14.310x Lecture 12

Statistics---criteria for assessing estimators

Recall that an estimator is a random variable. So it has a distribution. Our criteria for assessing estimators will be based on characteristics of their distributions.

Statistics---criteria for assessing estimators

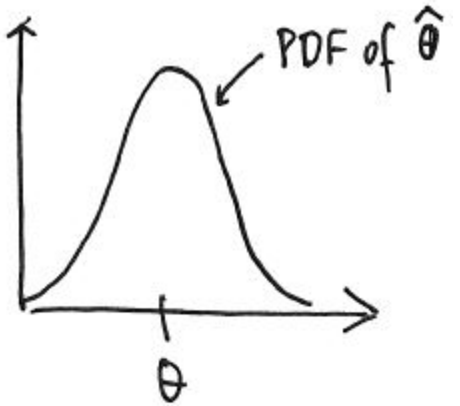
Recall that an estimator is a random variable. So it has a distribution. Our criteria for assessing estimators will be based on characteristics of their distributions.

An estimator is <u>unbiased</u> for θ if $E(\hat{\theta}) = \theta$ for all θ in Θ.

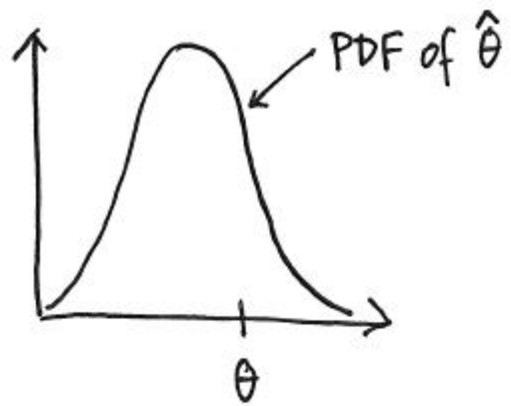
Statistics---criteria for assessing estimators

Recall that an estimator is a random variable. So it has a distribution. Our criteria for assessing estimators will be based on characteristics of their distributions.

An estimator is <u>unbiased</u> for θ if $E(\hat{\theta}) = \theta$ for all θ in Θ.



unbiased



biased

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$\hat{\theta}_2 = 2 \, \frac{1}{n} \sum_{i=1}^{n} X_i$$

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$\hat{\theta}_2 = 2 \, \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$E(\hat{\theta}_2) = 2 \, \frac{1}{n} \sum E(X_i)$$

$$= 2 \, \frac{1}{n} \sum \theta/2$$

$$= \theta$$

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$\hat{\theta}_2 = 2\,\frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E(\hat{\theta}_2) = 2\,\frac{1}{n}\sum E(X_i)$$

$$= 2\,\frac{1}{n}\sum \theta/2$$

$$= \theta$$

So unbiased for θ

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$\hat{\theta}_1 = \max\{x_1, \dots, x_n\}$$

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$\hat{\theta}_1 = \max\{X_1, \ldots, X_n\}$$

Can't use properties of E like we just did to calculate the expectation here---we'll do it directly.

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$\hat{\theta}_1 = \max\{X_1, \ldots, X_n\}$$

Can't use properties of E like we just did to calculate the expectation here---we'll do it directly. First we need the PDF:

$$f_{\hat{\theta}_1}(x) = n f_X(x) [F_X(x)]^{n-1}$$

$$= n \frac{1}{\theta} \left(\frac{x}{\theta}\right)^{n-1}$$

$$= n \, x^{n-1} / \theta$$

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$\hat{\theta}_1 = \max\{X_1, \dots, X_n\}$$

So, $E(\hat{\theta}_1) = \int_0^\theta x\, n\, x^{n-1}/\theta\, dx$

$$= \frac{n}{\theta^n} \int_0^\theta x^n\, dx$$

$$= \frac{n}{\theta^n} \frac{x^{n+1}}{n+1} \Big]_0^\theta$$

$$= \frac{n}{n+1}\theta$$

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$\hat{\theta}_1 = \max\{X_1, \ldots, X_n\}$$

So, $$E(\hat{\theta}_1) = \int_0^\theta x\, n\, x^{n-1}/\theta\, dx$$

$$= \frac{n}{\theta^n} \int_0^\theta x^n\, dx$$

$$= \frac{n}{\theta^n} \left. \frac{x^{n+1}}{n+1} \right]_0^\theta$$

$$= \frac{n}{n+1}\theta$$ So biased for θ

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$\hat{\theta}_1 = \max\{X_1,\dots,X_n\}$$

So, $E(\hat{\theta}_1) = \int_0^\theta x\, n\, x^{n-1}/\theta\, dx$

$$= \frac{n}{\theta^n}\int_0^\theta x^n\, dx$$

$$= \frac{n}{\theta^n}\, \frac{x^{n+1}}{n+1}\Big]_0^\theta$$

$$= \frac{n}{n+1}\theta$$

So biased for θ

Not so surprising, if you think about it. The estimator will always be <= θ, = with zero probability.

Statistics---criteria for assessing estimators

<u>Thm</u> The sample mean for an i.i.d. sample is unbiased for the population mean.

<u>Pf</u> Already did it when we calculated the expectation of the sample mean.

<u>Thm</u> The sample variance for an i.i.d. sample is unbiased for the population variance, where the sample variance is

$$S^2 = \frac{1}{n-1} \sum (X_i - \overline{X}_n)^2$$

Statistics---criteria for assessing estimators

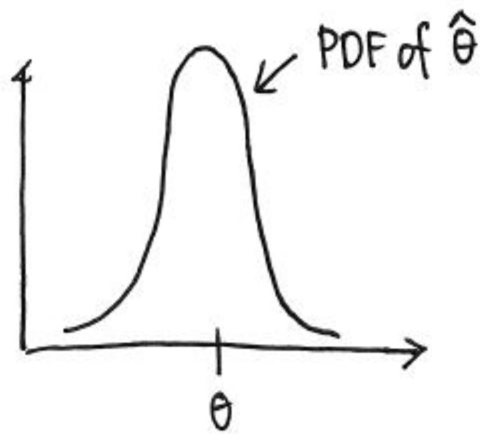
Given two unbiased estimators, $\hat{\theta}_1$ & $\hat{\theta}_2$, $\hat{\theta}_1$ is more <u>efficient</u> than $\hat{\theta}_2$ if, for a given sample size,

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$

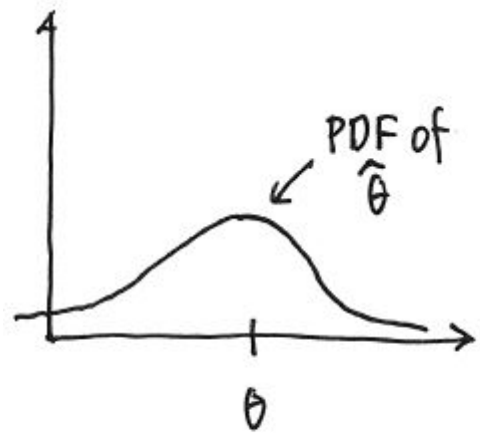
Statistics---criteria for assessing estimators

Given two unbiased estimators, $\hat{\theta}_1$ & $\hat{\theta}_2$, $\hat{\theta}_1$ is more <u>efficient</u> than $\hat{\theta}_2$ if, for a given sample size,

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$



efficient



inefficient

Statistics---criteria for assessing estimators

Given two unbiased estimators, $\hat{\theta}_1$ & $\hat{\theta}_2$, $\hat{\theta}_1$ is more <u>efficient</u> than $\hat{\theta}_2$ if, for a given sample size,

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$

Note that we have defined efficiency here just for unbiased estimators. The notion of efficiency can exist for broader classes of estimators as well, but we won't give a formal definition.

Statistics---criteria for assessing estimators

Sometimes we are interested in trading off bias and variance/efficiency. In other words, we might be willing to accept a little bit of bias in our estimator if we can have one that has a much lower variance. This is where mean squared error comes in.

$$MSE(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = Var(\hat{\theta}) + [E(\hat{\theta})-\theta]^2$$

Statistics---criteria for assessing estimators

Sometimes we are interested in trading off bias and variance/efficiency. In other words, we might be willing to accept a little bit of bias in our estimator if we can have one that has a much lower variance. This is where <u>mean squared error</u> comes in.

$$MSE(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = Var(\hat{\theta}) + [E(\hat{\theta})-\theta]^2$$

This is "bias" squared = 0 for unbiased estimators.

Statistics---criteria for assessing estimators

Sometimes we are interested in trading off bias and variance/efficiency. In other words, we might be willing to accept a little bit of bias in our estimator if we can have one that has a much lower variance. This is where <u>mean squared error</u> comes in.

$$MSE(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = Var(\hat{\theta}) + [E(\hat{\theta})-\theta]^2$$

This is "bias" squared = 0 for unbiased estimators.

Choosing a minimum mean squared error estimator is an explicit way to trade off bias and variance in an estimator. Not the only way, but a decent one.

Statistics---criteria for assessing estimators

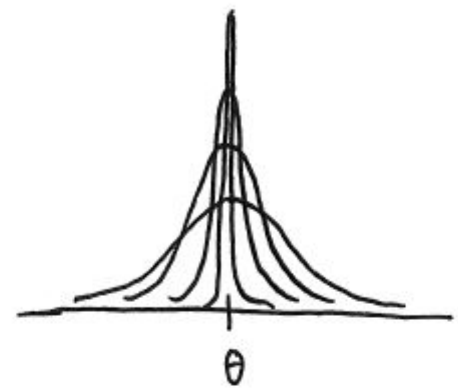
Finally, I will mention one additional criterion. $\hat{\theta}$ is a <u>consistent</u> estimator for θ if

$$\lim_{n \to \infty} P(|\theta - \hat{\theta}_n| < \delta) = 1$$

Statistics---criteria for assessing estimators

Finally, I will mention one additional criterion. $\hat{\theta}$ is a **consistent** estimator for θ if

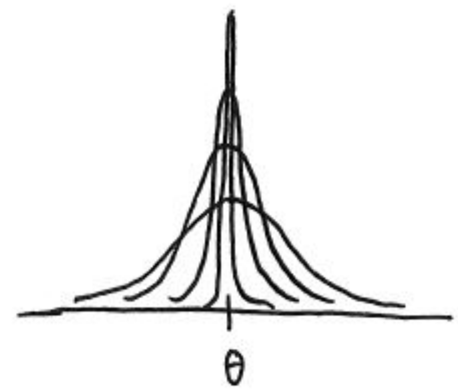
$$\lim_{n\to\infty} P(|\theta - \hat{\theta}_n| < \delta) = 1$$



Statistics---criteria for assessing estimators

Finally, I will mention one additional criterion. $\hat{\theta}$ is a <u>consistent</u> estimator for θ if

$$\lim_{n \to \infty} P(|\theta - \hat{\theta}_n| < \delta) = 1$$



Roughly, an estimator is consistent if its distribution collapses to a single point at the true parameter as $n \to \infty$.

Statistics---criteria for assessing estimators

These criteria are probably the most important reasons for choosing an estimator, but we also might consider how easy the estimator is to compute, how <u>robust</u> it is to assumptions we've made (i.e., whether the estimator will still do a decent job if we've assumed the wrong distribution), etc.

Statistics---criteria for assessing estimators

These criteria are probably the most important reasons for choosing an estimator, but we also might consider how easy the estimator is to compute, how <u>robust</u> it is to assumptions we've made (i.e., whether the estimator will still do a decent job if we've assumed the wrong distribution), etc.

For instance, it turns out that the 2-times-the-sample-median estimator I mentioned will have less bias than 2 times the sample mean if we've misspecified the tail probabilities of the underlying distribution.

Statistics--frameworks for finding estimators

We now know how to figure out if an estimator is good once we have one, but how do we get one in the first place?

Statistics--frameworks for finding estimators

We now know how to figure out if an estimator is good once we have one, but how do we get one in the first place?

There are two main frameworks for deriving estimators, the Method of Moments and Maximum Likelihood Estimation. (We've seen examples of both.)

Statistics--frameworks for finding estimators

We now know how to figure out if an estimator is good once we have one, but how do we get one in the first place?

There are two main frameworks for deriving estimators, the Method of Moments and Maximum Likelihood Estimation. (We've seen examples of both.)

A third framework is to think of something clever. (We've seen a couple of examples of this, too.)

Statistics--frameworks for finding estimators

The Method of Moments (developed in 1894 by Karl Pearson, the father of mathematical statistics):

First have to define moments.

<u>population moments (about the origin)</u>: $E(X)$, $E(X^2)$, $E(X^3)$, . . .

<u>sample moments</u>: $(1/n)\Sigma X_i$, $(1/n)\Sigma X_i^2$, $(1/n)\Sigma X_i^3$, . .

Statistics--frameworks for finding estimators

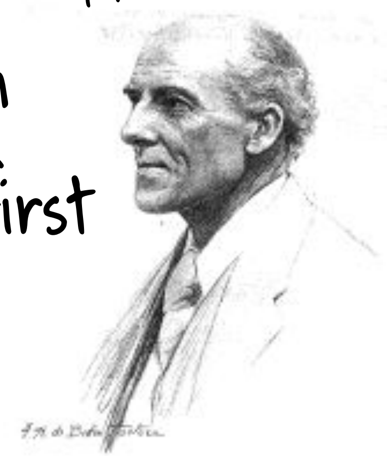
The Method of Moments (developed in 1894 by Karl Pearson, the father of mathematical statistics):

First have to define moments.

<u>population moments (about the origin)</u>: $E(X)$, $E(X^2)$, $E(X^3)$, . . .

<u>sample moments</u>: $(1/n)\Sigma X_i$, $(1/n)\Sigma X_i^2$, $(1/n)\Sigma X_i^3$, . .

To estimate a parameter, equate the first population moment (a function of the parameter), to the first sample moment, and solve for the parameter.



Statistics---method of moments

We've seen an example, $\hat{\theta}_2$ in the uniform example.

The first population moment, $E(X)$, of a $U[0,\theta]$, is $\theta/2$.

The first sample moment is $(1/n)\Sigma X_i$.

So equate the population and sample moments, stick a hat on θ, and solve for $\hat{\theta}$.

$$\hat{\theta}/2 = (1/n)\Sigma X_i$$

so,

$$\hat{\theta} = (2/n)\Sigma X_i$$

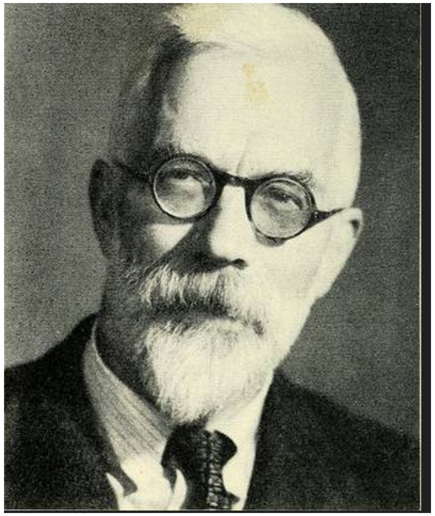
Statistics---method of moments

What if you have more than one parameter to estimate? No problem---just use as many sample and population moments as necessary. Each one is called a "moment condition." If you have k parameters to estimate, you will have k moment conditions. In other words, you will have k equations in k unknowns to solve.

Statistics--frameworks for finding estimators

Maximum Likelihood Estimation (of unclear origin going back centuries, but idea usually attributed to Lagrange, circa 1770, and analytics to R.A. Fisher, circa 1930):

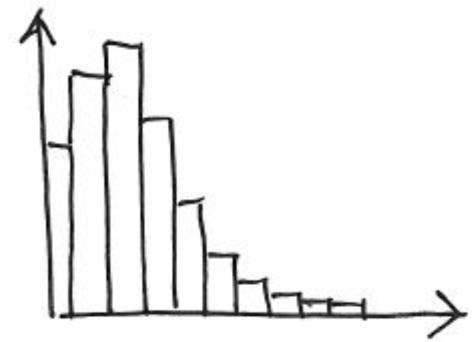
The maximum likelihood estimator of a parameter θ is the value $\hat{\theta}$ which most likely would have generated the observed sample.

Statistics---maximum likelihood

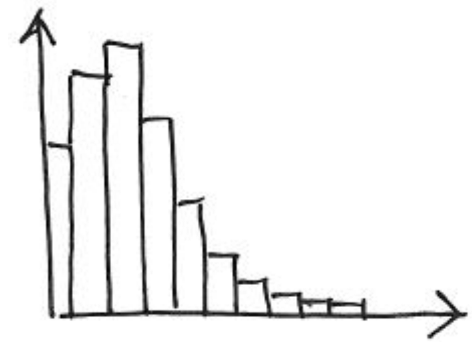
Here's a histogram of our data:

(Remember we think of the histogram as the empirical counterpart of the PDF of a random variable.)



Statistics---maximum likelihood

Here's a histogram of our data:



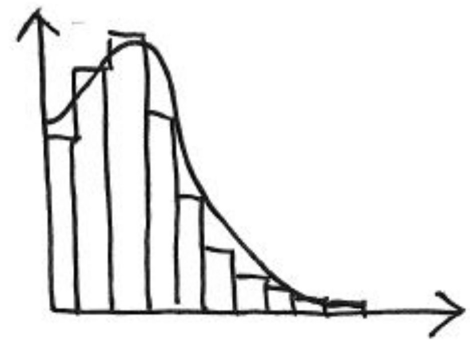
(Remember we think of the histogram as the empirical counterpart of the PDF of a random variable.)

Here are some options of PDFs that could have given rise to our data:

(Where did we get these? Well, we assumed a particular "family" of distributions and varied the parameter(s).)

Statistics---maximum likelihood

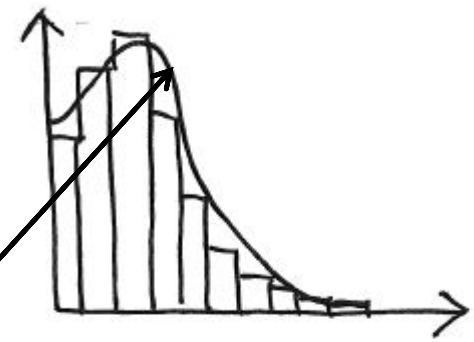
Which of those possible PDFs is most likely to have produced our data? The parameter(s) which describe it are the maximum likelihood estimate(s).



Statistics---maximum likelihood

Which of those possible PDFs is most likely to have produced our data?

The parameter(s) which describe it are the maximum likelihood estimate(s).



$\hat{\theta}$ is the value of the parameter associated with this particular "best fit" member of the family of distributions.

Statistics---maximum likelihood

Conceptually, makes sense. Operationally, how do we find the one of a bunch of PDFs that is most likely to have produced our data?

Statistics---maximum likelihood

Conceptually, makes sense. Operationally, how do we find the one of a bunch of PDFs that is most likely to have produced our data?

We have to sort of reinterpret the joint PDF of the data, or random sample. We have to think of it as a function of its parameters and maximize it over those parameters.

Statistics---maximum likelihood

Conceptually, makes sense. Operationally, how do we find the one of a bunch of PDFs that is most likely to have produced our data?

We have to sort of reinterpret the joint PDF of the data, or random sample. We have to think of it as a function of its parameters and maximize it over those parameters.

In other words, we define a function $L(\theta|x)$, the likelihood function, which is simply the joint PDF of the data, $\prod_i f(x|\theta)$ for an i.i.d. random sample.

Statistics---maximum likelihood

So $L(\theta|x) = \Pi_i f(x|\theta)$ and we just maximize L over θ in Θ. (We can use any monotonic transformation of L and it will still be maximized by the same θ. Computationally, it is often easier to take the log of L and maximize that because then the product becomes a sum, which is easier to deal with.)

Statistics---maximum likelihood

So $L(\theta|x) = \prod_i f(x|\theta)$ and we just maximize L over θ in Θ. (We can use any monotonic transformation of L and it will still be maximized by the same θ. Computationally, it is often easier to take the log of L and maximize that because then the product becomes a sum, which is easier to deal with.)

It's good practice to write down joint PDFs, maybe take the logs, take the derivatives with respect to θ, set the derivatives equal to zero, and solve for the maximum likelihood estimators. You may do that if you would like, but we won't do it here.

Statistics---maximum likelihood

Instead we will do a couple of examples that do not involve serious computation to find the maximum but rather just some clever reasoning.

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$f_X(x) = \begin{cases} 1/\theta & x \text{ in } [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

For the MLE, obviously wouldn't pick any $\hat{\theta} < X_{(n)}$. Why?

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$f_X(x) = \begin{cases} 1/\theta & x \text{ in } [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

For the MLE, obviously wouldn't pick any $\hat{\theta} < X_{(n)}$ because such a value would be impossible (probability 0), so can't maximize the likelihood function.

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$f_X(x) = \begin{cases} 1/\theta & x \text{ in } [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

So, write down the likelihood function:

$$L(\theta) = \begin{cases} (1/\theta)^n & x_i \text{ in } [0,\theta],\ i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$f_X(x) = \begin{cases} 1/\theta & x \text{ in } [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

In general, how do we get the likelihood function when we have an i.i.d. random sample?

So, write down the likelihood function:

$$L(\theta) = \begin{cases} (1/\theta)^n & x_i \text{ in } [0,\theta],\ i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$f_X(x) = \begin{cases} 1/\theta & x \text{ in } [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

In general, how do we get the likelihood function when we have an i.i.d. random sample?
It's the product of the n f_X's.

So, write down the likelihood function:

$$L(\theta) = \begin{cases} (1/\theta)^n & x_i \text{ in } [0,\theta],\ i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$f_X(x) = \begin{cases} 1/\theta & x \text{ in } [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

So, write down the likelihood function:

$$L(\theta) = \begin{cases} (1/\theta)^n & x_i \text{ in } [0,\theta],\ i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$f_X(x) = \begin{cases} 1/\theta & x \text{ in } [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

This is the same as saying that the nth order statistic is less than θ.

So, write down the likelihood function:

$$L(\theta) = \begin{cases} (1/\theta)^n & x_i \text{ in } [0,\theta],\ i = 1, . ., n \\ 0 & \text{otherwise} \end{cases}$$

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$f_X(x) = \begin{cases} 1/\theta & x \text{ in } [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

So, write down the likelihood function:

$$L(\theta) = \begin{cases} (1/\theta)^n & X_{(n)} <= \theta \\ 0 & \text{otherwise} \end{cases}$$

Can write in terms of order statistics instead.

Statistics---example

X_i i.i.d. $U[0,\theta]$

$$f_X(x) = \begin{cases} 1/\theta & x \text{ in } [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

So, write down the likelihood function:

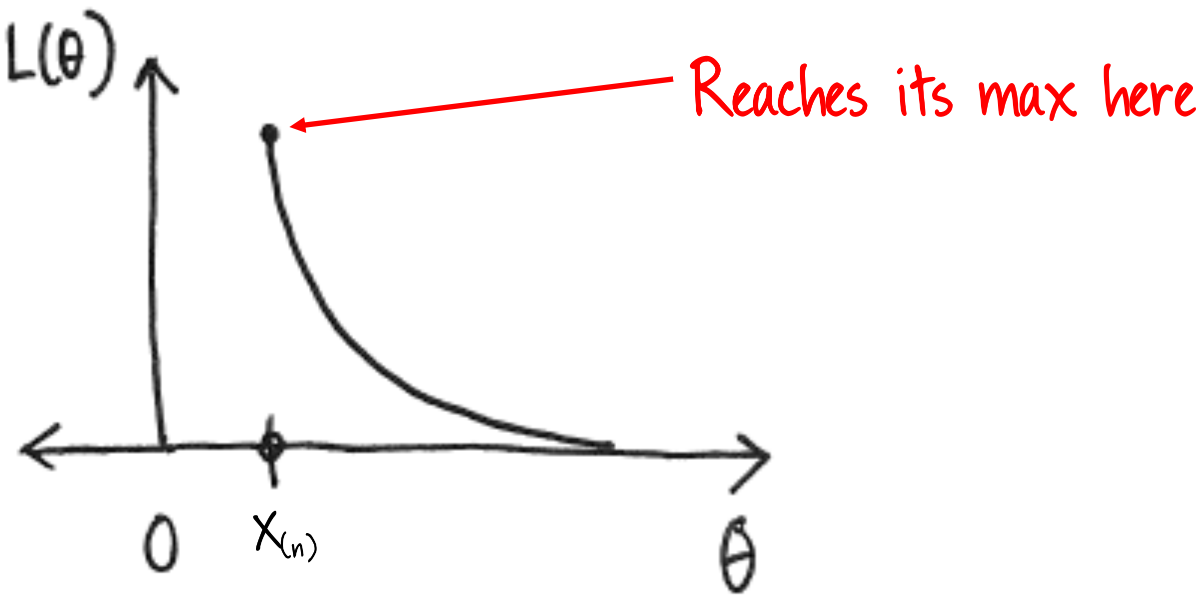
$$L(\theta) = \begin{cases} (1/\theta)^n & X_{(n)} <= \theta \\ 0 & \text{otherwise} \end{cases}$$

So, $\hat{\theta} = \max\{x_1,\ldots,x_n\}$

Statistics---example

Let's look at it graphically.

The likelihood function is 0 up until the nth order statistic, the smallest value it could be. Then it has this $(1/\theta)^n$ shape:



Statistics---example

X_i i.i.d. $U[\theta-1/2, \theta+1/2]$

$$f_X(x) = \begin{cases} 1 & x \text{ in } [\theta-1/2, \theta+1/2] \\ 0 & \text{otherwise} \end{cases}$$

Statistics---example

X_i i.i.d. $U[\theta-1/2, \theta+1/2]$

$$f_X(x) = \begin{cases} 1 & x \text{ in } [\theta-1/2, \theta+1/2] \\ 0 & \text{otherwise} \end{cases}$$

So, write down the likelihood function:

$$L(\theta) = \begin{cases} 1 & \theta \text{ in } [X_{(n)}-1/2, X_{(1)}+1/2] \\ 0 & \text{otherwise} \end{cases}$$

Statistics---example

X_i i.i.d. $U[\theta-1/2,\theta+1/2]$

$$f_X(x) = \begin{cases} 1 & x \text{ in } [\theta-1/2,\theta+1/2] \\ 0 & \text{otherwise} \end{cases}$$

Again, can write in terms of order statistics instead.

So, write down the likelihood function:

$$L(\theta) = \begin{cases} 1 & \theta \text{ in } [X_{(n)}-1/2, X_{(1)}+1/2] \\ 0 & \text{otherwise} \end{cases}$$

Statistics---example

X_i i.i.d. $U[\theta-1/2,\theta+1/2]$

$$f_X(x) = \begin{cases} 1 & x \text{ in } [\theta-1/2,\theta+1/2] \\ 0 & \text{otherwise} \end{cases}$$

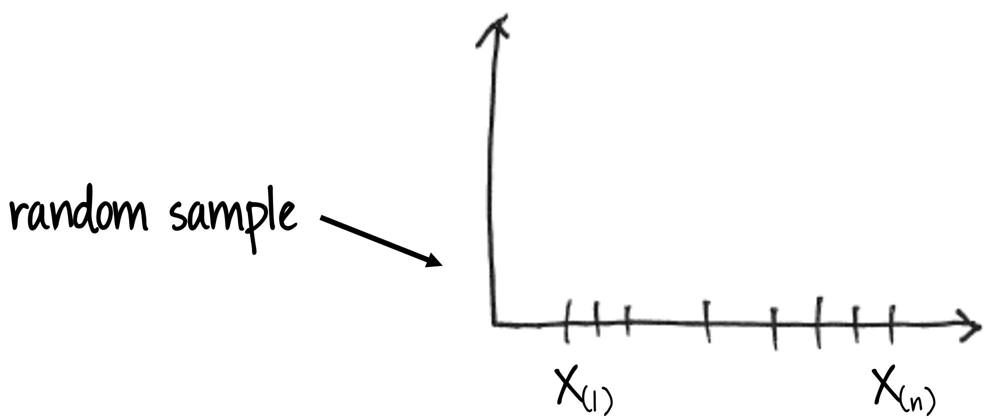
So, write down the likelihood function:

$$L(\theta) = \begin{cases} 1 & \theta \text{ in } [X_{(n)}-1/2,X_{(1)}+1/2] \\ 0 & \text{otherwise} \end{cases}$$

So, maximized for any value in that interval.

Statistics---example

Let's look at this one graphically, too.

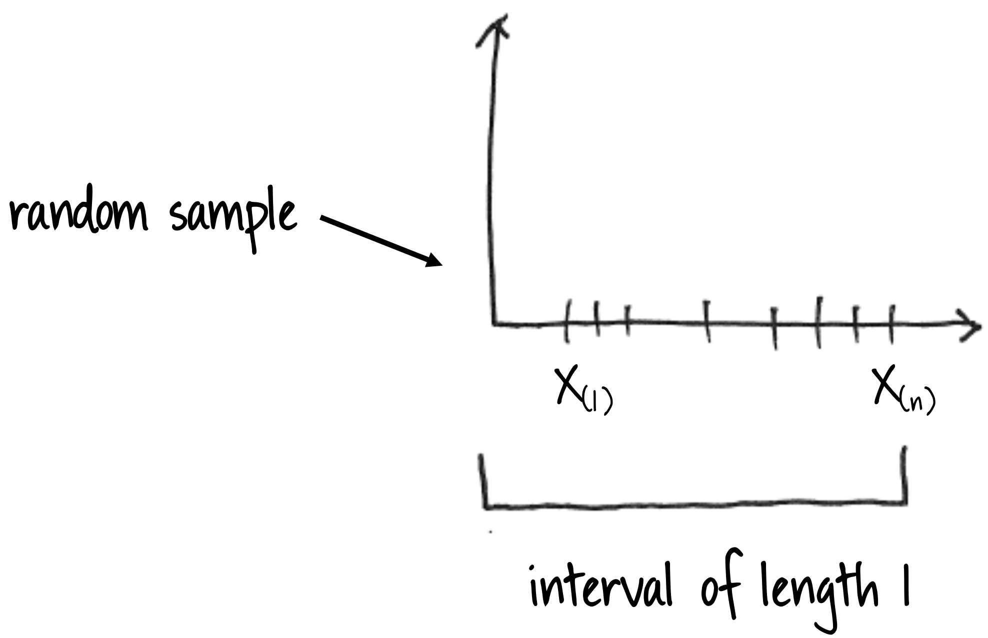


The interval that is length 1 centered at θ is here somewhere. And it must encompass all of the data.

Statistics---example

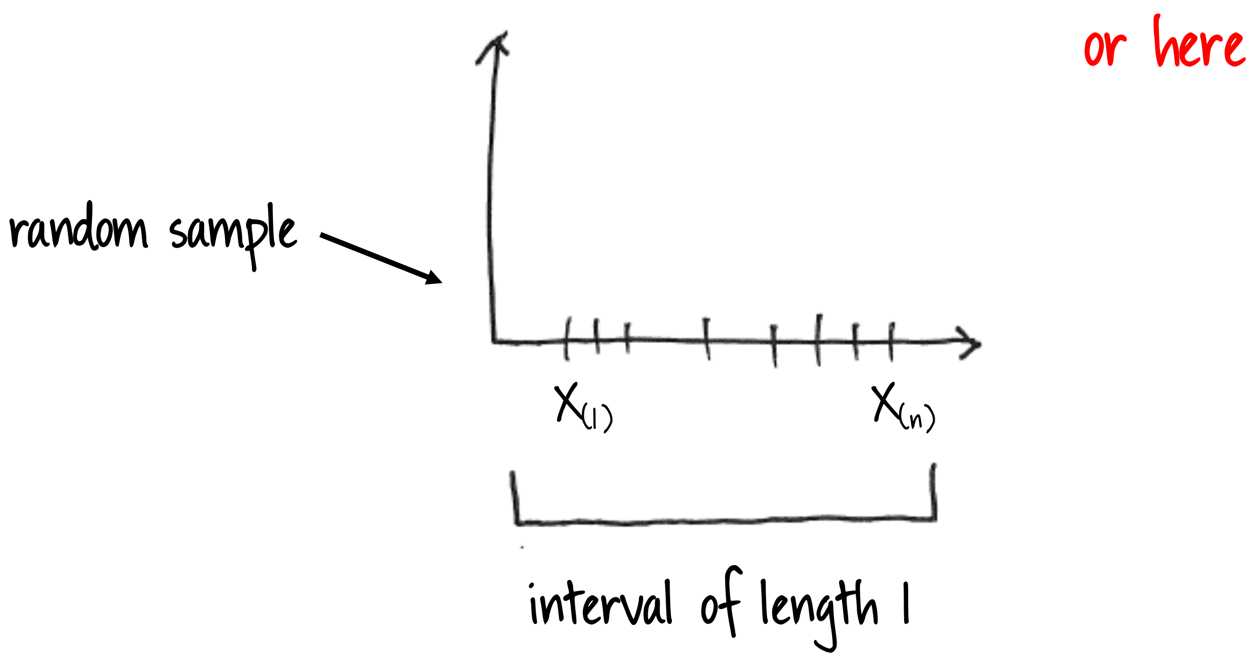
Let's look at this one graphically, too.

interval could be here



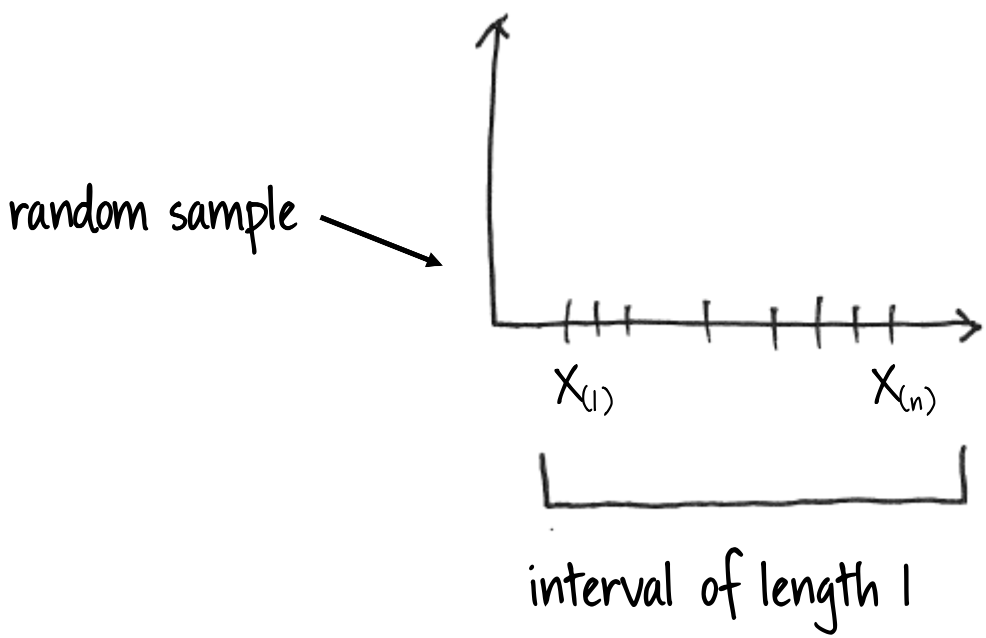
Statistics---example

Let's look at this one graphically, too.



Statistics---example

Let's look at this one graphically, too.

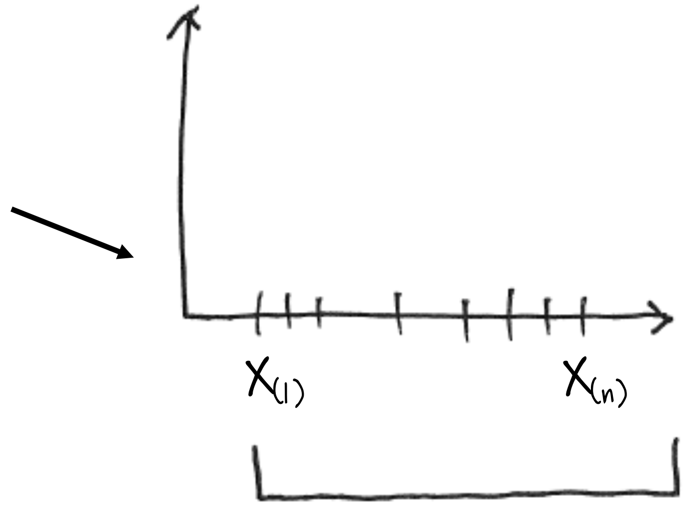


or here

Statistics---example

Let's look at this one graphically, too.

random sample



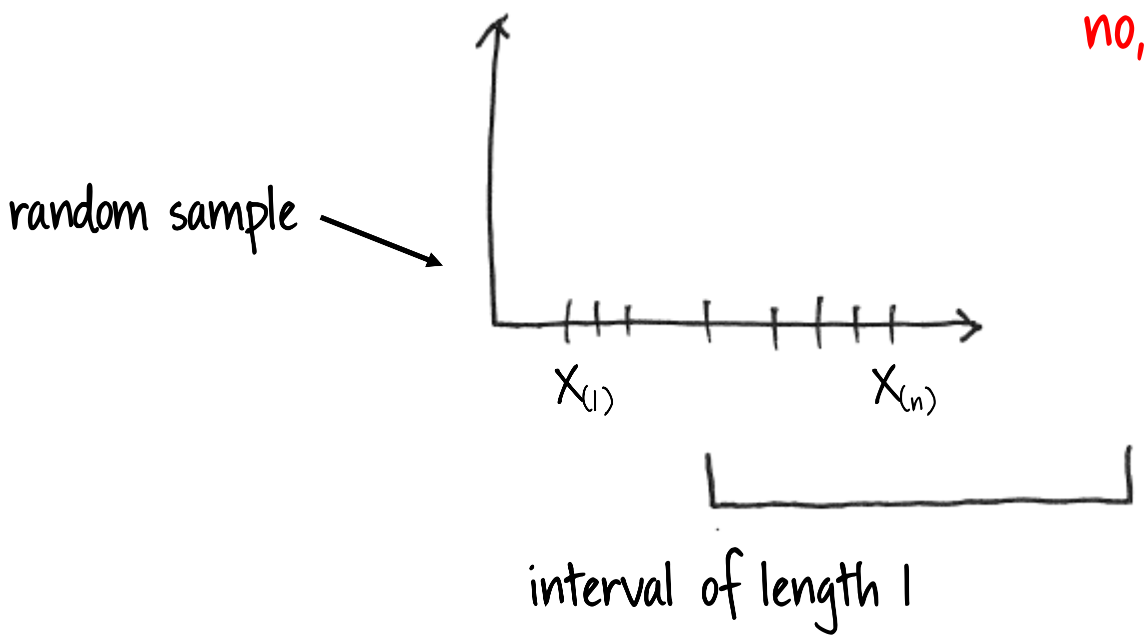
interval of length 1

or here
and, in fact, all of these possibilities are equally likely.

Statistics---example

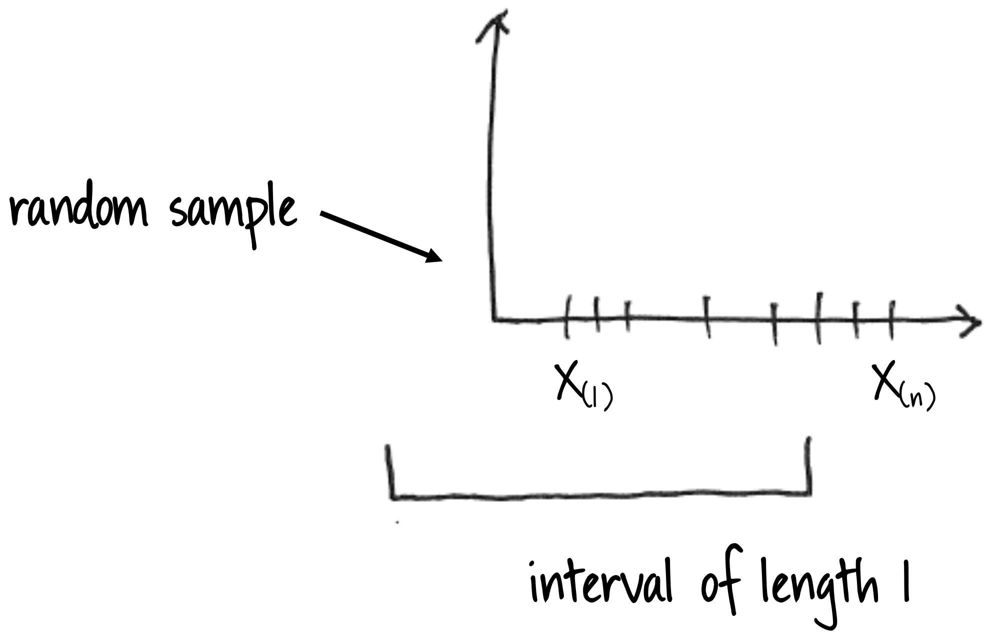
Let's look at this one graphically, too.

no, not here



Statistics---example

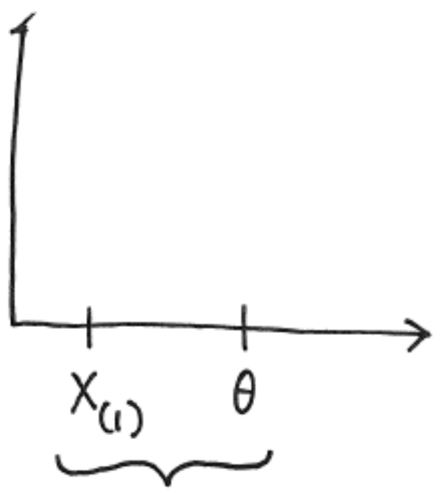
Let's look at this one graphically, too.



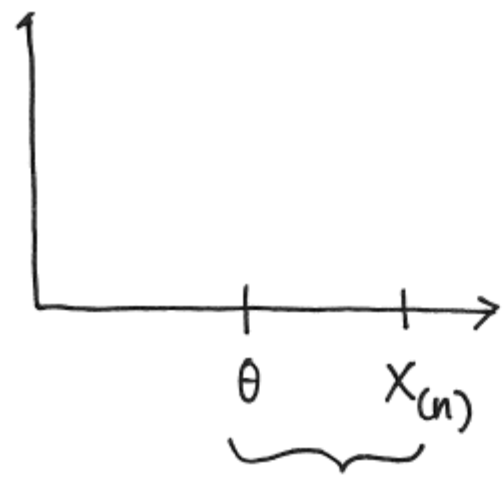
no, not here

Statistics---example

So, in other words,



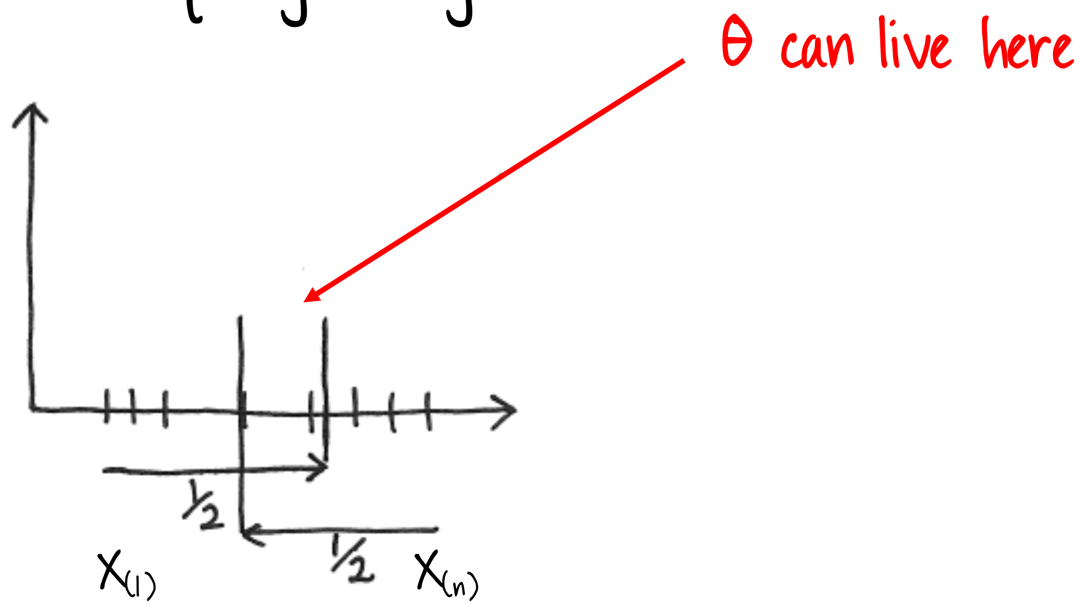
θ can be at most 1/2 above the 1st order statistic.



θ can be at most 1/2 below the nth order statistic.

Statistics---example

So, that gives us a window in which θ can live, and all values of θ in that window are equally likely.



$\hat{\theta}$ can be any value in $[X_{(n)}-1/2, X_{(1)}+1/2]$

Statistics---maximum likelihood

Maximum likelihood estimators have some favorable properties:

1. If there is an efficient estimator in a class of consistent estimators, MLE will produce it.
2. Under certain regularity conditions, MLEs will have asymptotically normal distributions (like a CLT for MLEs).

Statistics---maximum likelihood

Does this mean that maximum likelihood is always the right thing to do?

1. They can be biased (we saw an example).
2. They might be difficult to compute.
3. They can be sensitive to incorrect assumptions about the underlying distribution, more so than other estimators.

Summary to date

Probability basics

Introduced concept and talked about simple sample spaces, independent events, conditional probabilities, Bayes Rule

Random variables

Defined a random variable, discussed ways to represent distributions (PF, PDF, CDF), covered random variable versions of concepts above

Functions of random variables

Saw some basic strategies and several important examples

Summary to date

Moments

Defined moments of distributions and learned many techniques and properties to help compute moments of functions of random variables

Special distributions

Binomial, hypergeometric, geometric, negative binomial, Poisson, exponential, uniform, normal

Estimation

CLT, had general discussion and discussion about sample mean, criteria for assessing, frameworks for deriving