Estimation

Mwebesa Edson
(MBiostat, BSC STAT, BSC Edu-Math)
mwebesa.edson@gmail.com

OUTLINE

- POINT ESTIMATION (Properties and methods)
- INTERVAL ESTIMATION (For one mean, two means, variance, proportions)
- Sample size
- Simple linear regression
- More regression

Introduction

- In this section, we'll find good "**point estimates**" and "**confidence intervals**" for the usual population parameters, including:
- a population mean μ
- the difference in two population means, $\mu_1 - \mu_2$, say
- a population variance σ^2
- the ratio of two population variances, σ_1^2 / σ_2^2, say
- a population proportion p
- the difference in two population proportions, $p_1 - p_2$, say

- We will work on not only obtaining formulas for the estimates and intervals, but also on arguing that they are "good" in some way... unbiased, for example. We'll also address practical matters, such as how sample size affects the length of our derived confidence intervals. And, we'll also work on deriving good point estimates and confidence intervals for a least squares regression line through a set of (x,y) data points.

Point Estimation

- In this lesson, we'll learn two methods, namely the **method of maximum likelihood** and the **method of moments**, for deriving formulas for "good" point estimates for population parameters.

- We'll also learn one way of assessing whether a point estimate is "good." We'll do that by defining what a means for an estimate to be unbiased.

Objectives

- To learn how to find a maximum likelihood estimator of a population parameter.
- To learn how to find a method of moments estimator of a population parameter.
- To learn how to check to see if an estimator is unbiased for a particular parameter.
- To understand the steps involved in each of the proofs in the lesson.
- To be able to apply the methods learned in the lesson to new problems.

Definitions

- Let us denote the *n* random variables arising from a random sample as subscripted uppercase letters:

$$X_1, X_2, \ldots, X_n$$

- The corresponding observed values of a specific random sample are then denoted as subscripted lowercase letters:

$$x_1, x_2, \ldots, x_n$$

Definitions II

Definition. The range of possible values of the parameter θ is called the **parameter space** Ω (the greek letter "omega").

For example, if μ denotes the mean grade point average of all college students, then the parameter space (assuming a 4-point grading scale) is:

$$\Omega = \{\mu: 0 \leq \mu \leq 4\}$$

And, if p denotes the proportion of students who smoke cigarettes, then the parameter space is:

$$\Omega = \{p: 0 \leq p \leq 1\}$$

Definitions III

Definition. The function of $X_1, X_2, ..., X_n$, that is, the statistic $u(X_1, X_2, ..., X_n)$, used to estimate θ is called a **point estimator** of θ.

For example, the function:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a point estimator of the population mean μ. The function:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

(where $X_i = 0$ or 1) is a point estimator of the population proportion p. And, the function:

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

is a point estimator of the population variance σ^2.

Definitions IV

Definition. The function $u(x_1, x_2, ..., x_n)$ computed from a set of data is an observed **point estimate** of θ.

For example, if x_i are the observed grade point averages of a sample of 88 students, then:

$$\bar{x} = \frac{1}{88}\sum_{i=1}^{88} x_i = 3.12$$

is a point estimate of μ, the mean grade point average of all the students in the population.

And, if $x_i = 0$ if a student has no tattoo, and $x_i = 1$ if a student has a tattoo, then:

$$\hat{p} = 0.11$$

is a point estimate of p, the proportion of all students in the population who have a tattoo.

Maximum Likelihood Estimation

Definition. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ with probability density (or mass) function $f(x_i; \theta_1, \theta_2, \ldots, \theta_m)$. Suppose that $(\theta_1, \theta_2, \ldots, \theta_m)$ is restricted to a given parameter space Ω. Then:

(1) When regarded as a function of $\theta_1, \theta_2, \ldots, \theta_m$, the joint probability density (or mass) function of $X_1, X_2, \ldots, X_n$:

$$L(\theta_1, \theta_2, \ldots, \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_m)$$

$((\theta_1, \theta_2, \ldots, \theta_m)$ in $\Omega)$ is called the **likelihood function**.

(2) If:

$$[u_1(x_1, x_2, \ldots, x_n), u_2(x_1, x_2, \ldots, x_n), \ldots, u_m(x_1, x_2, \ldots, x_n)]$$

is the m-tuple that maximizes the likelihood function, then:

$$\hat{\theta}_i = u_i(X_1, X_2, \ldots, X_n)$$

is the **maximum likelihood estimator** of θ_i, for $i = 1, 2, \ldots, m$.

(3) The corresponding observed values of the statistics in (2), namely:

$$[u_1(x_1, x_2, \ldots, x_n), u_2(x_1, x_2, \ldots, x_n), \ldots, u_m(x_1, x_2, \ldots, x_n)]$$

are called the **maximum likelihood estimates** of θ_i, for $i = 1, 2, \ldots, m$.

NOTE

$$L(\theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1; \theta) \cdot f(x_2; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

- This definition of the terms
- (1) likelihood function,
- (2) maximum likelihood estimators, and
- (3) maximum likelihood estimates.

- So HOW DOES IT WORK?

EXAMPLES

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ where:

- $X_i = 0$ if a randomly selected student does not own a sports car, and
- $X_i = 1$ if a randomly selected student does own a sports car.

Assuming that the X_i are independent Bernoulli random variables with unknown parameter p, find the maximum likelihood estimator of p, the proportion of students who own a sports car.

Solution

Solution. If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$$

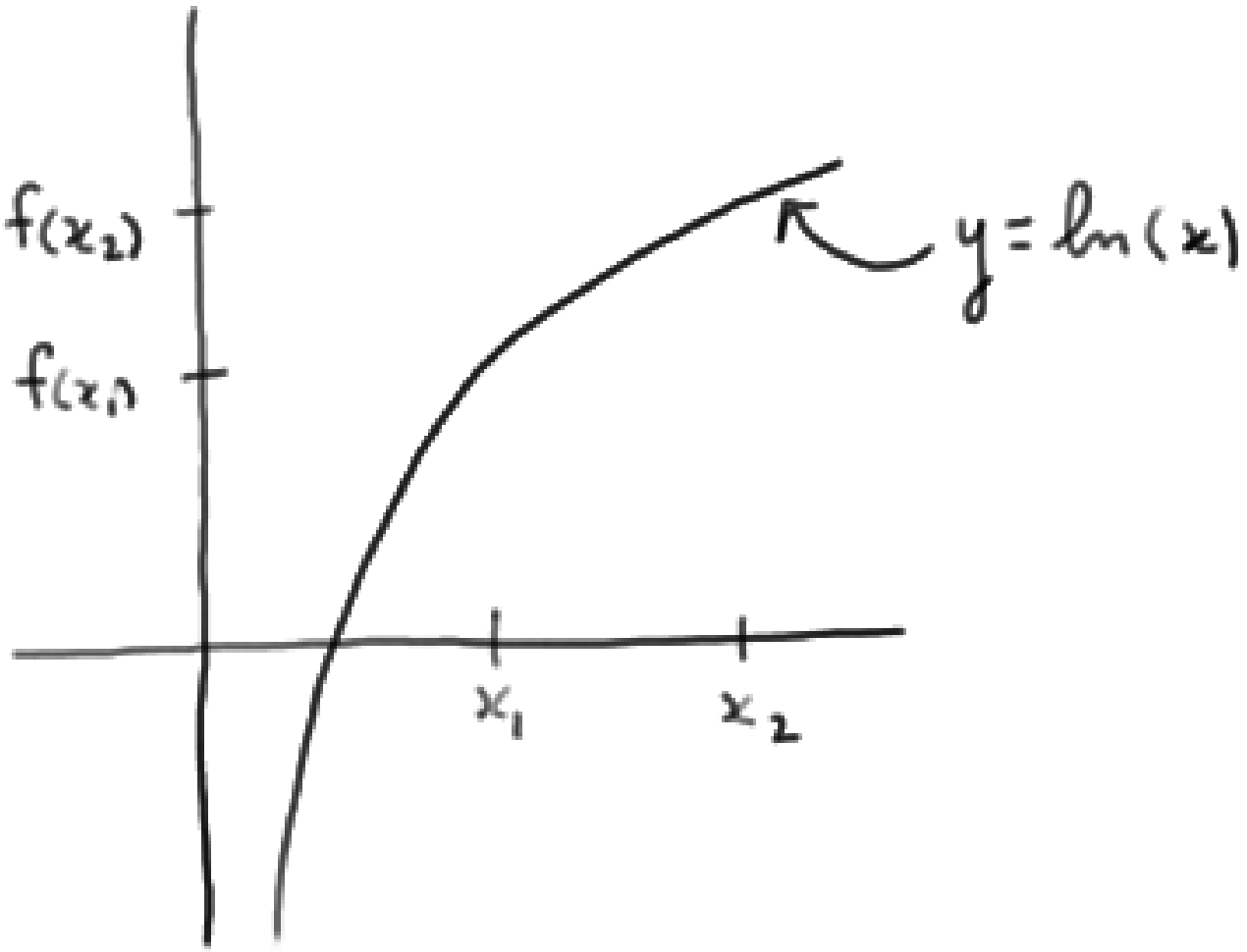
for $x_i = 0$ or 1 and $0 < p < 1$. Therefore, the likelihood function $L(p)$ is, by definition:

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

for $0 < p < 1$. Simplifying, by summing up the exponents, we get :

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

- Now, in order to implement the method of maximum likelihood, we need to find the p that maximizes the likelihood $L(p)$. We need to put on our calculus hats now, since in order to maximize the function, we are going to need to differentiate the likelihood function with respect to p. In doing so, we'll use a "trick" that often makes the differentiation a bit easier. Note that the natural logarithm is an increasing function of x:


f(x2)
f(x1)
y = ln(x)
x1
x2

That is, if $x_1 < x_2$, then $f(x_1) < f(x_2)$. That means that the value of p that maximizes the natural logarithm of the likelihood function $ln(L(p))$ is also the value of p that maximizes the likelihood function $L(p)$. So, the "trick" is to take the derivative of $ln(L(p))$ (with respect to p) rather than taking the derivative of $L(p)$. Again, doing so often makes the differentiation much easier. (By the way, throughout the remainder of this course, I will use either $ln(L(p))$ or $log(L(p))$ to denote the natural logarithm of the likelihood function.)

In this case, the natural logarithm of the likelihood function is:

$$\log L(p) = (\sum x_i)\log(p) + (n - \sum x_i)\log(1 - p)$$

Now, taking the derivative of the log likelihood, and setting to 0, we get:

$$\frac{\partial \log L(p)}{\partial p} = \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p} \overset{\text{SET}}{\equiv} 0$$

Now, multiplying through by $p(1-p)$, we get:

$$(\sum x_i)(1-p) - (n - \sum x_i)p = 0$$

Upon distributing, we see that two of the resulting terms cancel each other out:

$$\sum x_i - p\sum x_i - np + p\sum x_i = 0$$

leaving us with:

$$\sum x_i - np = 0$$

Now, all we have to do is solve for p. In doing so, you'll want to make sure that you always put a hat ("^") on the parameter, in this case p, to indicate it is an estimate:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

or, alternatively, an estimator:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Oh, and we should technically verify that we indeed did obtain a maximum. We can do that by verifying that the second derivative of the log likelihood with respect to p is negative. It is, but you might want to do the work to convince yourself!

Example II

- Suppose the weights of randomly selected American female college students are normally distributed with unknown mean μ and standard deviation σ. A random sample of 10 American female college students yielded the following weights (in pounds):

  115 122 130 127 149 160 152 138 149 180

- Based on the definitions given above, identify the likelihood function and the maximum likelihood estimator of μ, the mean weight of all American female college students. Using the given sample, find a maximum likelihood estimate of μ as well.

Solution. The probability density function of X_i is:

$$f(x_i;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

for $-\infty < x < \infty$. The parameter space is $\Omega = \{(\mu, \sigma)$: $-\infty < \mu < \infty$ and $0 < \sigma < \infty\}$. Therefore, (you might want to convince yourself that) the likelihood function is:

$$L(\mu,\sigma)=\sigma^{-n}(2\pi)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

for $-\infty < \mu < \infty$ and $0 < \sigma < \infty$. It can be shown (we'll do so in the next example!), upon maximizing the likelihood function with respect to μ, that the maximum likelihood estimator of μ is:

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}$$

Based on the given sample, a maximum likelihood estimate of μ is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(115 + \cdots + 180) = 142.2$$

pounds. Note that the only difference between the formulas for the maximum likelihood estimator and the maximum likelihood estimate is that:

- the estimator is defined using capital letters (to denote that its value is random), and
- the estimate is defined using lowercase letters (to denote that its value is fixed and based on an obtained sample)

Example III

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and variance σ^2. Find maximum likelihood estimators of mean μ and variance σ^2.

Solution. In finding the estimators, the first thing we'll do is write the probability density function as a function of $\theta_1 = \mu$ and $\theta_2 = \sigma^2$:

$$f(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{\theta_2}\sqrt{2\pi}} \exp\left[-\frac{(x_i - \theta_1)^2}{2\theta_2}\right]$$

for $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$. We do this so as not to cause confusion when taking the derivative of the likelihood with respect to σ^2. Now, that makes the likelihood function:

$$L(\theta_1, \theta_2) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2) = \theta_2^{-n/2}(2\pi)^{-n/2} \exp\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2\right]$$

and therefore the log of the likelihood function:

$$\log L(\theta_1, \theta_2) = -\frac{n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum(x_i - \theta_1)^2}{2\theta_2}$$

Now, upon taking the partial derivative of the log likelihood with respect to θ_1, and setting to 0, we see that a few things cancel each other out, leaving us with:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} = \frac{-2\sum(x_i - \theta_1)(-1)}{2\theta_2} \overset{\text{SET}}{\equiv} 0$$

Now, multiplying through by θ_2, and distributing the summation, we get:

$$\sum x_i - n\theta_1 = 0$$

Now, solving for θ_1, and putting on its hat, we have shown that the maximum likelihood estimate of θ_1 is:

$$\hat{\theta}_1 = \hat{\mu} = \frac{\sum x_i}{n} = x$$

Now for θ_2. Taking the partial derivative of the log likelihood with respect to θ_2, and setting to 0, we get:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{\Sigma(x_i - \theta_1)^2}{2\theta_2^2} \overset{\text{SET}}{\equiv} 0$$

Multiplying through by $2\theta_2^2$:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_2} = \left[-\frac{n}{2\theta_2} + \frac{\Sigma(x_i - \theta_1)^2}{2\theta_2^2} \overset{\text{SET}}{\equiv} 0 \right] \times 2\theta_2^2$$

we get:

$$-n\theta_2 + \sum(x_i - \theta_1)^2 = 0$$

And, solving for θ_2, and putting on its hat, we have shown that the maximum likelihood estimate of θ_2 is:

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{\sum(x_i - x)^2}{n}$$

(I'll again leave it to you to verify, in each case, that the second partial derivative of the log likelihood is negative, and therefore that we did indeed find maxima.) In summary, we have shown that the maximum likelihood estimators of μ and variance σ^2 for the normal model are:

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

respectively.

- Note that the maximum likelihood estimator of σ^2 for the normal model is not the sample variance S^2. They are, in fact, competing estimators. So how do we know which estimator we should use for σ^2 ? Well, one way is to choose the estimator that is "unbiased." Let's go learn about unbiased estimators now.

Unbiased Estimation

- In MLE, we showed that if X_i are Bernoulli random variables with parameter p, then:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the maximum likelihood estimator of p. And, if X_i are normally distributed random variables with mean μ and variance σ^2, then:

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

are the maximum likelihood estimators of μ and σ^2, respectively. A natural question then is whether or not these estimators are "good" in any sense. One measure of "good" is "unbiasedness."

Definition. If the following holds:

$$E[u(X_1, X_2, \ldots, X_n)] = \theta$$

then the statistic $u(X_1, X_2, \ldots, X_n)$ is an **unbiased estimator** of the parameter θ. Otherwise, $u(X_1, X_2, \ldots, X_n)$ is a **biased estimator** of θ.

Let $\hat{\Theta} = h(X_1, X_2, \cdots, X_n)$ be a point estimator for θ. The **bias** of point estimator $\hat{\Theta}$ is defined by

$$B(\hat{\Theta}) = E[\hat{\Theta}] - \theta.$$

Let $\hat{\Theta} = h(X_1, X_2, \cdots, X_n)$ be a point estimator for a parameter θ. We say that $\hat{\Theta}$ is an **unbiased** of estimator of θ if

$$B(\hat{\Theta}) = 0, \qquad \text{for all possible values of } \theta.$$

All Things Together

Definition: The bias of an estimator $\hat{\theta}$ of a parameter θ is the difference between the expected value of $\hat{\theta}$ and θ; that is, $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$. An estimator whose bias is identically equal to 0 is called unbiased estimator and satisfies $E(\hat{\theta}) = \theta$ for all θ.

Example

If X_i is a Bernoulli random variable with parameter p, then:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the maximum likelihood estimator (MLE) of p. Is the MLE of p an unbiased estimator of p?

Solution. Recall that if X_i is a Bernoulli random variable with parameter p, then $E(X_i) = p$.

Therefore:

$$E(\hat{p}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n}(np) = p$$

The first equality holds because we've merely replaced p-hat with its definition. The second equality holds by the rules of expectation for a linear combination. The third equality holds because $E(X_i) = p$. The fourth equality holds because when you add the value p up n times, you get np. And, of course, the last equality is simple algebra.

In summary, we have shown that:

$$E(\hat{p}) = p$$

Therefore, the maximum likelihood estimator is an unbiased estimator of p.

EXAMPLE II

If X_i are normally distributed random variables with mean μ and variance σ^2, then:

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

are the maximum likelihood estimators of μ and σ^2, respectively. Are the MLEs unbiased for their respective parameters?

Solution. Recall that if X_i is a normally distributed random variable with mean μ and variance σ^2, then $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Therefore:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1} \mu = \frac{1}{n}(n\mu) = \mu$$

The first equality holds because we've merely replaced X-bar with its definition. Again, the second equality holds by the rules of expectation for a linear combination. The third equality holds because $E(X_i) = \mu$. The fourth equality holds because when you add the value μ up n times, you get $n\mu$. And, of course, the last equality is simple algebra.

In summary, we have shown that:

$$E(\bar{X}) = \mu$$

Therefore, the maximum likelihood estimator of μ is unbiased. Now, let's check the maximum likelihood estimator of σ^2. First, note that we can rewrite the formula for the MLE as:

$$\hat{\sigma}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \bar{X}^2$$

because:

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

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\cdot\left(\frac{1}{n}\sum x_i\right) + \frac{1}{n}(n\bar{x}^2)$$

$$\frac{1}{n}\sum x_i \rightarrow \bar{x}$$

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

Then, taking the expectation of the MLE, we get:

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

as illustrated here:

$$\begin{aligned} E(\hat{\sigma}^2) &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - X^2\right] = \left[\frac{1}{n}\sum_{i=1}^{n} E(X_i^2)\right] - E(X^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right) \\ &= \frac{1}{n}(n\sigma^2+n\mu^2) - \frac{\sigma^2}{n} - \mu^2 \\ &= \sigma^2 - \frac{\sigma^2}{n} = \frac{n\sigma^2-\sigma^2}{n} = \frac{(n-1)\sigma^2}{n} \end{aligned}$$

The first equality holds from the rewritten form of the MLE. The second equality holds from the properties of expectation. The third equality holds from manipulating the alternative formulas for the variance, namely:

$$Var(X) = \sigma^2 = E(X^2) - \mu^2 \text{ and } \quad Var(\bar{X}) = \frac{\sigma^2}{n} = E(\bar{X}^2) - \mu^2$$

The remaining equalities hold from simple algebraic manipulation. Now, because we have shown:

$$E(\hat{\sigma}^2) \neq \sigma^2$$

the maximum likelihood estimator of σ^2 is a biased estimator.

EXAMPLE III

If X_i are normally distributed random variables with mean μ and variance σ^2, what is an unbiased estimator of σ^2? Is S^2 unbiased?

Solution. Recall that if X_i is a normally distributed random variable with mean μ and variance σ^2, then:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Also, recall that the expected value of a chi-square random variable is its degrees of freedom. That is, if:

$$X \sim \chi^2_{(r)}$$

then $E(X) = r$. Therefore:

$$E(S^2) = E\left[\frac{\sigma^2}{n-1} \cdot \frac{(n-1)S^2}{\sigma^2}\right] = \frac{\sigma^2}{n-1} E\left[\frac{(n-1)S^2}{\sigma^2}\right] = \frac{\sigma^2}{n-1} \cdot (n-1) = \sigma^2$$

The first equality holds because we effectively multiplied the sample variance by 1. The second equality holds by the law of expectation that tells us we can pull a constant through the expectation. The third equality holds because of the two facts we recalled above. That is:

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1$$

And, the last equality is again simple algebra.

- In summary, we have shown that, if X_i is a normally distributed random variable with mean μ and variance σ^2, then S^2 is an unbiased estimator of σ^2. It turns out, however, that S^2 is *always* an unbiased estimator of σ^2, that is, for *any* model, not just the normal model. (Show that as a homework.) And, although S^2 is always an unbiased estimator of σ^2, S is *not* an unbiased estimator of σ. (Show that as a homework, too.)

Another Example

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample. Show that the sample mean

$$\hat{\Theta} = \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

is an unbiased estimator of $\theta = EX_i$.

Solution

We have

$$\begin{aligned} B(\hat{\Theta}) &= E[\hat{\Theta}] - \theta \\ &= E\left[\overline{X}\right] - \theta \\ &= EX_i - \theta \\ &= 0. \end{aligned}$$

Definition

The **mean squared error** (MSE) of a point estimator $\hat{\Theta}$, shown by $MSE(\hat{\Theta})$, is defined as

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

Note that $\hat{\Theta} - \theta$ is the error that we make when we estimate θ by $\hat{\Theta}$. Thus, the MSE is a measure of the distance between $\hat{\Theta}$ and θ, and a smaller MSE is generally indicative of a better estimator.

Proof of MSE($\widehat{\boldsymbol{\theta}}$) = $\boldsymbol{Var}\left(\widehat{\boldsymbol{\theta}}\right) + [\boldsymbol{B}(\widehat{\boldsymbol{\theta}})]^2$

$$MSE_{\hat{\theta}} = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 = Var(\hat{\theta}) + (Bias\ of\ \hat{\theta})^2$$

This is so because

$$\begin{aligned} E(\hat{\theta} - \theta)^2 &= E(\hat{\theta}^2) + E(\theta^2) - 2\theta E(\hat{\theta}) \\ &= Var(\hat{\theta}) + [E(\hat{\theta})]^2 + \theta^2 - 2\theta E(\hat{\theta}) \\ &= Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \end{aligned}$$

RECALL, $\boldsymbol{B}\left(\widehat{\boldsymbol{\theta}}\right) = \boldsymbol{E}\left(\widehat{\boldsymbol{\theta}}\right) - \boldsymbol{\theta}$

For an unbiased estimator $\hat{\theta}$, we have

$$MSE_{\hat{\theta}} = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta})$$

and so, if an estimator is unbiased, its MSE is equal to its variance.

EXAMPLE

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from a distribution with mean $EX_i = \theta$, and variance $\mathrm{Var}(X_i) = \sigma^2$. Consider the following two estimators for θ:

1. $\hat{\Theta}_1 = X_1$.
2. $\hat{\Theta}_2 = \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$.

Find $MSE(\hat{\Theta}_1)$ and $MSE(\hat{\Theta}_2)$ and show that for $n > 1$, we have

$$MSE(\hat{\Theta}_1) > MSE(\hat{\Theta}_2).$$

Solution

We have

$$\begin{aligned} MSE(\hat{\Theta}_1) &= E\big[(\hat{\Theta}_1 - \theta)^2\big] \\ &= E[(X_1 - EX_1)^2] \\ &= \mathrm{Var}(X_1) \\ &= \sigma^2. \end{aligned}$$

To find $MSE(\hat{\Theta}_2)$, we can write

$$\begin{aligned} MSE(\hat{\Theta}_2) &= E\big[(\hat{\Theta}_2 - \theta)^2\big] \\ &= E[(\overline{X} - \theta)^2] \\ &= \mathrm{Var}(\overline{X} - \theta) + \big(E[\overline{X} - \theta]\big)^2. \end{aligned}$$

The last equality results from $EY^2 = \mathrm{Var}(Y) + (EY)^2$, where $Y = \overline{X} - \theta$. Now, note that

$$\mathrm{Var}(\overline{X} - \theta) = \mathrm{Var}(\overline{X})$$

since θ is a constant. Also, $E[\overline{X} - \theta] = 0$. Thus, we conclude

$$\begin{aligned} MSE(\hat{\Theta}_2) &= \mathrm{Var}(\overline{X}) \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Thus, we conclude for $n > 1$,

$$MSE(\hat{\Theta}_1) > MSE(\hat{\Theta}_2).$$

Conclusion on MLE and Unbiasedness

- Sometimes it is impossible to find maximum likelihood estimators in a convenient closed form. Instead, numerical methods must be used to maximize the likelihood function.

- In such cases, we might consider using an alternative method of finding estimators, such as the "method of moments." Let's go take a look at that method now.

Method of Moments

- In short, the method of moments involves equating sample moments with theoretical moments.

- So, let's start by making sure we recall the definitions of theoretical moments, as well as learn the definitions of sample moments.

Definitions.

(1) $E(X^k)$ is the k^{th} **(theoretical) moment** of the distribution (**about the origin**), for $k = 1, 2, \ldots$

(2) $E\left[(X-\mu)^k\right]$ is the k^{th} **(theoretical) moment** of the distribution (**about the mean**), for $k = 1, 2, \ldots$

(3) $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ is the k^{th} **sample moment**, for $k = 1, 2, \ldots$

(4) $M_k^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$ is the k^{th} **sample moment about the mean**, for $k = 1, 2, \ldots$

One Form of the Method

- The basic idea behind this form of the method is to:

(1) Equate the first sample moment about the origin $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ to the first theoretical moment $E(X)$.

(2) Equate the second sample moment about the origin $M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ to the second theoretical moment $E(X^2)$.

(3) Continue equating sample moments about the origin, M_k, with the corresponding theoretical moments $E(X^k)$, $k = 3, 4, \ldots$ until you have as many equations as you have parameters.

(4) Solve for the parameters.

- The resulting values are called **method of moments estimators**. It seems reasonable that this method would provide good estimates, since the empirical distribution converges in some sense to the probability distribution.
- Therefore, the corresponding moments should be about equal.
- In some cases, rather than using the sample moments about the origin, it is easier to use the sample moments about the mean. Doing so, provides us with an alternative form of the method of moments.

Another Form of the Method

- The basic idea behind this form of the method is to:

(1) Equate the first sample moment about the origin $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ to the first theoretical moment $E(X)$.

(2) Equate the second sample moment about the mean $M_2^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ to the second theoretical moment about the mean $E[(X-\mu)^2]$.

(3) Continue equating sample moments about the mean M_k^* with the corresponding theoretical moments about the mean $E[(X-\mu)^k]$, $k = 3, 4, \ldots$ until you have as many equations as you have parameters.

(4) Solve for the parameters.

EXAMPLE

Let $X_1, X_2, ..., X_n$ be gamma random variables with parameters α and θ, so that the probability density function is:

$$f(x_i) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta}$$

for $x > 0$. Therefore, the likelihood function:

$$L(\alpha, \theta) = \left(\frac{1}{\Gamma(\alpha)\theta^{\alpha}}\right)^{n} (x_1 x_2 \cdots x_n)^{\alpha-1} \exp\left[-\frac{1}{\theta}\sum x_i\right]$$

is difficult to differentiate because of the gamma function $\Gamma(\alpha)$. So, rather than finding the maximum likelihood estimators, what are the method of moments estimators of α and θ?

Solution. The first theoretical moment about the origin is:

$$E(X_i) = \alpha\theta$$

And the second theoretical moment about the mean is:

$$Var(X_i) = E(X_i - \mu)^2 = \alpha\theta^2$$

Again, since we have two parameters for which we are trying to derive method of moments estimators, we need two equations. Equating the first theoretical moment about the origin with the corresponding sample moment, we get:

$$E(X) = \alpha\theta = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

And, equating the second theoretical moment about the mean with the corresponding sample moment, we get:

$$Var(X) = \alpha\theta^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Now, we just have to solve for the two parameters α and θ. Let's start by solving for α in the first equation ($E(X)$). Doing so, we get:

$$\alpha = \frac{\bar{X}}{\theta}$$

Now, substituting $\alpha = \frac{\bar{X}}{\theta}$ into the second equation ($Var(X)$), we get:

$$\alpha\theta^2 = \left(\frac{\bar{X}}{\theta}\right)\theta^2 = \bar{X}\theta = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Now, solving for θ in that last equation, and putting on its hat, we get that the method of moment estimator for θ is:

$$\hat{\theta}_{MM} = \frac{1}{n\bar{X}} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

And, substituting that value of θ back into the equation we have for α, and putting on its hat, we get that the method of moment estimator for α is:

$$\hat{\alpha}_{MM} = \frac{\bar{X}}{\hat{\theta}_{MM}} = \frac{\bar{X}}{(1/n\bar{X}) \sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{n\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

The Gamma distribution

The gamma function is a real-valued non-negative function defined on $(0, \infty)$ in the following manner

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx \quad , \quad \alpha > 0 \, .$$

The Gamma function enjoys some nice properties. Two of these are listed below:

$$(a)\ \Gamma(\alpha + 1) = \alpha\, \Gamma(\alpha) \quad , \quad (b)\ \Gamma(n) = (n-1)! \ \ (n \ \text{ integer}) \, .$$

Property (b) is an easy consequence of Property (a). Start off with $\Gamma(n)$ and use Property (a) recursively along with the fact that $\Gamma(1) = 1$ (why?). Another important fact is that $\Gamma(1/2) = \sqrt{\pi}$ (Prove this at home!).

Definition

A random variable Y is said to have a *gamma distribution with parameters* $\alpha > 0$ and $\beta > 0$ if and only if the density function of Y is

$$f(y) = \begin{cases} \dfrac{y^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, & 0 \leq y < \infty, \\ 0, & \text{elsewhere,} \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}\,dy.$$

The quantity $\Gamma(\alpha)$ is known as the *gamma function*. Direct integration will verify that $\Gamma(1) = 1$. Integration by parts will verify that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ for any $\alpha > 1$ and that $\Gamma(n) = (n - 1)!$, provided that n is an integer.

Beta Distribution

- Definition

A random variable Y is said to have a *beta probability distribution with parameters* $\alpha > 0$ *and* $\beta > 0$ if and only if the density function of Y is

$$f(y) = \begin{cases} \dfrac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}, & 0 \le y \le 1, \\ 0, & \text{elsewhere,} \end{cases}$$

where

$$B(\alpha,\beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

If Y is a beta-distributed random variable with parameters $\alpha > 0$ and $\beta > 0$, then

$$\mu = E(Y) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \sigma^2 = V(Y) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

By definition,

$$
\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} y f(y)\, dy \\
&= \int_{0}^{1} y \left[\frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} \right] dy \\
&= \frac{1}{B(\alpha,\beta)} \int_{0}^{1} y^{\alpha}(1-y)^{\beta-1}\, dy \\
&= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \qquad \text{(because } \alpha > 0 \text{ implies that } \alpha+1 > 0\text{)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha}{(\alpha+\beta)}.
\end{aligned}
$$

If X follows the Gamma(α, β) distribution, the mean and variance of X can be explicitly expressed in terms of the parameters:

$$\mathbb{E}(X) = \alpha\beta \quad \text{and} \quad \text{Var}(X) = \alpha\beta^2 \,.$$

We outline the computation of a general moment $\mathbb{E}(X^k)$, where k is a positive integer.

We have,

$$\begin{aligned}
\mathbb{E}(X^k) &= \int_0^\infty x^k \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1} dx \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty e^{-x/\beta} x^{k+\alpha-1} dx \\
&= \frac{\beta^{\alpha+k}}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\beta^\alpha} \\
&= \frac{(\alpha+k-1)\cdots(\alpha)\,\Gamma(\alpha)}{\Gamma(\alpha)} \\
&= \beta^k \Pi_{i=1}^k (\alpha+i-1).
\end{aligned}$$

The formulae for the mean and the variance should follow directly from the above computation. Note that in the above derivation, we have used the fact that

$$\int_0^\infty e^{-x/\beta} x^{k+\alpha-1} dx = \Gamma(\alpha+k)\beta^{\alpha+k}.$$

This is an immediate consequence of the fact that the gamma density with parameters $(\alpha+k, \beta)$ integrates to 1.

- Assignment:
- Derive the mean and variance of Gamma Distribution by Integration
- Show how Chi-Square and Exponential Distributions are derived from Gamma Distribution. Determine their respective means and variances

OTHER PROPERTIES OF ESTIMATORS

- We have so far looked at UNBIASEDNESS
- Others Include

1. Sufficiency
2. Efficiency
3. Consistency

Sufficient Statistics

Objectives

In this Section, our goals are:

- To learn a formal definition of sufficiency.
- To learn how to apply the Factorization Theorem to identify a sufficient statistic.
- To learn how to apply the Exponential Criterion to identify a sufficient statistic.
- To extend the definition of sufficiency for one parameter to two (or more) parameters.

Definition of Sufficiency

Definition. Let $X_1, X_2, ..., X_n$ be a random sample from a probability distribution with unknown parameter θ. Then, the statistic:

$$Y = u(X_1, X_2, \ldots, X_n)$$

is said to be **sufficient** for θ if the conditional distribution of $X_1, X_2, ..., X_n$, given the statistic Y, does not depend on the parameter θ.

Factorization Theorem

- While the definition of sufficiency provided on the previous page may make sense intuitively, it is not always all that easy to find the conditional distribution of $X1, X2, \ldots, Xn$ given Y.
- Not to mention that we'd have to find the conditional distribution of $X1, X2, \ldots, Xn$ given Y for every Y that we'd want to consider a possible sufficient statistic!
- Therefore, using the formal definition of sufficiency as a way of identifying a sufficient statistic for a parameter θ can often be a daunting road to follow. Thankfully, a theorem often referred to as the Factorization Theorem provides an easier alternative

Factorization Theorem. Let $X_1, X_2, ..., X_n$ denote random variables with joint probability density function or joint probability mass function $f(x_1, x_2, ..., x_n; \theta)$, which depends on the parameter θ. Then, the statistic $Y = u(X_1, X_2, \ldots, X_n)$ is sufficient for θ if and only if the p.d.f (or p.m.f.) can be factored into two components, that is:

$$f(x_1, x_2, \ldots, x_n; \theta) = \phi[u(x_1, x_2, \ldots, x_n); \theta] h(x_1, x_2, \ldots, x_n)$$

where:

- φ is a function that depends on the data $x_1, x_2, ..., x_n$ only through the function $u(x_1, x_2, ..., x_n)$, and
- the function $h(x_1, x_2, ..., x_n)$ does not depend on the parameter θ

Example

Let $X_1, X_2, ..., X_n$ denote a random sample from a Poisson distribution with parameter $\lambda > 0$. Find a sufficient statistic for the parameter λ.

Solution. Because $X_1, X_2, ..., X_n$ is a random sample, the joint probability mass function of $X_1, X_2, ..., X_n$ is, by independence:

$$f(x_1, x_2, \ldots, x_n; \lambda) = f(x_1; \lambda) \times f(x_2; \lambda) \times \ldots \times f(x_n; \lambda)$$

Inserting what we know to be the probability mass function of a Poisson random variable with parameter λ, the joint p.m.f. is therefore:

$$f(x_1, x_2, \ldots, x_n; \lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \times \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \times \ldots \times \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

Now, simpliyfing, by adding up all n of the λs in the exponents, as well as all n of the x_i's in the exponents, we get:

$$f(x_1, x_2, \ldots, x_n; \lambda) = \left(e^{-n\lambda}\lambda^{\Sigma x_i}\right) \times \left(\frac{1}{x_1!x_2!\ldots x_n!}\right)$$

Hey, look at that! We just factored the joint p.m.f. into two functions, one (φ) being only a function of the statistic $Y = \sum_{i=1}^{n} X_i$ and the other (h) not depending on the parameter λ:

$$f(x_1, x_2, \ldots, x_n; \lambda) = \left(e^{-n\lambda}\lambda^{\Sigma x_i}\right) \times \left(\frac{1}{x_1!x_2!\cdots x_n!}\right)$$

$\phi[u(\Sigma x_i); \lambda]$ $\quad h(x_1, x_2, \ldots, x_n)$

Therefore, the Factorization Theorem tells us that $Y = \sum_{i=1}^{n} X_i$ is a sufficient statistic for λ. But, wait a second! We can also write the joint p.m.f. as:

$$f(x_1, x_2, \ldots, x_n; \lambda) = \left(e^{-n\lambda}\lambda^{n\bar{x}}\right) \times \left(\frac{1}{x_1!x_2!\ldots x_n!}\right)$$

Therefore, the Factorization Theorem tells us that $Y = \bar{X}$ is also a sufficient statistic for λ!

If you think about it, it makes sense that $Y = \bar{X}$ and $Y = \sum_{i=1}^{n} X_i$ are both sufficient statistics, because if we know $Y = \bar{X}$, we can easily find $Y = \sum_{i=1}^{n} X_i$. And, if we know $Y = \sum_{i=1}^{n} X_i$, we can easily find $Y = \bar{X}$.

Example II

Let $X_1, X_2, ..., X_n$ be a random sample from a normal distribution with mean μ and variance 1. Find a sufficient statistic for the parameter μ.

Solution. Because $X_1, X_2, ..., X_n$ is a random sample, the joint probability density function of $X_1, X_2, ..., X_n$ is, by independence:

$$f(x_1, x_2, \ldots, x_n; \mu) = f(x_1; \mu) \times f(x_2; \mu) \times \ldots \times f(x_n; \mu)$$

Inserting what we know to be the probability density function of a normal random variable with mean μ and variance 1, the joint p.d.f. is:

$$f(x_1, x_2, \ldots, x_n; \mu) = \frac{1}{(2\pi)^{1/2}} exp\left[-\frac{1}{2}(x_1 - \mu)^2\right] \times \frac{1}{(2\pi)^{1/2}} exp\left[-\frac{1}{2}(x_2 - \mu)^2\right] \times \ldots \times \frac{1}{(2\pi)^{1/2}} exp\left[-\frac{1}{2}(x_n - \mu)^2\right]$$

Collecting like terms, we get:

$$f(x_1, x_2, \ldots, x_n; \mu) = \frac{1}{(2\pi)^{n/2}} exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

A trick to making the factoring of the joint p.d.f. an easier task is to add 0 to the quantity in parentheses in the summation. That is:

$$f(x_1,x_2,\ldots,x_n;\mu)=\frac{1}{(2\pi)^{n/2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2\right]$$

$\bar{x}-\bar{x}\to 0$

Now, squaring the quantity in parentheses, we get:

$$f(x_1,x_2,\ldots,x_n;\mu)=\frac{1}{(2\pi)^{n/2}}exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\bar{x})^2+2(x_i-\bar{x})(\bar{x}-\mu)+(\bar{x}-\mu)^2\right]\right]$$

And then distributing the summation, we get:

$$f(x_1,x_2,\ldots,x_n;\mu)=\frac{1}{(2\pi)^{n/2}}exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2-(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x})-\frac{1}{2}\sum_{i=1}^{n}(\bar{x}-\mu)^2\right]$$

But, the middle term in the exponent is 0, and the last term, because it doesn't depend on the index i, can be added up n times:

$$f(x_1,x_2,\ldots,x_n;\mu)=\frac{1}{(2\pi)^{n/2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2-(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x})-\frac{1}{2}\sum_{i=1}^{n}(\bar{x}-\mu)^2\right]$$

$\sum_{i=1}^{n}(x_i-\bar{x})\to 0$

$\sum_{i=1}^{n}(\bar{x}-\mu)^2\to n(\bar{x}-\mu)^2$

So, simplifying, we get:

$$f(x_1, x_2, \ldots, x_n; \mu) = \left\{ exp\left[-\frac{n}{2}(\bar{x} - \mu)^2 \right] \right\} \times \left\{ \frac{1}{(2\pi)^{n/2}} exp\left[-\frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \right\}$$

In summary, we have factored the joint p.d.f. into two functions, one (ϕ) being only a function of the statistic $Y = \bar{X}$ and the other (h) not depending on the parameter μ:

$$f(x_1, x_2, \ldots, x_n; \mu) = \left\{ \exp\left[-\frac{n}{2}(\bar{x} - \mu)^2 \right] \right\} \times \left\{ \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \right\}$$

$\phi[u(\bar{x}); \mu]$ $\qquad$ $h(x_1, x_2, \ldots, x_n)$

Example

Let $X_1, X_2, ..., X_n$ be a random sample from an exponential distribution with parameter θ. Find a sufficient statistic for the parameter θ.

Solution. Because $X_1, X_2, ..., X_n$ is a random sample, the joint probability density function of $X_1, X_2, ..., X_n$ is, by independence:

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) \times f(x_2; \theta) \times \ldots \times f(x_n; \theta)$$

Inserting what we know to be the probability density function of an exponential random variable with parameter θ, the joint p.d.f. is:

$$f(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{\theta} exp\left(\frac{-x_1}{\theta}\right) \times \frac{1}{\theta} exp\left(\frac{-x_2}{\theta}\right) \times \ldots \times \frac{1}{\theta} exp\left(\frac{-x_n}{\theta}\right)$$

Now, simpliyfing, by adding up all n of the θs and the n x_i's in the exponents, we get:

$$f(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{\theta^n} exp\left(-\frac{1}{\theta} \sum_{i=1}^{n} x_i\right)$$

We have again factored the joint p.d.f. into two functions, one (φ) being only a function of the statistic $Y = \sum_{i=1}^{n} X_i$ and the other (h) not depending on the parameter θ:

$$f(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta} \sum_{i=1}^{n} x_i\right) \times 1$$

$\phi[u(x_i); \theta]$ $\quad h(x_1, x_2, \ldots, x_n)$

Therefore, the Factorization Theorem tells us that $Y = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ. And, since $Y = \bar{X}$ is a one-to-one function of $Y = \sum_{i=1}^{n} X_i$, it implies that $Y = \bar{X}$ is also a sufficient statistic for θ.

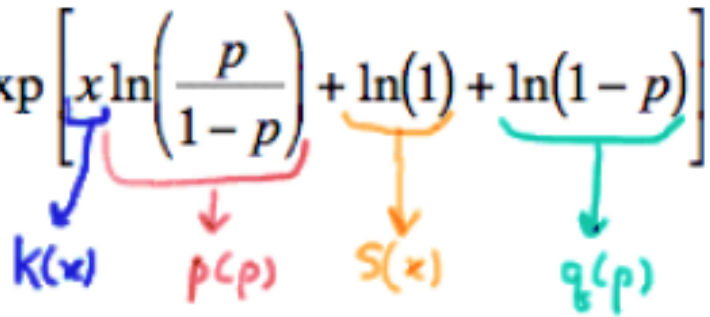
Exponential Form

You might not have noticed that in all of the examples we have considered so far in this lesson, every p.d.f. or p.m.f. could we written in what is often called **exponential form**, that is:

$$f(x;\theta) = exp\left[K(x)p(\theta) + S(x) + q(\theta)\right]$$

with (1) $K(x)$ and $S(x)$ being functions only of x, (2) $p(\theta)$ and $q(\theta)$ being functions only of the parameter θ, and (3) the support being free of the parameter θ. First, we had Bernoulli random variables with p.m.f. written in exponential form as:

$$f(x;p) = p^x(1-p)^{1-x} = \exp\left[x\ln\left(\frac{p}{1-p}\right) + \ln(1) + \ln(1-p)\right]$$



with (1) $K(x)$ and $S(x)$ being functions only of x, (2) $p(p)$ and $q(p)$ being functions only of the parameter p, and (3) the support $x = 0, 1$ not depending on the parameter p. Okay, we just skipped a lot of steps in that second equality sign, that is, in getting from point A (the typical p.m.f.) to point B (the p.m.f. written in exponential form). So, let's take a look at that more closely. We start with:

$$f(x;p) = p^x(1-p)^{1-x}$$

Is the p.m.f. in exponential form? Doesn't look like it to me! We clearly need an "exp" to appear up front. The only way we are going to get that without changing the underlying function is by taking the inverse function, that is, the natural log ("ln"), at the same time. Doing so, we get:

$$f(x;p) = exp\left[\ln(p^x(1-p)^{1-x})\right]$$

Is the p.m.f. now in exponential form? Nope, not yet, but at least it's looking more hopeful. All of the steps that follow now involve using what we know about the properties of logarithms. Recognizing that the natural log of a product is the sum of the natural logs, we get:

$$f(x;p) = exp\left[\ln(p^x) + \ln(1-p)^{1-x}\right]$$

Is the p.m.f. now in exponential form? Nope, still not yet, because $K(x)$, $p(p)$, $S(x)$, and $q(p)$ can't yet be identified as following exponential form, but we are certainly getting closer. Recognizing that the log of a power is the power times the log of the base, we get:

$$f(x;p) = exp\left[x\ln(p) + (1-x)\ln(1-p)\right]$$

This is getting tiring. Is the p.m.f. in exponential form yet? Nope, afraid not yet. Let's distribute that $(1-x)$ in that last term. Doing so, we get:

$$f(x;p) = exp\left[x\ln(p) + \ln(1-p) - x\ln(1-p)\right]$$

Is the p.m.f. now in exponential form? Let's take a closer look. Well, in the first term, we can identify the $K(x)p(p)$ and in the middle term, we see a function that depends only on the parameter p:

$$f(x;p) = \exp\left[x\ln(p) + \ln(1-p) - x\ln(1-p)\right]$$

Now, all we need is the last term to depend only on x and we're as good as gold. Oh, rats! The last term depends on both x and p. So back to work some more! Recognizing that the log of a quotient is the difference between the logs of the numerator and denominator, we get:

$$f(x;p) = exp\left[x\ln\left(\frac{p}{1-p}\right) + \ln(1-p)\right]$$

Is the p.m.f. now in exponential form? So close! Let's just add 0 in (by way of the natural log of 1) to make it obvious. Doing so, we get:

$$f(x;p) = exp\left[x\ln\left(\frac{p}{1-p}\right) + \ln(1) + \ln(1-p)\right]$$

Yes, we have finally written the Bernoulli p.m.f. in exponential form:

$$f(x;p) = \exp\left[x\ln\left(\frac{p}{1-p}\right) + \ln(1) + \ln(1-p)\right]$$

$K(x)$ $p(p)$ $S(x)$ $q(p)$

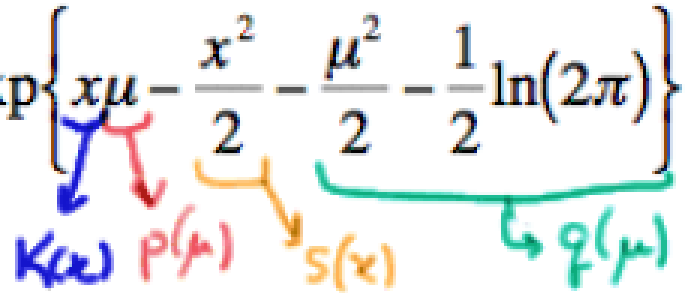
Whew! So, we've fully explored writing the Bernoulli p.m.f. in exponential form! Let's get back to reviewing all of the p.m.f.'s we've encountered in this lesson. We had Poisson random variables whose p.m.f. can be written in exponential form as:

$$f(x;\lambda)=\frac{e^{-\lambda}\lambda^x}{x!}=\exp[x\ln\lambda-\ln x!-\lambda]$$

$K(x)$ $p(\lambda)$ $S(x)$ $q(\lambda)$

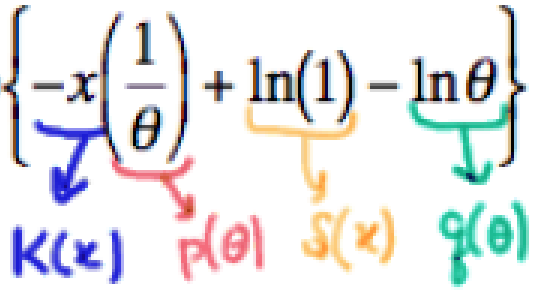
with (1) $K(x)$ and $S(x)$ being functions only of x, (2) $p(\lambda)$ and $q(\lambda)$ being functions only of the parameter λ, and (3) the support $x = 0, 1, 2, \ldots$ not depending on the parameter λ. Then, we had $N(\mu, 1)$ random variables whose p.d.f. can be written in exponential form as:

$$f(x;\mu)=\frac{1}{\sqrt{2\pi}}e^{-(x-\mu)^2/2}=\exp\left\{x\mu-\frac{x^2}{2}-\frac{\mu^2}{2}-\frac{1}{2}\ln(2\pi)\right\}$$



with (1) $K(x)$ and $S(x)$ being functions only of x, (2) $p(\mu)$ and $q(\mu)$ being functions only of the parameter μ, and (3) the support $-\infty < x < +\infty$ not depending on the parameter μ. Then, we had exponential random variables random variables whose p.d.f. can be written in exponential form as:

$$f(x;\theta)=\frac{1}{\theta}e^{-x/\theta}=\exp\left\{-x\left(\frac{1}{\theta}\right)+\ln(1)-\ln\theta\right\}$$



with (1) $K(x)$ and $S(x)$ being functions only of x, (2) $p(\theta)$ and $q(\theta)$ being functions only of the parameter θ, and (3) the support $x \geq 0$ not depending on the parameter θ. Happily, it turns out that writing p.d.f.s and p.m.f.s in exponential form provides us yet a third way of identifying sufficient statistics for our parameters. The following theorem tells us how.

Exponential Criterion

Exponential Criterion. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with a p.d.f. or p.m.f. of the exponential form:

$$f(x;\theta) = exp[K(x)p(\theta) + S(x) + q(\theta)]$$

with a support that does not depend on θ. Then, the statistic:

$$\sum_{i=1}^{n} K(X_i)$$

is sufficient for θ.

Proof. Because $X_1, X_2, ..., X_n$ is a random sample, the joint p.d.f. (or joint p.m.f.) of $X_1, X_2, ..., X_n$ is, by independence:

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) \times f(x_2; \theta) \times \ldots \times f(x_n; \theta)$$

Inserting what we know to be the p.m.f. or p.d.f. in exponential form, we get:

$$f(x_1, \ldots, x_n; \theta) = \exp\left[K(x_1)p(\theta) + S(x_1) + q(\theta)\right] \times \ldots \times \exp\left[K(x_n)p(\theta) + S(x_n) + q(\theta)\right]$$

Collecting like terms in the exponents, we get:

$$f(x_1, \ldots, x_n; \theta) = \exp\left[p(\theta)\sum_{i=1}^{n} K(x_i) + \sum_{i=1}^{n} S(x_i) + nq(\theta)\right]$$

which can be factored as:

$$f(x_1, \ldots, x_n; \theta) = \left\{ \exp\left[p(\theta) \sum_{i=1}^{n} K(x_i) + nq(\theta) \right] \right\} \times \left\{ \exp\left[\sum_{i=1}^{n} S(x_i) \right] \right\}$$

We have factored the joint p.m.f. or p.d.f. into two functions, one (φ) being only a function of the statistic $Y = \sum_{i=1}^{n} K(X_i)$ and the other (h) not depending on the parameter θ:

$$f(x_1,\ldots,x_n;\theta) = \left\{ \exp\left[p(\theta) \sum_{i=1}^{n} K(x_i) + nq(\theta) \right] \right\} \times \left\{ \exp\left[\sum_{i=1}^{n} S(x_i) \right] \right\}$$

$\phi[u(\Sigma K(x_i)); \theta]$ $\quad$ $h(x_1, \ldots, x_n)$

Therefore, the Factorization Theorem tells us that $Y = \sum_{i=1}^{n} K(X_i)$ is a sufficient statistic for θ.

Let $X_1, X_2, ..., X_n$ be a random sample from a geometric distribution with parameter p. Find a sufficient statistic for the parameter p.

Solution. The probability mass function of a geometric random variable is:

$$f(x;p) = (1-p)^{x-1}p$$

for $x = 1, 2, 3,...$ The p.m.f. can be written in exponential form as:

$$f(x;p) = \exp\left[x\log(1-p) + \log(1) + \log\left(\frac{p}{1-p}\right)\right]$$

Therefore, $Y = \sum_{i=1}^{n} X_i$ is sufficient for p. Easy as pie!

Examples of Exponential Family of Distributions

- Exponential families include many of the most common distributions. Among many others, exponential families includes the following:

Normal Distribution
Exponential Distribution
Gamma Distribution
Chi – Square Distribution
Beta Distribution

Bernoulli Distribution
Poisson Distribution
Geometric Distribution
Binomial Distribution
Etc.

Relative Efficiency

- It usually is possible to obtain more than one unbiased estimator for the same target parameter θ.
- If $\hat{\theta}_1$ and $\hat{\theta}_2$ denote two unbiased estimators for the same parameter θ, we prefer to use the estimator with the smaller variance.
- That is, if both estimators are unbiased, $\hat{\theta}_1$ is *relatively more efficient* than $\hat{\theta}_2$ if $V(\hat{\theta}_2) > V(\hat{\theta}_1)$.
- In fact, we use the ratio $V(\hat{\theta}_2)/V(\hat{\theta}_1)$ to define the *relative efficiency* of two unbiased estimators.

Definition

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, then the *efficiency* of $\hat{\theta}_1$ relative to $\hat{\theta}_2$, denoted $\text{eff}(\hat{\theta}_1, \hat{\theta}_2)$, is defined to be the ratio

$$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}.$$

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased estimators for θ, the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$, $eff\ (\hat{\theta}_1, \hat{\theta}_2)$, is greater than 1 only if $V(\hat{\theta}_2) > V(\hat{\theta}_1)$. In this case, $\hat{\theta}_1$ is a better unbiased estimator than $\hat{\theta}_1$.

For example, if $eff\ (\hat{\theta}_1, \hat{\theta}_2) = 1.8$, then $V(\hat{\theta}_2) = (1.8)V(\hat{\theta}_1)$, and $\hat{\theta}_1$ is preferred to $\hat{\theta}_2$. Similarly, if $eff\ (\hat{\theta}_1, \hat{\theta}_2)$ is less than 1—say, 0.73—then $V(\hat{\theta}_2) = (0.73)V(\hat{\theta}_1)$, and $\hat{\theta}_2$ is preferred to $\hat{\theta}_1$.

Consistency

- A sequence of estimators $\hat{\theta}_n$ that converges in probability to the unknown value of the parameter being estimated, as $n \rightarrow \infty$, is called a consistent sequence of estimators, i.e., $\hat{\theta}_n$ is consistent if and only if for every $\epsilon > 0$,

$$\mathbb{P}(|\widehat{\theta}_n - \boldsymbol{\theta}| > \epsilon) \rightarrow 0, \qquad as \;\; n \rightarrow \infty.$$

- Here we denote an estimator by $\hat{\theta}_n$ to denote that we used n data points to form the estimators.

Theorem

- Let $\hat{\theta}_n$ be an unbiased estimator for θ. If $Var(\hat{\theta}_n) \rightarrow 0$, as $n \rightarrow \infty$, then $\hat{\theta}_n$ is a consistent estimator for θ.

- Alternatively

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

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

Example

Example: Let $X_1, X_2, \ldots, X_n$ be the indicators of n Bernoulli trials with success probability θ.

Find a consistent estimator for θ.

- We will Theorem 5.2 to find a consistent estimator for θ. Consider

$$\hat{\theta}_n = \overline{X}_n,$$

Recall that $\mathbb{E}(\hat{\theta}_n) = \mathbb{E}(\overline{X}_n) = \mathbb{E}(X_1) = \theta$ and $\text{Var}(\hat{\theta}_n) = \text{Var}(\overline{X}_n) = \frac{\text{Var}(X_1)}{n} = \frac{\theta(1-\theta)}{n}$. Since $\hat{\theta}_n$ is unbiased $\text{MSE}(\hat{\theta}_n) = \text{Var}(\hat{\theta}_n)$ and as $n \to \infty$, $\text{Var}(\hat{\theta}_n)$ decreases to 0. Thus $\hat{\theta}_n$ is a consistent estimator for θ.

Example

Example: Let $X_1, X_2, \ldots, X_n$ be from a $N(\mu, \sigma^2)$ distribution. Find consistent estimators for μ and σ^2.

Solution:

Consider

$$\hat{\mu}_n = \overline{X}_n \,,$$

and

$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X}_n)^2.$$

- We will use the Theorem to show that the above estimators are consistent.

Recall that $\mathbb{E}(\hat{\mu}_n) = \mathbb{E}(\overline{X}_n) = \mathbb{E}(X_1) = \mu$ and $\text{Var}(\hat{\mu}_n) = \text{Var}(\overline{X}_n) = \frac{\text{Var}(X_1)}{n} = \frac{\sigma^2}{n}$. Since $\hat{\mu}_n$ is unbiased $\text{MSE}(\hat{\mu}_n) = \text{Var}(\hat{\mu}_n)$ and as $n \to \infty$, $\text{Var}(\hat{\mu}_n)$ decreases to 0. Thus $\hat{\mu}_n$ is a consistent estimator for μ.

Recall that we have that $\mathbb{E}(s^2) = \sigma^2$ and $\text{Var}(s^2) = \frac{2\sigma^4}{n-1}$. Since s^2 is unbiased $\text{MSE}(s^2) = \text{Var}(s^2)$ and as $n \to \infty$, $\text{Var}(s^2)$ decreases to 0. Thus by Theorem 5.2, s^2 is a consistent estimator for σ^2.

Fisher's Information and Cramer-Rao Inequality

- The Cramer-Rao Inequality provides us with a lower bound on the variance of an unbiased estimator for a parameter.

Cramér-Rao Inequality. *Let $f(x;\theta)$ be a probability density with continuous parameter θ. Let $X_1,\ldots,X_n$ be independent random variables with density $f(x;\theta)$, and let $\hat{\Theta}(X_1,\ldots,X_n)$ be an unbiased estimator of θ. Assume that $f(x;\theta)$ satisfies two conditions:*

1. *We have*

$$\frac{\partial}{\partial\theta}\left[\int\cdots\int\hat{\Theta}(x_1,\ldots,x_n)\prod_{i=1}^{n}f(x_i;\theta)\,dx_i\right] = \int\cdots\int\hat{\Theta}(x_1,\ldots,x_n)\frac{\partial\prod_{i=1}^{n}f(x_i;\theta)}{\partial\theta}\,dx_1\cdots dx_n, \qquad (2.1)$$

2. *For each θ, the variance of $\hat{\Theta}(X_1, \ldots, X_n)$ is finite.*

Then

$$\operatorname{var}(\hat{\Theta}) \geq \frac{1}{n\,\mathbb{E}\left[\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)^2\right]}, \tag{2.2}$$

where $\mathbb{E}$ denotes the expected value with respect to the probability density function $f(x;\theta)$.

Fisher Information

$$I(\theta) := \mathrm{Var}_\theta \left[\frac{\partial}{\partial \theta} \log f(X|\theta) \right] = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right]$$

is the Fisher information.

- We can thus write CRLB as

Theorm A Let $X_1, \ldots, X_n$ be i.i.d. with density function $f(x|\theta)$. Let

$$T = t(X_1, \ldots, X_n)$$

be an unbiased estimate of θ. Then, under smoothness assumptions on $f(x|\theta)$,

$$\mathrm{Var}(T) \geq \frac{1}{nI(\theta)}.$$

Example 1: Let X be $N(\theta, \sigma^2)$, where $-\infty < \theta < \infty$. and σ^2 is known. Then

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]$$

and

$$ln[f(x;\theta)] = -\frac{1}{2} ln(2\pi\sigma^2) - \frac{(x-\theta)^2}{2\sigma^2}.$$

Differentiating with respect to θ we have

$$\frac{\partial ln[f(x;\theta)]}{\partial\theta} = \frac{x-\theta}{\sigma^2}$$

and

$$\frac{\partial^2 ln[f(x;\theta)]}{\partial\theta^2} = \frac{-1}{\sigma^2}.$$

No matter which version of $I(\theta)$ we use, we see that

$$I(\theta) = E\left(\left[\frac{\partial ln[f(X;\theta)]}{\partial \theta}\right]^2\right)$$
$$= -E\left[\frac{\partial^2 ln[f(X;\theta)]}{\partial \theta^2}\right] = \frac{1}{\sigma^2}.$$

Derive the CRLB for this information

Example 2: Let X be binomial $b(1, \theta)$. Then

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}$$

and

$$ln[f(x;\theta)] = xln(\theta) + (1-x)ln(1-\theta).$$

$$\frac{\partial ln[f(x;\theta)]}{\partial \theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

$$\frac{\partial^2 ln[f(x;\theta)]}{\partial \theta^2} = \frac{-x}{\theta^2} - \frac{1-x}{(1-\theta)^2}.$$

$$I(\theta) = -E\left[\frac{-X}{\theta^2} - \frac{1-X}{(1-\theta)^2}\right]$$

$$= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$$

Derive the CRLB for this information

Now suppose that we have a random sample $X_1, X_2, ..., X_n$ from a distribution with pdf $f(x;\theta)$. The likelihood function is given by

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

and

$$ln[L(\theta)] = \sum_{i=1}^{n} ln[f(x_i;\theta)]$$

which implies that

$$\frac{\partial ln[L(\theta)]}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial ln[f(x_i;\theta)]}{\partial\theta}.$$

Thus, the natural definition of Fisher information in a sample of size n is

$$I_n(\theta) = E\left(\left[\frac{\partial ln[L(\theta)]}{\partial\theta}\right]^2\right).$$

Notice that for $i \neq j$, cross-product terms in this expectation are 0. By independence,

$$E\left[\frac{\partial ln[f(X_i;\theta)]}{\partial\theta}\frac{\partial ln[f(X_j;\theta)]}{\partial\theta}\right]$$

$$= E\left[\frac{\partial ln[f(X_i;\theta)]}{\partial\theta}\right] E\left[\frac{\partial ln[f(X_j;\theta)]}{\partial\theta}\right] = 0$$

It follows that

$$I_n(\theta) = \sum_{i=1}^{n} E\left(\left[\frac{\partial ln[f(X_i;\theta)]}{\partial\theta}\right]^2\right) = nI(\theta)$$

Theorem A Cramer-Rao Inequality
Let $X_1, \ldots, X_n$ be i.i.d with density function $f(x;\theta)$.

Let $T = u(X_1, X_2, ..., X_n)$ be an estimator of θ. We allow that T might be biased, and denote its expectation by

$$E[T] = E[u(X_1, ..., X_n)] = k(\theta).$$

It turns out that we can bound $Var(T)$ from below using the **Cramer-Rao inequality**,

$$Var(T) \geq \frac{[k'(\theta)]^2}{nI(\theta)}.$$

If $T = u(X_1, X_2, ..., X_n)$ is an unbiased estimator of θ, then $k(\theta) = \theta$ and $k'(\theta) = 1$. In this case, the Cramer-Rao inequality becomes

$$Var(T) \geq \frac{1}{nI(\theta)}.$$

Recall from Examples 1 and 2 that $\frac{1}{nI(\theta)}$ equals σ^2/n and $\theta(1-\theta)/n$, respectively. Thus, we see that in both cases the sample mean $\bar{X}$ achieves the Rao-Cramer lower bound.

Definition Let T be an unbiased estimator of θ. The statistic T is called an **efficient estimator** of θ if and only if the variance of T attains the Cramer-Rao lower bound.

Definition The ratio of the Rao-Cramer lower bound to the actual variance of an unbiased estimator of θ is called the **efficiency** of that estimator.

Example 3 Let $X_1, X_2, ..., X_n$ be a random sample from a Poisson distribution with mean $\theta > 0$. We have seen that $\bar{X}$ is the maximum likelihood estimator of θ.

$$f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!}$$

$$ln[f(x;\theta)] = xln(\theta) - \theta - ln(x!)$$

$$\frac{\partial ln[f(x;\theta)]}{\partial \theta} = \frac{(x-\theta)}{\theta}$$

$$E\left(\left[\frac{\partial ln[f(X;\theta)]}{\partial \theta}\right]^2\right) = \frac{\sigma^2}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta}$$

We see that the Rao-Cramer lower bound is θ/n, which is the variance of $\bar{X}$. Hence $\bar{X}$ is an efficient estimator of θ.

Summary

- **Methods of estimation**
- 1. MLE
- 2. Method of Moments

- **Properties of Estimators**
- 1. Unbiasedness
- 2. Sufficiency
- 3. Efficiency
- 4. Consistency

- THANK YOU

Properties of MLE Estimators

- *Maximum Likelihood Estimation* (MLE) is a widely used statistical estimation method. In this section, we will study its properties: **efficiency, consistency and asymptotic normality**

Efficient Estimator

Definition 1. EFFICIENT ESTIMATOR
An estimator $\widehat{\theta}(y)$ is efficient if it achieves equality in CRLB.

- That is

$$\mathbf{Var}\left(\widehat{\theta}(Y)\right) \geqslant \frac{1}{I(\theta)},$$

Example 1.
Question: $\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_n\}$ are i.i.d. Gaussian random variables with distribution $N(\theta, \sigma^2)$. Determine the maximum likelihood estimator of θ. Is the estimator efficient?

Solution: Let $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ be the observation, then

$$\begin{aligned} f(\boldsymbol{y}\,;\theta) &= \prod_{k=1}^{n} f(y_k\,;\theta) \\ &= \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_k-\theta)^2}{2\sigma^2}\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{\sum_{k=1}^{n}(y_k-\theta)^2}{2\sigma^2}\right\}. \end{aligned}$$

Take the log of both sides of the above equation, we have

$$\log f(\boldsymbol{y}\,;\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{k=1}^{n}(y_k-\theta)^2}{2\sigma^2}.$$

Since $\log f(\boldsymbol{y}\,;\theta)$ is a quadratic concave function of θ, we can obtain the MLE by solving the following equation.

$$\frac{\partial \log f(\boldsymbol{y}\,;\theta)}{\partial\theta} = \frac{2\sum_{k=1}^{n}(y_k-\theta)}{2\sigma^2} = 0.$$

Therefore, the MLE is $\widehat{\theta}_{MLE}(\boldsymbol{y}) = \frac{1}{n}\sum_{k=1}^{n} y_k$.

Now let us check whether the estimator is efficient or not. It is easy to check that the MLE is an unbiased estimator ($\mathbb{E}[\widehat{\theta}_{MLE}(\boldsymbol{y})] = \theta$). To determine the CRLB, we need to calculate the Fisher information of the model.

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{y}\,;\theta)\right] = \frac{n}{\sigma^2} \tag{4}$$

According to Equation 3, we have

$$\mathrm{Var}\left(\widehat{\theta}_{MLE}(\boldsymbol{Y})\right) \geqslant \frac{1}{I(\theta)} = \frac{\sigma^2}{n}. \tag{5}$$

And the variance of the MLE is

$$\mathrm{Var}\left(\widehat{\theta}_{MLE}(\boldsymbol{Y})\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right) = \frac{\sigma^2}{n}. \tag{6}$$

So CRLB equality is achieved, thus the MLE is efficient.

Minimum Variance Unbiased Estimator (MVUE)

- *Minimum Variance Unbiased Estimator (MVUE)* is an unbiased estimator whose variance is lower than any other unbiased estimator for all possible values of parameter θ. That is

$$\mathrm{Var}(\widehat{\theta}_{MVUE}(\boldsymbol{Y})) \leqslant \mathrm{Var}(\widehat{\theta}(\boldsymbol{Y}))$$

for any unbiased $\widehat{\theta}(\boldsymbol{Y})$ of any θ.

Proposition 1. UNBIASED AND EFFICIENT ESTIMATORS
If an estimator $\widehat{\theta}(y)$ is unbiased and efficient, then it must be MVUE.

Consistency of MLE

Definition 2. CONSISTENCY
Let $\{Y_1, \cdots, Y_n\}$ be a sequence of observations. Let $\widehat{\theta}_n$ be the estimator using $\{Y_1, \cdots, Y_n\}$. We say that $\widehat{\theta}_n$ is consistent if $\widehat{\theta}_n \xrightarrow{P} \theta$, i.e.,

$$\mathbb{P}\left(|\widehat{\theta}_n - \theta| > \varepsilon\right) \to 0, \text{ as } n \to \infty \tag{14}$$

Remark: A sufficient condition to have Equation 14 is that

$$\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] \to 0, \quad \text{as } n \to \infty.$$

Proof.
According to Chebyshev's inequality, we have

$$\mathbb{P}\left(|\widehat{\theta}_n - \theta| \geq \varepsilon\right) \leq \frac{\mathbb{E}\left[(\widehat{\theta}_n - \theta)^2\right]}{\varepsilon^2}$$

Since $\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] \to 0$, the we have

$$0 \leq \mathbb{P}\left(|\widehat{\theta}_n - \theta| \geq \varepsilon\right) \leq \frac{\mathbb{E}\left[(\widehat{\theta}_n - \theta)^2\right]}{\varepsilon^2} \to 0.$$

Therefore, $\mathbb{P}\left(|\widehat{\theta}_n - \theta| > \varepsilon\right) \to 0, \text{ as } n \to \infty.$

Example 3.
$\{Y_1, Y_2, \cdots, Y_n\}$ are i.i.d. Gaussian random variables with distribution $N(\theta, \sigma^2)$. Is the MLE using $\{Y_1, Y_2, \cdots, Y_n\}$ consistent?

Solution:
From Example 1., we know that the MLE is

$$\widehat{\theta}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k.$$

Since

$$\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] = \text{Var}\left(\widehat{\theta}_n\right) = \frac{\sigma^2}{n}, \text{ (see Equation 6),}$$

so $\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] \to 0$. Therefore $\widehat{\theta}_n \xrightarrow{p} \theta$, i.e. $\widehat{\theta}_n$ is consistent.

Asymptotic Normality of MLE

- READ THIS ONE. (Definition and Proof)