Lecture 16: Properties of Estimators and Methods of Estimation

MATH 697

Omidali Jazi

October 25, 2018

McGill University

Goals for this Chapter

- Properties of Estimators
  - Efficiency
  - Consistency
  - Sufficiency
- The Rao–Blackwell Theorem and Minimum-Variance Unbiased Estimation
- The Method of Moments
- The Method of Maximum Likelihood

Introduction

- In the previous chpater, we focused on some estimators that merit consideration on the basis of intuition such as sample mean, sample proportion, and sample variance.
- As indicated, if $\hat{\theta}_1$ and $\hat{\theta}_2$ denote two unbiased estimators for the same parameter θ, we would prefer to use the estimator with the smaller variance.
- In this chapter, we introduce a standard procedure to find the Best estimator.

Relative Efficiency

Definition 1

Given two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of a parameter θ, with variances $V(\hat{\theta}_1)$ and $V(\hat{\theta}_2)$, respectively. The relative efficiency of $\hat{\theta}_1$ and $\hat{\theta}_2$ is defined by

$$RE(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}$$

Relative Efficiency

- Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be unbiased estimators of θ. If $RE(\hat{\theta}_1, \hat{\theta}_2)$ is greater than 1, then $V(\hat{\theta}_2) > V(\hat{\theta}_1)$. In this case, $\hat{\theta}_1$ is a better unbiased estimator than $\hat{\theta}_2$
- Note that, if at least one of the estimators is biased, the relative efficiency is defined by the ratio of the MSE of the estimators.

Example

Suppose that $Y_1, Y_2, ..., Y_n$ denote a random sample from a Poisson distribution with mean λ. Consider $\hat{\lambda}_1 = (Y_1 + Y_2)/2$ and $\hat{\lambda}_2 = \bar{Y}$.

- Derive the relative efficiency of $\hat{\lambda}_1$ and $\hat{\lambda}_2$.
- Solution: Both $\hat{\lambda}_1$ and $\hat{\lambda}_1$ are unbiased estimators of λ. Further, $V(\hat{\lambda}_1) = \frac{\lambda}{2}$ and $V(\hat{\lambda}_2) = \frac{\lambda}{n}$. Thus,

$$RE(\hat{\lambda}_1, \hat{\lambda}_2) = \frac{2}{n}$$

This implies that when $n > 2$, $\hat{\lambda}_2$ is a more efficient estimator.

Consistency

- Let us examine the probability that the distance between the estimator and the target parameter, will be less than some arbitrary positive real number ε.
- We use subscript n on estimators to explicitly convey their dependence on the value of the sample size n.
- Recall the convergence in probability for a sequence of variables.
- Recall the weak law of large numbers for the sample mean.

Consistency

Theorem 2

The estimator $\hat{\theta}_n$ is said to be a consistent estimator of θ if, for any positive number ϵ,

$$\lim_{n\to\infty} P(\left|\hat{\theta}_n - \theta\right| \leq \varepsilon) = 1$$

or, equivalently,

$$\lim_{n\to\infty} P(\left|\hat{\theta}_n - \theta\right| > \varepsilon) = 0$$

Useful Result to Show Consistency

Theorem 3

An unbiased estimator $\hat{\theta}_n$ for θ is a consistent estimator of θ if

$$\lim_{n \to \infty} V(\hat{\theta}_n) = 0$$

- The Theorem implies that a sufficient condition for an unbiased estimator to be consistent is that the variance of the estimator tends to 0 when n is large. (Not all unbiased estimators are consistent)

Example

Let $Y_1, Y_2, ..., Y_n$ denote a random variable from a distribution with the mean μ and variance $\sigma^2 < \infty$.

- Show that $\bar{Y}_n$ is a consistent estimator of μ.
- Solution: Firstly, note that $\bar{Y}_n$ is an unbiased estimator of μ regradless of the distribution of the underlying population. Secondly, $V(\bar{Y}_n) = \frac{V(Y)}{n}$ which tends to 0 when n is large. Therefore, $\bar{Y}_n$ is a consistent estimator of μ.

Useful Results to Show the Consistency of Functions of Consistent Estimators

Theorem 4

Suppose that $\hat{\theta}_n$ converges in probability to θ and that $\hat{\theta}'_n$ converges in probability to θ'.

1. *$\hat{\theta}_n \pm \hat{\theta}'_n$ converge in probability to $\theta \pm \theta'$.*
2. *$\hat{\theta}_n \times \hat{\theta}'_n$ converges in probability to $\theta \times \theta'$.*
3. *If $\theta' \neq 0$, $\hat{\theta}_n / \hat{\theta}'_n$ converges in probability to θ / θ'.*
4. *If $g(\cdot)$ is a real-valued function that is continuous at θ, then $g(\hat{\theta}_n)$ converges in probability to $g(\theta)$.*

Example

Show that S_n^2 is a consistent estimator of $\sigma^2 = V(Y_i)$.

- Solution: First note that S_n^2 can be expanded as

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} Y_i^2 - n\bar{Y}_n^2\right\}$$
$$= \frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}_n^2\right\}$$

- It follows that when $n \to \infty$, $\frac{n}{n-1} \to 1$ and by the weak law of large numbers, the first term converges to $E(Y^2)$ in probability. Further, since $g(x) = x^2$ is a continuous function, by Thorem 4, the second term converges to μ^2 in probability. This completes the proof.

Asymptotic Normality of U_n/W_n

Theorem 5

Suppose that U_n has a distribution function that converges to a standard normal distribution function as $n \to \infty$. If W_n converges in probability to 1, then the distribution function of U_n/W_n converges to a standard normal distribution function.

Theoretical Justification for the Large-sample C.I for μ

Suppose that $Y_1, Y_2, ..., Y_n$ is a random sample of size n from a distribution with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$.

- show that the distribution function of $\frac{\bar{Y}_n - \mu}{S/\sqrt{n}}$ converges to a standard normal distribution function.
- Solution: First, from the previous example, S_n^2 is consistent estimator of σ^2 which implies that $S = \sqrt{S_n^2}$ is a consistent estimator of σ (even though, it is a biased estimator). That is, $\frac{S}{\sigma}$ conveges to 1 in probability.
  Secondly, the distribution function of $U_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}$ converges to the standard normal distribution following the CLT.
  Thus, by Thorem 5, the distribution function of $\frac{\sqrt{n}(\bar{Y}_n - \mu)/\sigma}{S/\sigma} = \frac{\bar{Y}_n - \mu}{S/\sqrt{n}}$ converges to the standard normal distribution as $n \to \infty$.

Sufficiency

- Up to this point, we have chosen estimators on the basis of intuition.
- At this stage, the actual sample values are no longer important; rather, we summarize the information in the sample that relates to the parameters of interest by using the statistics.
- Has this process of summarizing or reducing the data to the two statistics, $\bar{Y}$ and S^2, retained all the information about μ and σ^2 in the original set of n sample observations? Or has some information about these parameters been lost or obscured through the process of reducing the data.

Sufficiency

- We present methods for finding statistics that in a sense summarize all the information in a sample about a target parameter. Such statistics are said to have the property of sufficiency; or more simply, they are called sufficient statistics.
- As we will see in the next section, "good" estimators are (or can be made to be) functions of any sufficient statistic. Indeed, sufficient statistics often can be used to develop estimators that have the minimum variance among all unbiased estimators.

Sufficiency

Definition 6

Let $Y_1, Y_2, ..., Y_n$ denote a random sample from a probability distribution with unknown parameter θ. Then the statistic $U = g(Y_1, Y_2, ..., Y_n)$ is said to be sufficient for θ if the conditional distribution of $Y_1, Y_2, ..., Y_n$, given U, does not depend on θ.

Example

Let us consider the outcomes of n trials of a Bernoulli experiment, $Y_1, Y_2, ..., Y_n$ where

$$Y_i = \begin{cases} 1 & \text{If the } i\text{th trial is a success} \\ 0 & \text{If the } i\text{th trial is a failure} \end{cases}$$

Suppose that we are given a value of $T = \sum_{i=1}^{n} Y_i$, the number of successes among the n trials. If we know the value of T, can we gain any further information about p by looking at other functions of $Y_1, Y_2, ..., Y_n$?

Example

- By Definition 6,

$$P(Y_1 = y_1, ..., Y_n = y_n | T = t) = \frac{P(Y_1 = y_1, ..., Y_n = y_n, T = t)}{P(T = t)}$$

which equals

$$\frac{P(Y_1 = y_1, ..., Y_n = y_n)}{P(T = t)} \quad if \quad \sum_{i=1}^{n} Y_i = t$$

and 0 otherwise. It follows that $T \sim Binom(n, p)$ which entails that

$$P(Y_1 = y_1, ..., Y_n = y_n | T = t) = \frac{p^{\sum_{i=1}^{n} y_i} (1-p)^{n - \sum_{i=1}^{n} y_i}}{\binom{n}{t} p^t (1-p)^{(n-t)}}$$

$$= \frac{1}{\binom{n}{t}}$$

which does not depend on p. Thus, $T = \sum_{i=1}^{n} Y_i$ is a sufficient statistic for p.

Likelihood Function

Definition 7

Let $y_1, y_2, ..., y_n$ be sample observations taken on corresponding random variables $Y_1, Y_2, ..., Y_n$ whose distribution depends on a parameter θ. Then,

- If $Y_1, Y_2, ..., Y_n$ are discrete random variables, the likelihood of the sample, $L(y_1, y_2, ..., y_n|\theta)$, is defined to be the joint probability of $y_1, y_2, ..., y_n$.
- If $Y_1, Y_2, ..., Y_n$ are continuous random variables, the likelihood $L(y_1, y_2, ..., y_n|\theta)$ is defined to be the joint density evaluated at $y_1, y_2, ..., y_n$.

Factorization Criterion

Theorem 8

Let U be a statistic based on the random sample $Y_1, Y_2, ..., Y_n$. Then U is a sufficient statistic for the estimation of a parameter θ if and only if the likelihood $L(\theta) = L(y_1, y_2, ..., y_n|\theta)$ can be factored into two nonnegative functions,

$$L(\theta) = L(y_1, y_2, ..., y_n|\theta) = g(u, \theta) \times h(y_1, ..., y_n)$$

where $g(u, \theta)$ is a function of u and θ and $h(y_1, y_2, ..., y_n)$ is not a function of θ.

Example

Let $Y_1, Y_2, ..., Y_n$ be a random sample from $Exp(\lambda)$.

- Show that $\bar{Y}$ is a sufficient statistic for parameter λ.
- Solution: Using the factorization criterion

$$L(y_1, ..., y_n|\lambda) = f(y_1, ..., y_n|\lambda) = \frac{1}{\lambda}e^{-\frac{1}{\lambda}y_1} \times ... \times \frac{1}{\lambda}e^{-\frac{1}{\lambda}y_n}$$

$$= \frac{1}{\lambda^n}e^{-\frac{1}{\lambda}\sum_{i=1}^{n} y_i}$$

Because $g(u, \lambda) = \frac{1}{\lambda^n}e^{-\frac{1}{\lambda}\sum_{i=1}^{n} y_i}$ and $h(y_1, ..., y_n) = 1$. Thus, $U = \sum_{i=1}^{n} Y_i$ is a sufficient statistic for λ. That is, $\bar{Y}$ is a sufficient statistic for λ.

Example

Let $Y_1, Y_2, ..., Y_n$ be a random sample from the uniform distribution on $(0, \theta)$.

- Find a sufficient statistic for θ.
- Solution: Using the factorization criterion

$$\begin{aligned} L(y_1, ..., y_n|\theta) = f(y_1, ..., y_n|\theta) &= \frac{1}{\theta} I(0 < y_1 < \theta) \times ... \times \frac{1}{\theta} I(0 < y_n < \theta) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 < y_i < \theta) \\ &= \frac{1}{\theta^n} I(0 < y_{(n)} < \theta) \end{aligned}$$

where the indicator function $I(t) = 1$ if $t > 0$ and 0 otherwise. Further, $y_{(n)} = \max(y_1, ..., y_n)$.
Note that the last row followed because when for every i, $0 < y_i < \theta$, then $0 < \max(y_1, ..., y_n) < \theta$.
Therefore, $g(u, \theta) = \frac{1}{\theta^n} I(0 < y_{(n)} < \theta)$ and $h(y_1, ..., y_n) = 1$ which entails that $U = Y_{(n)}$ is a sufficient statistic for θ.

Remarks

- According to Definition 6, the random sample itself is a sufficient statistic (Trivial Case).
- Sufficient statistics reduce the data $Y_1, ..., Y_n$ which without losing any information remain sufficient for the population parameter.
- Generally, we would like to find a sufficient statistic that reduces the data in the sample as much as possible.
- Although many statistics are sufficient for the parameter θ associated with a specific distribution, application of the factorization criterion typically leads to a statistic that provides the "best" summary of the information in the data.

How to obtain Minimum Variance Estimators

- If $\hat{\theta}$ is an unbiased estimator for θ and if U is a statistic that is sufficient for θ, then there is a function of U that is also an unbiased estimator for θ and has no larger variance than $\hat{\theta}$.
- If we seek unbiased estimators with small variances, we can restrict our search to estimators that are functions of sufficient statistics.

The Rao–Blackwell Theorem

Theorem 9

Let $\hat{\theta}$ be an unbiased estimator for θ such that $V(\hat{\theta}) < \infty$. If U is a sufficient statistic for θ, define $\hat{\theta}^* = E(\hat{\theta}|U)$. Then, for all θ,

1. $E(\hat{\theta}^*) = \theta$
2. $V(\hat{\theta}^*) \leq V(\hat{\theta})$

Minimum-Variance Unbiased Estimation

- The Theorem implies that an unbiased estimator for θ with a small variance is or can be made to be a function of a sufficient statistic. If we have an unbiased estimator for θ, we might be able to improve it by using the result.
- Because many statistics are sufficient for a parameter θ associated with a distribution, which sufficient statistic should we use when we apply this theorem?
- For the distributions that we discuss in this text, the factorization criterion typically identifies a statistic U that best summarizes the information in the data about the parameter θ. Such statistics are called minimal sufficient statistics
- We apply the Theorem using U, we not only get an estimator with a smaller variance but also actually obtain an unbiased estimator for θ with minimum variance. Such an estimator is called a Minimum-Variance Unbiased Estimator (MVUE).

Example

Let $Y_1, Y_2, ..., Y_n$ denote a random sample from $Exp(\beta)$.

- Find the MVUE of $V(Y_i)$.
- Solution: First note that $V(Y_i) = \beta^2$. Further, from the previous example, $\bar{Y}$ is the minimal sufficient statistic for β. The MVUE is the unbiased estimator of β^2 based on $\bar{Y}$. Consider $\bar{Y}^2$. Then,

$$E(\bar{Y}^2) = V(\bar{Y}) + [E(\bar{Y})]^2 = \frac{\beta^2}{n} + \beta^2$$
$$= \beta^2(\frac{1}{n} + 1) = \beta^2(\frac{n+1}{n})$$

That is, $\frac{n}{n+1}\bar{Y}^2$ is the MVUE of $V(Y_i) = \beta^2$.

Example

Let $Y_1, Y_2, ..., Y_n$ be independent Bernoulli random variables with parameter p.

- Find the MVUE of $p(1-p)$
- Solution: In this example, we first find an unbiased estimator of $p(1-p)$ and improve the estimator by the conditional expection given the sufficient statistic (the Rao-Blackwell theorem). Consider

$$T = \begin{cases} 1 & Y_1 = 1 \text{ and } Y_2 = 0 \\ 0 & \text{Otherwise} \end{cases}$$

Note that T is an unbiased estimator of $p(1-p)$ since

$$E(T) = 1 \times P(Y_1 = 1, Y_2 = 0) + 0 = p(1-p)$$

- It follows that the conditional expection of T given the minimal sufficient statistic $U = \sum_{i=1}^{n} Y_i \sim binom(n, p)$ becomes

Example

$$
\begin{aligned}
E(T|U=u) &= P(Y_1=1, Y_2=0|U=u) \\
&= \frac{P(Y_1=1, Y_2=0, U=u)}{P(U=u)} = \frac{P(Y_1=1, Y_2=0, \sum_{i=1}^{n} Y_i = u)}{P(U=u)} \\
&= \frac{P(Y_1=1)P(Y_2=0)P(\sum_{i=3}^{n} Y_i = u-1)}{P(U=u)} \\
&= \frac{p(1-p)\begin{pmatrix} n-2 \\ u-1 \end{pmatrix} p^{u-1}(1-p)^{n-2-(u-1)}}{\begin{pmatrix} n \\ u \end{pmatrix} p^{u}(1-p)^{n-u}} = \frac{u(n-u)}{n(n-1)} \\
&= \frac{n\bar{y}(1-\bar{y})}{n-1}
\end{aligned}
$$

- That is, $\frac{n\bar{Y}(1-\bar{Y})}{n-1}$ is the MVUE of $p(1-p)$.
- As an exercise, use the approach in the previous example to show the answer.

The Method of Moments

- Recall that the kth moment of a random variable, taken about the origin, is

$$\mu'_k = E(Y^k)$$

- The corresponding kth sample moment is the average

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$$

- The method of moments is based on the intuitively appealing idea that sample moments should provide good estimates of the corresponding population moments.

The Method of Moments

- Choose as estimates those values of the parameters that are solutions of the equations $\mu'_k = m'_k$, for $k = 1, 2, ..., t$, where t is the number of parameters to be estimated.

Remark

- Sample moments are consistent estimators of the corresponding population moments. Because the estimators obtained from the method of moments obviously are functions of the sample moments, estimators obtained using the method of moments are usually consistent estimators of their respective parameters.

Example

Let $Y_1, \ldots, Y_n$ be random sample from $U(0, \theta)$.

- Use the method of moment to find an estimator for the parameter θ. Show that the resulting estimator is consistent.
- Solution: The first moment of random variable is $\mu_1' = E(Y) = \frac{\theta}{2}$ and The first moment of sample is $m_1' = \bar{Y}$. Solving the equation $\frac{\theta}{2} = \bar{Y}$ yields $\hat{\theta} = 2\bar{Y}$. The estimator is unbiased and $V(\hat{\theta}) = 4\frac{V(Y)}{n} = 4\frac{\theta^2}{12n}$ which tends to zero as $n \to \infty$. By Theorem 3, $\hat{\theta}$ is a consistent estimator of θ.
- What is the minimal sufficient statistic for θ.
- Solution: From the example in page 23, $Y_{(n)}$ is the minimal sufficient statistic for θ.
- What can you say about the efficiency of the method of moment's estimator for θ ?
- Solution: The estimator $\hat{\theta} = 2\bar{Y}$ is not a function of the minimal sufficient statistics. Thus, by the Rao-Blackwell theorem, it is not an efficient estimator.

Example

Let $Y_1, Y_2, ..., Y_n$ be random sample from $Gamma(\alpha, \beta)$ distribution.

- Find the method-of-moment's estimators for the unknown parameters α and β.
- Solution: There are two unknown parameters. Then, by the first two moments of random variables $E(Y)$ and $E(Y^2)$ and the first two moments of sample $\bar{Y}$ and $\bar{Y^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$, we need to solve the following equations

$$E(Y) = \alpha\beta = \bar{Y}$$
$$E(Y^2) = V(Y) + [E(Y)]^2 = \alpha\beta^2 + (\alpha\beta)^2 = \bar{Y^2}$$

The first equation leads to $\hat{\beta} = \frac{\bar{Y}}{\hat{\alpha}}$. Substituting in the second equation, after some algebraic operation, yields $\hat{\alpha} = \frac{n\bar{Y}^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$. Thus, $\hat{\beta} = \frac{\bar{Y}}{\hat{\alpha}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n\bar{Y}}$.

Remarks

- The method of moments finds estimators of unknown parameters by equating corresponding sample and population moments.
- Although the method is easy to employ and provides consistent estimators. the estimators derived by this method are often not functions of sufficient statistics. As a result, method-of-moments estimators are sometimes not very efficient.
- In many cases, the method-of-moments estimators are biased. The primary virtues of this method are its ease of use and that it sometimes yields estimators with reasonable properties.

Maximum Likelihood Estimation (MLE)

The Method of Maximum Likelihood

- We previously presented a method for deriving an MVUE for a target parameter: using the factorization criterion together with the Rao–Blackwell theorem which requires that we find some function of a minimal sufficient statistic that is an unbiased estimator for the target parameter.
- Although we have a method for finding a sufficient statistic, the determination of the function of the minimal sufficient statistic that gives us an unbiased estimator can be largely a matter of hit or miss.
- The method of moments is intuitive and easy to apply but does not usually lead to the best estimators.

Maximum Likelihood Estimation (MLE)

- Maximum Likelihood Estimation is a systematic technique for estimating parameters in a probability model from a data
- Suppose a sample $x_1, ..., x_n$ has been obtained from a probability model specified by mass or density function $f(x; \theta)$ depending on parameter(s) θ lying in parameter space Θ
- The **maximum likelihood estimate** or **MLE** is produced as follows:

Maximum Likelihood Estimation (MLE)

STEP 1 Write down the **likelihood function**, $L(\theta)$, where

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

that is, the product of the n mass/density function terms (where the ith term is the mass/density function evaluated at x_i) viewed as a function of θ

STEP 2 Take the natural log of the likelihood, and collect terms involving θ

Maximum Likelihood Estimation (MLE)

STEP 3 Find the value of $\theta \in \Theta$, $\hat{\theta}$, for which $\log L(\theta)$ is maximized, for example by differentiation. If θ is a single parameter, find $\hat{\theta}$ by solving

$$\frac{d}{d\theta} \{\log L(\theta)\} = 0$$

in the parameter space Θ. If θ is vector-valued, say $\theta = (\theta_1, ..., \theta_d)$, then find $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_d)$ by simultaneously solving the d equations given by

$$\frac{\partial}{\partial \theta_j} \{\log L(\theta)\} = 0 \qquad j = 1, \dots, d$$

in parameter space Θ.

Maximum Likelihood Estimation (MLE)

STEP 4 Check that the estimate $\hat{\theta}$ obtained in STEP 3 truly corresponds to a maximum in the (log) likelihood function by inspecting the second derivative of $\log L(\theta)$ with respect to θ. If

$$\frac{d^2}{d\theta^2} \{\log L(\theta)\} < 0$$

at $\theta = \hat{\theta}$, then $\hat{\theta}$ is confirmed as the m.l.e. of θ

This procedure is a systematic way of producing parameter estimates from sample data and a probability model; it can be shown that such an approach produces estimates that have good properties. After they have been obtained, the estimates can be used to carry out *prediction* of behaviour for future samples.

Maximum Likelihood Estimator

Definition 10 (Maximum Likelihood Estimator)

$$\widehat{\Theta}(\mathbf{x}) = \arg\max_{\Theta} L(\Theta; \mathbf{x}) \tag{1}$$

Invariance Principle: if $\widehat{\Theta}(\mathbf{x})$ is a MLE for Θ, then $g(\widehat{\Theta}(\mathbf{x}))$ is a MLE for $g(\theta)$

Example: Poisson MLE

A sample $y_1, ..., y_n$ is modelled by a Poisson distribution with parameter λ.

$$f(y; \lambda) = e^{-\lambda} \frac{\lambda^y}{y!} \qquad y = 0, 1, 2, ...$$

for some $\lambda > 0$.

- Find the MLE of λ and λ^2 **analytically**.
- Solution Note that the likelihood function is the joint probability mass function.

$$L(\lambda) = e^{-n\lambda} \lambda^{\sum_{i=1}^{n} y_i} \prod_{i=1}^{n} \frac{1}{y_i!}$$

Taking the natural log function, we obtain

$$\ell = \log L(\lambda) = -n\lambda + \sum_{i=1}^{n} y_i \log(\lambda) + \sum_{i=1}^{n} \log\left(\frac{1}{y_i!}\right)$$

By differentiating with respect to λ and setting to zero, we have $\hat{\lambda} = \bar{Y}$. Note that, here $\frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{\sum_{i=1}^{n} y_i}{\lambda^2} < 0$.

- Using the invariance property, the MLE of λ^2 is $\bar{Y}^2$.

Poisson MLE using `optim`

We can use the `stats::optim` function in `R` to find the MLE, provided we have a likelihood function. The `optim` can maximize (or minimize) an objective function using many different algorithms. This is referred to as **solving the objective function numerically**. Simulate some sample data generated from a Poisson distribution and solve for the MLE:

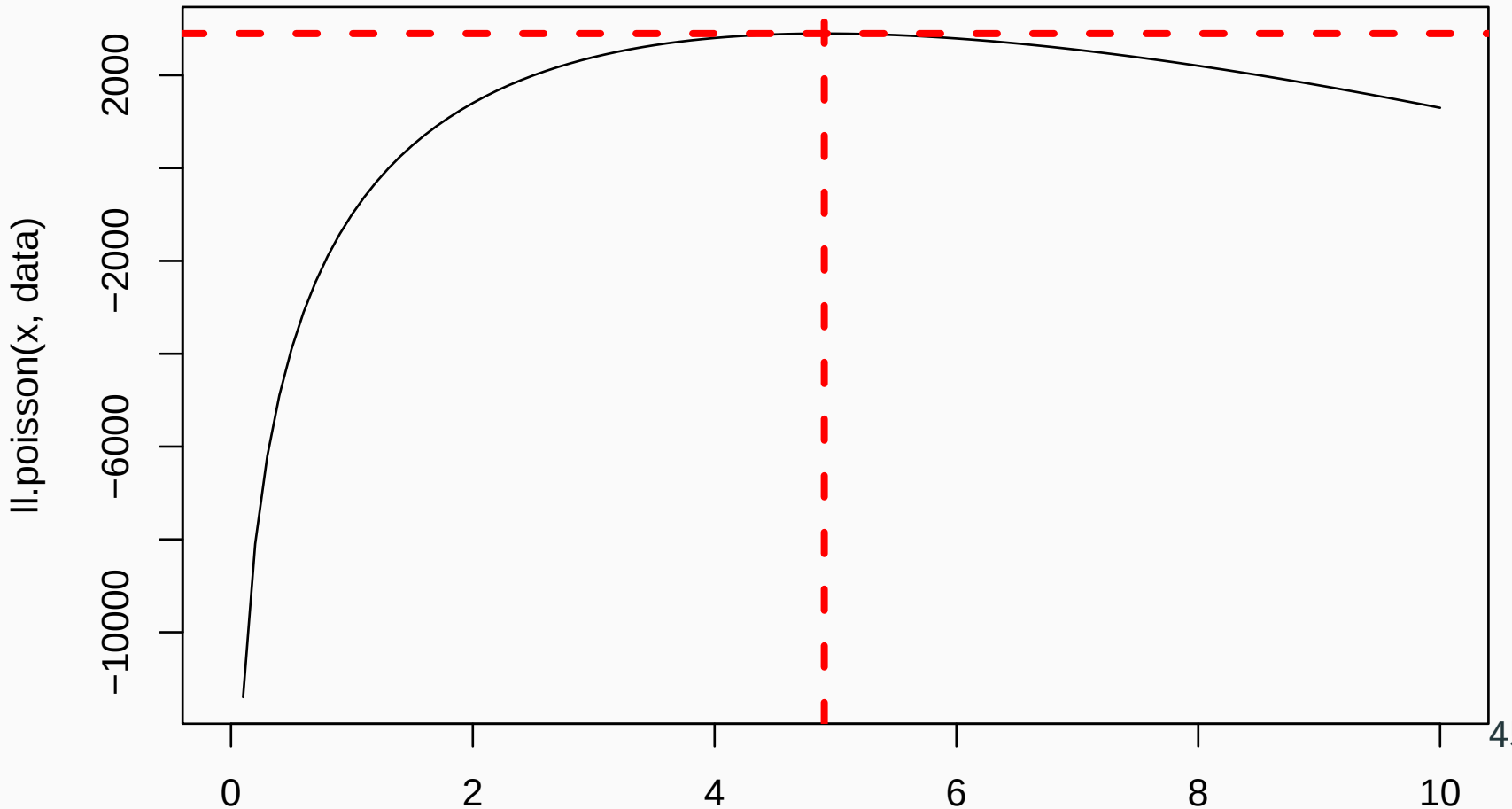
```
# define the objective function
ll.poisson <- function(lambda, x) {
 sum(x) * log(lambda) - length(x) * lambda
}
data <- rpois(1000, 5) # generate some data
# by default optim finds the min, but the negative min is the max
# therefore we need to use list(fnscale = -1)
opt <- optim(par = 2, fn = ll.poisson, method = "BFGS",
             control = list(fnscale = -1), x = data)

c(opt$par, mean(data))
```

```
## [1] 4.906999 4.907000
```

Poisson MLE using `optim`

We plot the objective function (in this case, it's the log-likelihood) and dotted red lines representing the value of the objective function at the value of λ that maximizes the log-likelihood

```
curve(ll.poisson(x, data), 0,10, xlab = "lambda")
abline(h = opt$value, v = opt$par, lty = 2, lwd = 3, col = "red")
```



Bernoulli MLE

A sample $x_1, ..., x_n$ is modelled by a Bernoulli distribution with unknown parameter denoted p

$$f(x;\theta) \equiv f(x;p) = p^x(1-p)^{1-x} \qquad x = 0, 1 ...$$

for some $p > 0$.

- Find the MLE of p.

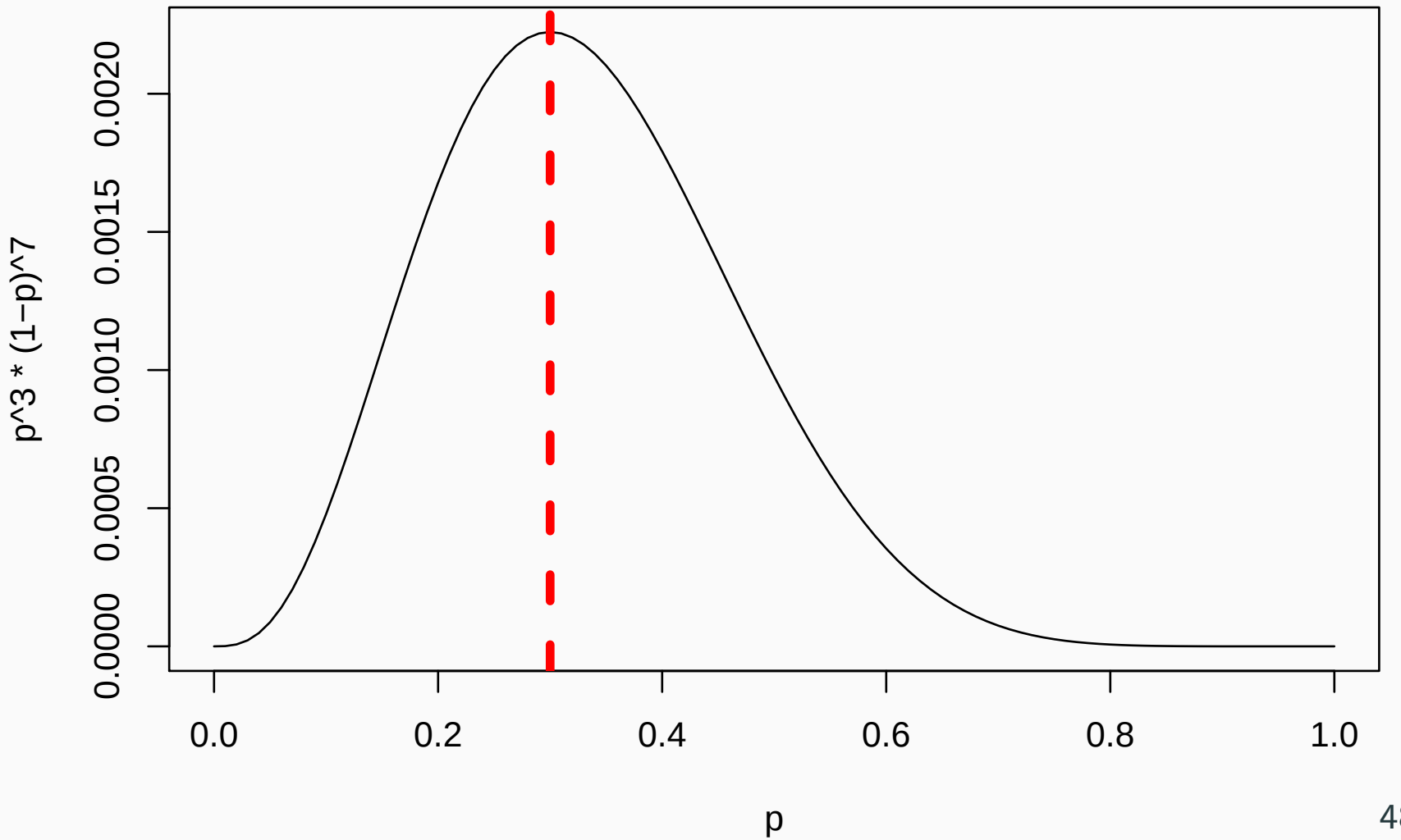
Bernoulli MLE Example

Example 11 (Bike Helmets)

A sample of ten new bike helmets manufactured by a company is obtained. Upon testing, it is found that the first, third, and tenth helmets are flawed, whereas the others are not. Let $p = P(\text{flawed helmet})$ and define $X_1, \ldots, X_{10}$ by $X_i = 1$ if the ith helmet is flawed and zero otherwise. Then the observed x_i's are $1, 0, 1, 0, 0, 0, 0, 0, 0, 1$. For what value of p is the observed sample **most likely to have occurred**? Would anything change if we had been told only that among the ten helmets there were three that were flawed?

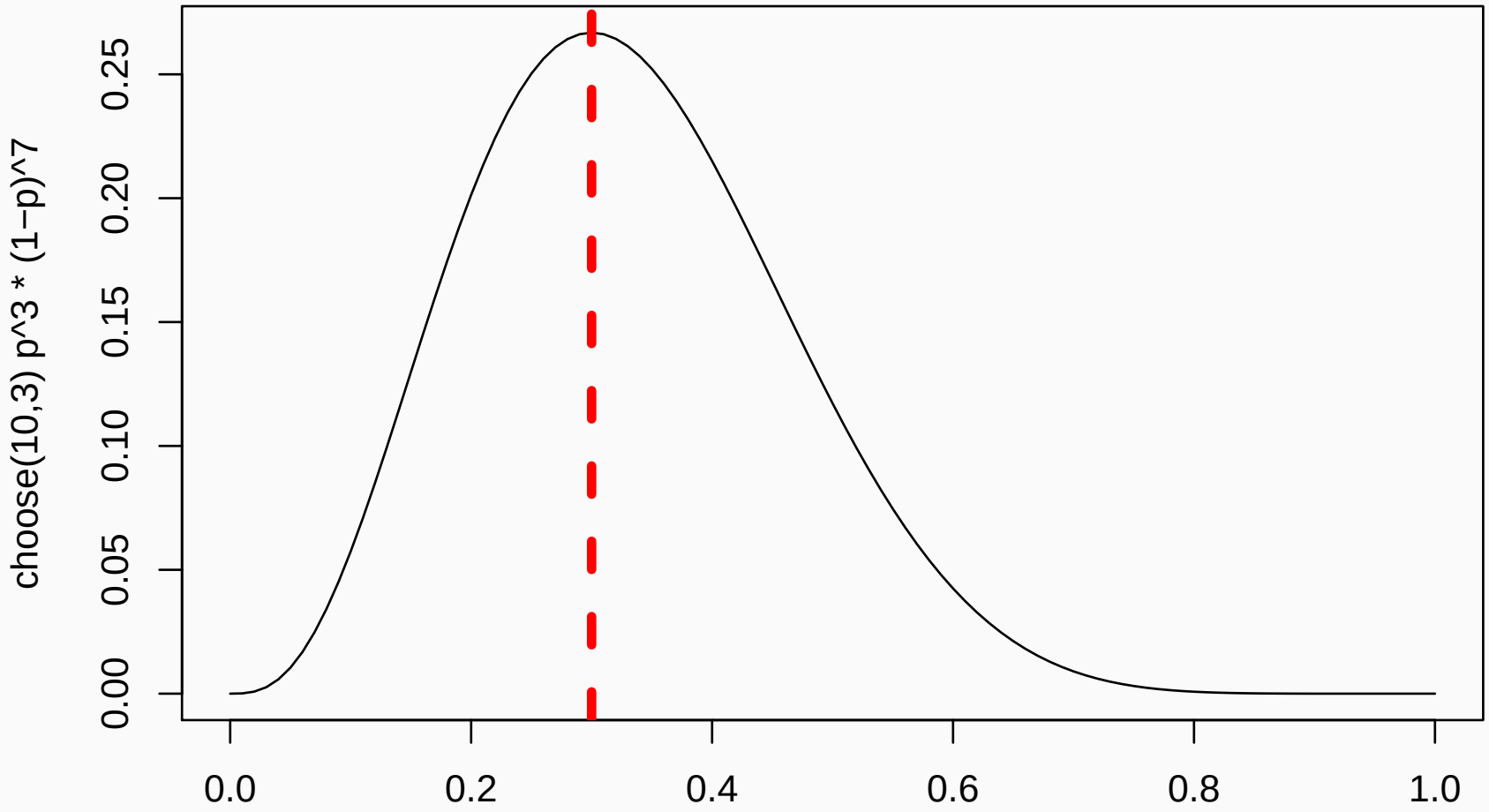
Example: Bernoulli MLE

```
bern <- function(x) x^3 * (1-x)^7
curve(bern(x), 0,1, ylab = "p^3 * (1-p)^7", xlab = "p")
abline(v = 0.3, lty = 2, col = "red", lwd = 4)
```



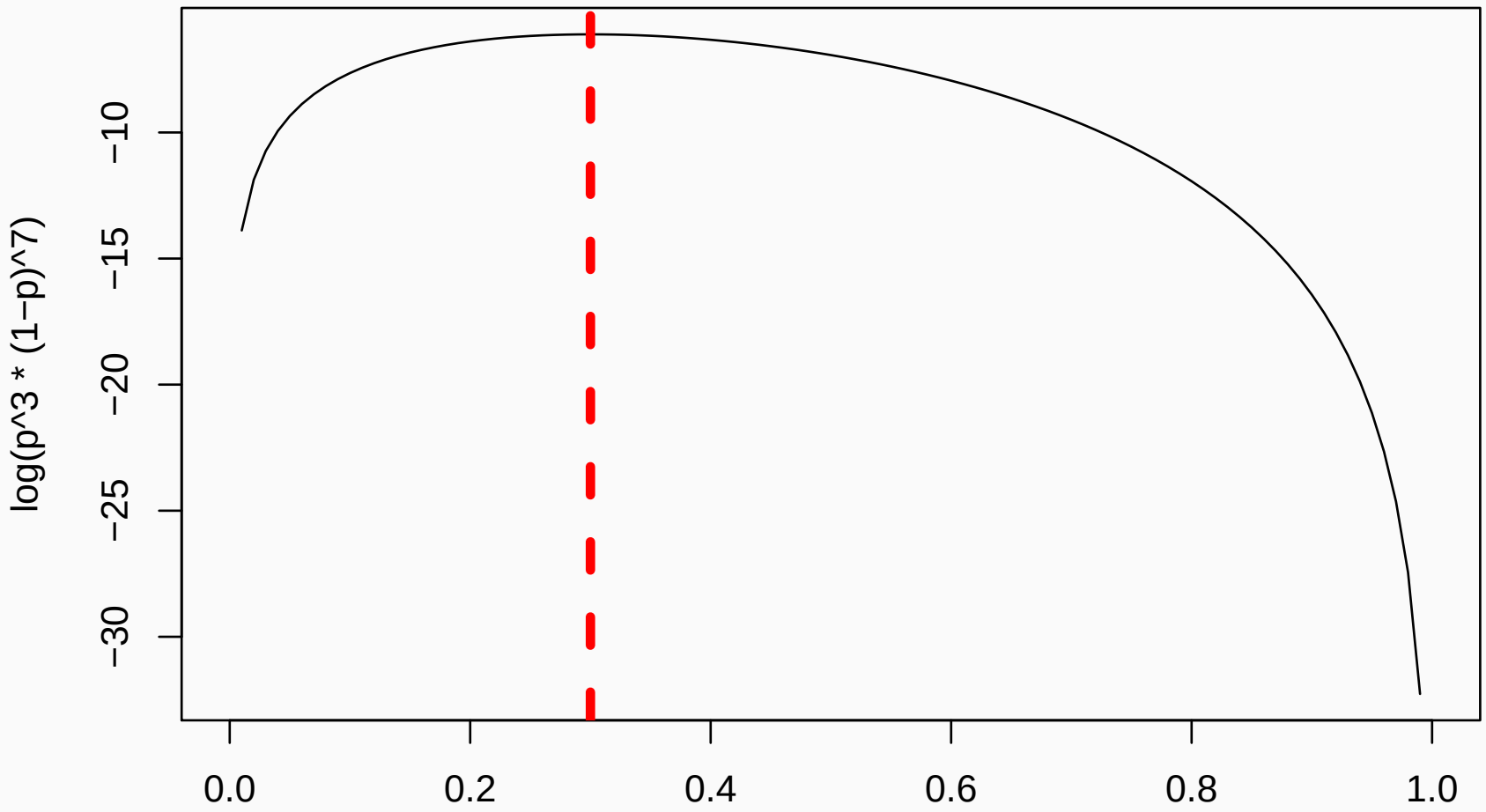
Example: Bernoulli MLE - Likelihood

```
binom <- function(x) choose(10,3) * x^3 * (1-x)^7
curve(binom(x), 0,1, ylab = "choose(10,3) p^3 * (1-p)^7", xlab = "p")
abline(v = 0.3, lty = 2, col = "red", lwd = 4)
```



Example: Bernoulli MLE - Log Likelihood

```
bern <- function(x) x^3 * (1-x)^7
curve(log(bern(x)), 0,1, ylab = "log(p^3 * (1-p)^7)", xlab = "p")
abline(v = 0.3, lty = 2, col = "red", lwd = 4)
```



Example: Normal MLE

A sample $y_1, ..., y_n$ is modelled by a Normal distribution with unknown parameters $\Theta \equiv (\mu, \sigma^2)$

$$f(y; \Theta) \equiv f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \qquad y \in \mathbb{R}$$

for some $\mu \in \mathbb{R}$ and $\sigma > 0$.

- Find the MLE of μ and σ^2.
- Solution The likelihood function becomes

$$L(\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

  Taking the natural log function, we obtain

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

  By taking the partial derivative with respect to μ and σ^2, setting to zero and solving them, we obtain $\hat{\mu} = \bar{Y}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

- Note that here $\hat{\sigma}^2$ is a biased estimator of σ^2.

Example: Normal MLE using `optim`

```r
ll.normal = function(theta, x) {
  # theta = (mu, sig2)
  # needs to be defined this way for optim to work
  mu = theta[1]
  sig2 = theta[2]
  n = length(x)
  (n / 2) * log(1 / sig2) - sum((x - mu) ^ 2) / (2 * sig2)
}

# generate N(2,16) data
data = rnorm(1000, mean=2, sd=4)
opt <- optim(par = c(0,1), fn = ll.normal, method = "BFGS",
             control = list(fnscale = -1), x = data)
opt$par
```

```
## [1]  1.972449 15.809416
```

```r
c(mean(data), sum((data - mean(data))^2)  / length(data) )
```

```
## [1]  1.972064 15.817615
```

Example: Uniform MLE

Let $Y_1, \ldots, Y_n$ be random sample from $U(0, \theta)$.

- Find the MLE of θ.
- Solution The likelihood function becomes

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 < y_i < \theta)$$

To maximize $L(\theta)$, first note that $\frac{1}{\theta^n}$ increases as θ decreases, and $\frac{1}{\theta^n}$ is maximized by selecting θ to be as small as possible, subject to the constraint that all of the y_i values are between zero and θ. The smallest value of θ that satisfies this constraint is the maximum observation in the set $y_1, y_2, \ldots, y_n$. That is, $\hat{\theta} = Y_{(n)}$.

- This MLE for θ is not an unbiased estimator of θ, but it can be adjusted to be unbiased.

What we have just learned

- Properties of Estimators
  - Efficiency
  - Consistency
  - Sufficiency
- The Rao–Blackwell Theorem and Minimum-Variance Unbiased Estimation
- The Method of Moments
- The Method of Maximum Likelihood