Chapter A of the statistics material. (We changed from estimating a population proportion using the sum of the data to estimating it using the p-hat - by dividing by 25.)
- Then we will see how to apply the same general idea to the MLE of theta from Example 2 above, and find an unbiased estimator of theta, based on that MLE.

Modifying the estimator

How can we “fix” a biased estimator?

Answer: Sometimes, if we compute the expected value of an estimator, the result is a simple function of the parameter, and so it suggests a simple modification of it in order to obtain an unbiased estimator of the parameter.

Previous example:
Recall the example in the first chapter,
where we estimated a population proportion *p* from
25 observations. (See the discussion on the next page.)

Modifying the Estimator from Sec. A.1

We had 25 independent $X_i \sim \text{Bernoulli}(p)$ and used a summary statistic $S = \sum_{i=1}^{25} X_i$ where $S \sim \text{Bin}(25, p)$. We noted that $E(S) = 25p$ and we wanted to estimate p so we decided to divide the summary statistic S by 25 to obtain the statistic

$$\hat{p} = \frac{S}{25}$$ so that

$$E(\hat{p}) = E\left(\frac{S}{25}\right)$$

$$= \left(\frac{25p}{25}\right)$$

$$= p$$

and we now know to call this an "unbiased estimator" of the population proportion p.

Modifying the Estimator: MLE for Uniform 1

- **Preliminary work for the MLE Uniform example**
  (Skip this now and come back to it if you wish)

- It can be shown that, for any continuous r.v.'s which are independent, identically distributed with cdf $F_X(x)$ and pdf $f_X(x)$, then the sampling dist'n of $W = X_{(n)}$ is

$$f_W(w) = n \cdot \left(F_X(w)\right)^{n-1} f_X(w).$$

- When you have the PDF of W, then you can find the expected value of W.

$$E(W) = \int_0^{\theta} w \cdot f_W(w)\, dw$$

That gets you to where the example for the MLE of θ of a $\text{Uniform}(0, \theta)$ begins.

Modifying the Estimator: MLE for Uniform 2

For independent random variables

$X_1, X_2, \ldots, X_n \sim \text{Uniform}(0, \theta) \quad \text{for } \theta > 0$

We have shown that the MLE of θ is

$W = X_{\max} = X_{(n)}$ ("order statistics" notation)

Find a statistic that is a function of this MLE and is an unbiased estimator of θ.

Assume, from the work described on the previous slide,

$$E_\theta(W) = \frac{n}{n+1}\theta$$

Stop the video and think about this.

- What will you do now to find an unbiased estimator of θ?
- Is it straightforward?
  Don't come back to the video until you are clear on whether you know the answer or are feeling confused.

Modifying the Estimator: MLE for Uniform 3

Answer to Modifying an Estimator: MLE for Uniform

If we define an estimator as $V = \frac{n+1}{n} W$ then

$$E_\theta(V) = \frac{n+1}{n} E_\theta(W)$$

$$= \left(\frac{n+1}{n}\right)\left(\frac{n}{n+1}\theta\right)$$

$$= \theta$$

Thus V is an unbiased estimator of θ.

“Unbiasing” the estimator

- The idea of “unbiasing” an estimator is very useful.
- To do this, we have to know the distribution of the estimator
  also known as the sampling dist’n of that statistic which is the estimator.
- For many potential interesting statistics, it is difficult to find the exact (theoretical) sampling distribution.
- It is noteworthy that you can do it as simply as this for order statistics of continuous random variables. (Similar formulas can be derived for other order statistics.)

Comparing estimators - MSE

MSE: Generally speaking, the estimator with smaller MSE is better.

What's tricky?

- The MSE is often a function of the unknown parameter. Thus, which estimator has smaller MSE may differ depending on where the actual parameter is.
- This can be OK if you have some prior reason to believe the population parameter value is in a particular region of the parameter space. It is not unusual for that to be the case in practical problems.

Comparing Estimators: Relative Efficiency

The Relative Efficiency of two estimators is defined as the ratio of the variances.

- Since this doesn't take into account any bias, it is most appropriate to use if both estimators are unbiased, or, at least, both estimators have the same, or approximately the same, bias.
- Generally speaking, we will form the ratio in such a way that the result is between 0 and 1, so that it is fairly easy to interpret. That means the "better" estimator (smaller variance) is in the numerator.

Relative Efficiency: Example

Our data (sample size n) are independent identically distributed from $\text{Normal}\left(\mu, \sigma^2\right)$. Let $W = \bar{X}_n$ and $V =$ sample median.

The relative efficiency of the median relative to the mean is $RE(V, W) = \dfrac{\text{var}(W)}{\text{var}(V)}$.

If we're estimating the mean, we might not know the variance either. We also don't have easily available the variance of the sample median, but one can find various references to that computation, which says that, **if *n* is large**, $\text{var}(V) = (1.253)^2 \dfrac{\sigma^2}{n}$ and, of course, $\text{var}(W) = \dfrac{\sigma^2}{n}$

so $RE(V, W) = \dfrac{\text{var}(W)}{\text{var}(V)} = \dfrac{\sigma^2/n}{(1.253)^2(\sigma^2/n)} = 0.637$

Thus, if the dist'n of the data is normally distributed, and if the sample size is large, then using the median of the data to estimate the center is about 64% as effective as using the mean to estimate the center.

Relative Efficiency of ONE estimator

- What could this mean?
- It is called the “Asymptotic Relative Efficiency (ARE.)”
- The one estimator should be unbiased.
- Then it is compared to a value which is the lower bound on the possible variance of an unbiased estimator from this particular distribution. (Comes from the Cramer-Rao Inequality.)

Mean Squared Error

We are also interested in how close the values of W come to the parameter θ as we evaluate W for various sets of sample values.

We might measure this by the average squared distance between the value of the estimator and the parameter value. That is called the Mean Squared Error (MSE.)

$$\text{MSE} = E(W-\theta)^2$$

It can be shown that

$$\text{MSE} = E(W-\theta)^2 = \text{Var}(W) + \left(\text{Bias}(W)\right)^2$$

Mean Squared Error

$$\mathrm{MSE} = E(W - \theta)^2$$

It can be shown that

$$\mathrm{MSE} = E(W - \theta)^2 = \mathrm{Var}(W) + \left(\mathrm{Bias}(W)\right)^2$$

Compare two estimators using MSE

Consider independent random variables

$$X_1, X_2, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$$

$$\text{for } -\infty < x < \infty \text{ and } -\infty < \mu < \infty \text{ and } \sigma > 0$$

It can be shown that the MLE's of the two parameters are

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

It can also be shown that this MLE of the variance is a biased estimator.

$E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$. Thus we often define a new estimator,

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \text{ because } E(S^2) = \sigma^2$$

MSE of the two estimators

We want to find the MSE of each of these two estimators.
From a result involving a chi-squared dist'n, we can show

$\mathrm{Var}\left(S^2\right) = \dfrac{2\sigma^4}{n-1}$ which can be used to show $\mathrm{Var}\left(\hat{\sigma}^2\right) = \dfrac{(2n-1)\sigma^4}{n^2}$

$$\mathrm{MSE}\left(S^2\right) = \mathrm{Var}\left(S^2\right) + \mathrm{Bias}\left(S^2\right)$$

$$= \frac{2\sigma^4}{n-1} \quad + \quad 0$$

$$= \frac{2\sigma^4}{n-1}$$

$$\mathrm{MSE}\left(\hat{\sigma}^2\right) = \mathrm{Var}\left(\hat{\sigma}^2\right) \; + \; \mathrm{Bias}\left(\hat{\sigma}^2\right)$$

$$= \frac{2(n-1)\sigma^4}{n^2} \; + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2$$

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

Actual comparison

$$\text{difference} = \frac{2\sigma^4}{n-1} - \frac{(2n-1)\sigma^4}{n^2}$$

$$= \left(\frac{2}{n-1} - \frac{(2n-1)}{n^2} \right) \sigma^4$$

$$= \left(\frac{2}{n-1} - \frac{\left(2 - \frac{1}{n}\right)}{n} \right) \sigma^4$$

Notice here that the second fraction compared to the first fraction has smaller numerator and larger denominator.

Thus, the second fraction is smaller than the first fraction. The smaller is the MSE of the MLE and NOT the unbiased estimator.

Conclusion of MSE example

This is a quite interesting result.

The biased estimator $\hat{\sigma}^2$ actually has a smaller MSE than the unbiased estimator S^2

Comparison of Estimators

- If we consider all possible estimators of a parameter, we don't have enough "structure" to obtain unique best estimators.
- If we restrict ourselves to only unbiased estimators, there are ways of finding unique best estimators in some families of distributions for some functions of the parameters.
- For the reasons mentioned at the beginning of this chapter, maximum likelihood estimators are very commonly used.

In what situation are we estimating?

- All of the theory mentioned here about estimation are strongly dependent on the distribution of the data (resulting in the distribution of the sample statistic.)

- In many situations, robustness of estimators against certain types of deviations from the assumptions is as important or more important than optimality results in the specific distribution.

Of what use to you are these ideas?

- In a situation you want to study, can you devise a plan to use simulation studies to compare estimators? (Think of our introduction “Tank Problem” discussion.)
- Have you learned something about how to use theory to investigate estimators when you have a good idea of the appropriate distribution of the data?