

Lab 3: Parameter Estimation

Objective

Explore properties of estimators and understand what makes an estimator preferred.

Definitions

- estimator vs. estimate
- maximum likelihood estimation: likelihood function, log-likelihood
- method of moments estimation
- bias, unbiased estimator
- efficiency of estimators
- mean square error (MSE)
- bias-variance trade-off

Introduction

In class this week, we went over two procedures for estimating parameters: **maximum likelihood estimation** and **method of moments**. There are other methods of estimation that may exist in a given context, including using "plug-in" estimators. This begs the question of which method is best. In this lab, you will explore properties of estimators and using these properties learn what criteria we think good estimators should satisfy. Each property provides a "sniff test": an estimator that fails these just doesn't smell right.

Activities

Getting Started: Navigate to your class folder structure. Within your "Labs" folder make a subfolder called "Lab3". Next, download the lab notebook .Rmd file for this lab from Blackboard and save it in your "Lab3" folder. There are no datasets used in this lab.

Within RStudio, navigate to your "Lab3" folder via the file browser in the lower right pane and then click "More > Set as working directory". Get set to write your observations and R commands in an R Markdown file by opening the "lab3_notebook.Rmd" file in RStudio. Remember to add your name as the author in line 3 of the document. For this lab, enter all of your commands into code chunks in the lab notebook. You can still experiment with code in an R script, if you want. To set up an R Script in RStudio, in the upper left corner click "File > New File > R script". A new tab should open up in the upper left pane of RStudio.

Maximum Likelihood Estimation: Before we get into the properties, let's revisit maximum likelihood estimation.

Likelihood Function, Maximum Likelihood Estimate

Suppose X_1, \dots, X_n represent a random sample from a probability distribution with associated parameter θ and pmf/pdf given by $f(x; \theta)$. The **likelihood function** $L(\theta) = L(\theta|x_1, \dots, x_n)$ gives the likelihood of θ , given the observed sample values x_1, \dots, x_n , and is calculated as follows:

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdots f(x_n; \theta)$$

The **maximum likelihood estimate** (MLE) $\hat{\theta}_{MLE}$ is a value of θ that maximizes the likelihood function, or equivalently that maximizes the log-likelihood function: $\ln L(\theta)$.

So, the likelihood function $L(\theta)$ is a function of the unknown parameter θ , and we estimate θ by maximizing $L(\theta)$. Remember that we can use calculus to find the value of θ that maximizes the likelihood by setting the derivative equal to 0, $L'(\theta) = 0$, and then solving for θ to find the MLE.

In practice, we usually maximize the **log-likelihood**, because taking the logarithm of a product results in a sum:

$$\ln L(\theta) = \ln [f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta)]$$

$$= \ln f(x_1; \theta) + \ln f(x_2; \theta) + \dots + \ln f(x_n; \theta) = \sum_{i=1}^n \ln f(x_i; \theta)$$

In most cases, we are able to find a closed-form expression for the MLE. However, this is not always possible, as the following example demonstrates.

Example: Suppose X_1, \dots, X_n are a random sample from the Cauchy distribution, which has pdf given by $f(x; \theta) = \frac{1}{\pi(1 + (x - \theta)^2)}$, for $x, \theta \in \mathbb{R}$. The likelihood function for θ is

$$L(\theta) = \frac{1}{\pi^n \prod_{i=1}^n [1 + (X_i - \theta)^2]}. \quad (1)$$

Thus, $L(\theta)$ will be maximized when $\prod_{i=1}^n [1 + (X_i - \theta)^2]$ is minimized, or equivalently when $\sum_{i=1}^n \ln(1 + (X_i - \theta)^2)$ is minimized. The value of θ that minimizes this expression must be determined by numerical methods.

Pause for Reflection #1:

1. On a separate piece of paper, write out the details for deriving the likelihood function $L(\theta)$ in Equation (1).
2. Next, explain why $L(\theta)$ will be maximized when the expression $\sum_{i=1}^n \ln(1 + (X_i - \theta)^2)$ is minimized. Type your response directly into your lab notebook in RStudio.
3. Finally, on the same piece of paper you worked out step 1, take the derivative (with respect to θ) of the sum expression given in step 2 to see why the mathematical approach of setting the derivative equal to 0 will not work in this example.

Take a picture of your written work for steps 1 and 3 and upload it to your Lab3 folder in order to include in your lab notebook.

Continuing with the Cauchy distribution, suppose you make the following observations for a random sample of size four: $x_1 = 1, x_2 = 2, x_3 = 2$, and $x_4 = 3$. Then, to maximize $L(\theta)$, you need to minimize

$$\ln(1 + (1 - \theta)^2) + \ln(1 + (2 - \theta)^2) + \ln(1 + (2 - \theta)^2) + \ln(1 + (3 - \theta)^2)$$

Since we cannot find the minimum of the above analytically, you will use the `optimize()` function in R. Type the following in a code chunk in your Lab 3 notebook and run each line:

```
x = c(1, 2, 2, 3)
g = function(theta) sum(log(1 + (x - theta)^2))
optimize(g, interval = c(0, 4))
```

Pause for Reflection #2

1. In your lab notebook, describe what each of the three lines of code above are doing. You may find it helpful to type `?c`, `help("function")`, and `?optimize` one at a time in the console window to pull up info in the Help window (lower right pane) for each of these commands.
2. Based on the results of this code, what is the maximum likelihood estimate of θ based on the given data?

Unbiasedness: The first property of estimators that we will consider is *bias*. An estimator $\hat{\theta}$ is biased if, on average, it tends to be too high or too low, relative to the true value of θ . Formally, this is defined using expected values:

Definition 3.1

The **bias** of an estimator $\hat{\theta}$ is given by

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

In other words, Definition 3.1 states that a statistic used to estimate a parameter is **biased** when the mean of its sampling distribution is not equal to the true value of the parameter. We will explore sampling distributions more in depth in next week. For now, we will use R to *approximate* sampling distributions.

We like an estimator to be, on average, equal to the parameter it is estimating. That is, we like estimators that are **unbiased**, or equivalently, $\text{Bias}(\hat{\theta}) = 0$. You will show in the homework that the sample mean is always an unbiased estimator of the population mean μ . It can also be shown that the sample proportion is also an unbiased estimator of the population proportion.

The case of the sample variance is less straightforward. Given a sample of values x_1, x_2, \dots, x_n , the "plug-in" estimator of the population variance σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

However, in Lab 1, we defined the sample variance as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

which is computed in R using the function `var()`. Notice the difference between the two estimators, namely, the division by " n " versus " $n - 1$ ". It turns out that the plug-in estimator $\hat{\sigma}^2$ is biased, but the sample variance s^2 is unbiased.

Let's explore this in the context of the standard normal distribution: $N(\mu = 0, \sigma = 1)$. In this context, we know the value of the parameter we are estimating, namely $\sigma^2 = 1$. So we know that in order for an estimator of σ^2 to be unbiased, its expected value needs to equal 1. You will run a simulation in R to see how the two estimators, $\hat{\sigma}^2$ and s^2 , perform. With the following code (already added to the lab notebook for you), you will draw random samples of size 15 from $N(0, 1)$. For each sample, you will compute $\hat{\sigma}^2$ and s^2 and record the values. You will repeat this 1000 times.

```
sample.var = numeric(1000)      # object to store sample variances
plugin = numeric(1000)          # object to store plug-in estimates
n = 15                          # set sample size
for (i in 1:1000)
{
  x = rnorm(n)                  # draw a random sample of size n from N(0,1) popul.
  sample.var[i] = var(x)         # compute and store sample variance of ith sample
  plugin[i] = ((n-1)/n)*var(x)   # compute and store plug-in estimate from ith samp.
}
```

We can now investigate the results of the simulation by finding the mean for the estimates of σ^2 based on the two estimators $\hat{\sigma}^2$ (`plugin.var`) and s^2 (`sample.var`). We can also visualize the results using histograms.

Pause for Reflection #3

1. In your lab notebook, calculate the respective means for the 1000 samples of $\hat{\sigma}^2$ and s^2 you found with the simulation.
2. Code has been provided in your lab notebook to create histograms of the simulated values for $\hat{\sigma}^2$ and s^2 . Run the code chunk and answer the following: Do the results you found support the claim that $\hat{\sigma}^2$ is a biased estimator of σ^2 and s^2 is unbiased? Why or why not?

Efficiency: What happens when you have two estimators that are both unbiased? Which one should you use? The next property we consider, *efficiency*, provides a criterion for comparing unbiased estimators that depends on their variance.

Definition 3.2

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of θ and $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$, then $\hat{\theta}_1$ is said to be more **efficient** than $\hat{\theta}_2$.

We will again explore this property with a simulation, this time in the context of the uniform distribution on the closed interval $[0, \beta]$.

At the start of lab, we will find the **method of moments estimator** for β to be $\hat{\beta}_1 = 2\bar{X}$, i.e., twice the mean of a given sample. It can be shown that this estimator is unbiased (a fact we will prove later). Using maximum likelihood estimation, we can find another unbiased estimator of β given by $\hat{\beta}_2 = ((n+1)/n)X_{\max}$, where X_{\max} denotes the largest value in a random sample (it is referred to as the *largest order statistic*).

Use the following code (already provided in your lab notebook) to run a simulation to see how these two estimators perform in the specific context of drawing random samples of size 25 from uniform $[0,12]$.

```
beta.1hat = numeric(1000)
beta.2hat = numeric(1000)
for (i in 1:1000)
{
  x = runif(25, 0, 12)           # draw a random sample of size 25 from uniform[0,12]
  beta.1hat[i] = 2 * mean(x)
  beta.2hat[i] = ((25 + 1)/25) * max(x)
}
# descriptive statistics
mean(beta.1hat)
sd(beta.1hat)
mean(beta.2hat)
sd(beta.2hat)
# graphical comparison
hist(beta.1hat, xlim = c(8,16), ylim = c(0,650), xlab = "2*mean")
hist(beta.2hat, xlim = c(8,16), ylim = c(0,650), xlab = "((25+1)/25)*max")
```

Pause for Reflection #4

1. Do the results support the claim that both $\hat{\beta}_1$ (beta.1hat) and $\hat{\beta}_2$ (beta.2hat) are unbiased estimators for β ? Why or why not?
2. Which estimator is more efficient, i.e., which estimator exhibits a smaller amount of variability?
3. Given these results, which estimator do you think you should use?

Mean Square Error: The final criterion we consider combines both bias and variance. This is useful for comparing estimators that are not both unbiased. We may prefer an estimator with small bias and small variance over one that is unbiased but with large variance. The following definition provides a way to quantify the preference.

Definition 3.3

The **mean square error** (MSE) of an estimator is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$.

MSE measures the average squared difference between the estimator and the parameter; it takes both the variability and bias of the estimator into account, as the following proposition shows.

Proposition 3.1

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

It follows from Proposition 3.1 that if $\hat{\theta}$ is unbiased, then $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta})$. So, for unbiased estimators, one is more efficient than a second if and only if its MSE is smaller. But, in general, when comparing two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ , we are faced with a trade-off between variability and bias.

Example: Let's explore this *bias-variance trade-off* in the context of the binomial distribution, where the number of trials n is known but the probability of "success" p is unknown. Let $X \sim \text{binomial}(n, p)$. The sample proportion X/n (the proportion of "successes" in n observed trials) is an unbiased estimator of p . Denote this estimator as \hat{p}_1 , then

$$E[\hat{p}_1] = E\left[\frac{X}{n}\right] = \frac{np}{n} = p.$$

Furthermore, the mean square error of the sample proportion is

$$\text{MSE}[\hat{p}_1] = \text{Var}(\hat{p}_1) = \text{Var}\left(\frac{X}{n}\right) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Consider the alternative estimator of p given by

$$\hat{p}_2 = \frac{X+1}{n+2},$$

which adds one artificial success and one failure to the real data.

Pause for Reflection #5

1. On a piece of paper, write out the details to derive the following:

$$E[\hat{p}_2] = \frac{np+1}{n+2} \text{ and } \text{Var}(\hat{p}_2) = \frac{np(1-p)}{(n+2)^2}.$$

2. Then, using Proposition 3.1, show that the mean square error for \hat{p}_2 is given by

$$\text{MSE}(\hat{p}_2) = \frac{np(1-p) + (1-2p)^2}{(n+2)^2}.$$

Take a picture of your written work for steps 1 and 2 and upload it to your Lab3 folder to include in your lab notebook. Refer to the code provided in the lab notebook for Reflection #1.

We can compare the two estimators \hat{p}_1 and \hat{p}_2 for p by comparing their mean squared errors. Note that we have the MSE for both estimators as a function of p . Thus, we can graphically compare the MSE for \hat{p}_1 and \hat{p}_2 by plotting curves in R using the following code. Note that we use $n = 16$ just to have a specific example to work with.

```
n = 16
curve(x*(1-x)/n, from=0, to=1, xlab="p", ylab="MSE")
curve((n*x*(1-x)+(1-2*x)^2)/(n+2)^2, add=TRUE, col="blue", lty=2)
```

The MSE for \hat{p}_1 is in solid black, and the MSE for \hat{p}_2 is the dashed blue curve.

Pause for Reflection #6

Inspect the graphs of the MSE curves and answer the following:

1. For approximately what values of p does \hat{p}_2 have smaller MSE than \hat{p}_1 ?
 2. For the values identified in step 1, even though \hat{p}_2 is biased, it has a smaller MSE than \hat{p}_1 . Comment on why \hat{p}_2 may be preferred over \hat{p}_1 as an estimator of p for these values.
 3. Now alter the code above to recreate the MSE graphs for the following sample sizes: $n = 30, n = 50, n = 100, n = 200$. What do you see is the effect of increasing the sample size?
-

Optional Reflection #7

Prove Proposition 3.1.

Lab 3: Parameter Estimation is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.