Notes on Statistics — A Classical Approach

By Photios G. Ioannou, Ph.D., Dipl.Eng., F.ASCE
Professor of Civil and Environmental Engineering
University of Michigan

Point Estimation

Consider a random variable X whose probability density function $f_X(x)$ is defined by certain parameters. For example, the parameters of the Normal distribution are m_X and σ_X^2; the parameter of the Exponential distribution is λ; the parameters of the Binomial are n and p.

In point estimation, we assume that the functional form of the distribution of X is known but its parameters are unknown. For example, we may know that $X \sim N(m_X, \sigma_X)$, but we may not know m_X, or σ_X^2, or both. Similarly, for the Exponential distribution we may not know λ. For the Binomial distribution, we may know n, but not know p.

In classical statistics, the unknown parameters of a distribution $f_X(x)$ are not considered random variables. They are considered unknown *constants*. Thus, our objective in *point estimation* is to use experimental observations (sample data) to estimate these unknown constant parameters and give them a *numerical value*.

Our only source of numerical information are experimental values of the random variable X. For example, we can toss a coin 10 times and observe how many tosses result in heads. In general, we can sample n experimental values x_i $(i = 1, \ldots, n)$ of the random variable X. For the coin example, each x_i can be either 0 or 1.

For now, we shall assume that X is sampled in such a way that the sample outcomes are mutually independent. Because of the independence assumption, and *before* we actually do any real-life sampling, we can think of any future sample of size n as a collection of random variables X_i $(i = 1, \ldots, n)$, where all X_i are independent and identically distributed (iid). This is

called a *random sample*. (Notice that in simulation experiments, however, we often produce *correlated samples* deliberately in an attempt to reduce the variance of our estimates.)

Continuing the coin example, let X be a Bernoulli variable:

$$p_X(x) = \begin{cases} p & x = 1 \quad (heads) \\ 1-p & x = 0 \quad (tails) \end{cases}$$

We can think of X as describing the outcomes of an "unbalanced" coin. If we toss the same coin n times, we will have a sample of n experimental values $x_1, x_2, \ldots, x_n$. *Before* we toss the coin, however, we may consider each toss as a random variable $X_1, X_2, \ldots, X_n$ where all X_i are independent and follow the same distribution $p_{X_i}(x) = p_X(x)$.

To assign a numerical value (*estimate*) to an unknown PDF parameter θ (in the above case, θ is p), we can only use those quantities that we will eventually be able to *observe*, *i.e.*, the random variables X_i $(i = 1, \ldots, n)$. This is true no matter how many unknown parameters θ_j a PDF has. If the distribution $f_X(x)$ has h unknown parameters, θ_j $(j = 1, \ldots, h)$, we need to use the same n outcomes of the random variables X_i to estimate all unknown parameters θ_j.

Statistics, Estimators, and Estimates

We shall now distinguish between *statistics*, *estimators* and *estimates*.

Statistic

In general, a *statistic* is defined to be <u>any function of the sample</u> X_i $(i = 1, \ldots, n)$ <u>that can be computed numerically</u>. To be computed numerically, a statistic must not involve any unknown quantities. It should only be a function of those quantities that we can eventually observe so that the statistic can give us a numerical value. A statistic would be of little use if it contained unknown quantities (such as unknown PDF parameters) that cannot be observed and evaluated *experimentally* because such a statistic would not be numerically computable.

For example, the following function <u>is not</u> a statistic because it requires the true mean m_X whose numerical value is unknown:

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

In contrast, the following function uses the sample average $\bar{X}$ (instead of the true mean m_X). Thus this function is a statistic:

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

Estimator

An *estimator* $\hat{\Theta}$ is a statistic that *we have chosen to use* for computing the value of an unknown constant parameter θ. For example, the sample average $\bar{X}$ is an *estimator* for the true mean m_X. From the definition of a statistic we see that given a sample of size n we can define a great variety of statistics (*i.e.*, functions of the random variables X_i). Obviously, an important question is which statistic (function) is the best estimator for a particular parameter θ. This question and the criteria for choosing estimators will be addressed below.

Estimate

The *estimate* $\hat{\theta}$ of an unknown parameter θ is the *numerical value* of an estimator *after* we observe the n experimental values $x_1, x_2, \ldots, x_n$ of a specific sample.

Sample Moments

Sample moments are used in many of the most common estimators. The definitions of sample moments are similar to those for the moments of a probability density function:

The k-th sample moment: $\overline{X}_{(k)} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

Sample average (1st moment): $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Second sample moment (average of squares): $\overline{X}_{(2)} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

Sample variance (where the divisor is n): $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$

Sample variance (where the divisor is $n-1$): $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$

The reasons for having two definitions for the sample variance, S_n^2 and S_{n-1}^2, are explained below. Clearly, the two sample variances, S_n^2 and S_{n-1}^2, are related as follows:

$$S_{n-1}^2 = \frac{n}{n-1} S_n^2$$

Also, an easy way to compute the sample variance S_n^2, and thus S_{n-1}^2, is to use the following formula which gives *the average of the squares* minus t*he square of the average*:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2 = \overline{X}_{(2)} - (\overline{X})^2$$

Moments of Estimators

Let $\hat{\Theta} = g(X_1, X_2, ..., X_n)$ be the estimator for the unknown PDF parameter θ. Since an estimator is a statistic, it is a function of the random variables X_i ($i = 1, \ldots, n$). Before we conduct any sampling, the estimator $\hat{\Theta} = g(X_1, X_2, ..., X_n)$ is a random variable with its own mean and variance and its own distribution. We can determine the moments of $\hat{\Theta}$ as moments of a function of random variables. Similarly, we can determine its PDF as a derived distribution.

To illustrate, let us determine some of the moments of the most common estimators:

$$\begin{aligned} E[\bar{X}] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n} nE[X] \\ &= E[X] \\ &= m_X \end{aligned}$$

$$\begin{aligned} Var[\bar{X}] &= Var[\frac{1}{n}\sum_{i=1}^{n} X_i] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \quad \text{(because the } X_i \text{ are } IID\text{)} \\ &= \frac{1}{n^2} nVar[X] \\ &= \frac{1}{n}\sigma_X^2 \end{aligned}$$

$$\begin{aligned} E[(\bar{X})^2] &= Var[\bar{X}] + (E[\bar{X}])^2 \\ &= \frac{1}{n}\sigma_X^2 + m_X^2 \quad \text{(this result is used below)} \end{aligned}$$

$$\begin{aligned} E[S_n^2] &= E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2] \\ &= E[\frac{1}{n}\sum_{i=1}^{n} X_i^2] - E[(\bar{X})^2] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - m_X^2 - \frac{1}{n}\sigma_X^2 \\ &= \frac{1}{n}(nE[X^2]) - m_X^2 - \frac{1}{n}\sigma_X^2 \\ &= \sigma_X^2 + m_X^2 - m_X^2 - \frac{1}{n}\sigma_X^2 \\ &= \frac{n-1}{n}\sigma_X^2 \end{aligned}$$

$$\begin{aligned} E[S_{n-1}^2] &= E[\frac{n}{n-1} S_n^2] \\ &= \sigma_X^2 \end{aligned}$$

Notice that in general the moments of estimators are functions of the moments of the underlying random variable, which in turn are functions of the parameters we need to estimate. Thus, the moments of the estimators are *not known* numerically and need to be estimated, too.

Desirable Properties of Estimators

In order for a statistic to be a good estimator it must have some desirable properties that help us to distinguish between "good" and "bad" estimators. The following is an overview of the generally acceptable properties that good estimators should have. Notice that there are no theoretical foundations for why these properties are better than any others. Their wide adoption in statistics is primarily due to the fact that they just make intuitive sense.

Unbiasedness

An estimator $\hat{\Theta}$ is *unbiased* if:

$$E[\hat{\Theta}] = \theta$$

An estimator $\hat{\Theta}$ is *asymptotically unbiased* if:

$$\lim_{n\to\infty} E[\hat{\Theta}] = \theta$$

For example, the sample average $\overline{X}$ is an unbiased estimator of the true mean m_X:

$$E[\overline{X}] = m_X$$

The sample variance S_n^2 is a biased estimator of σ_X^2, but it is asymptotically unbiased.

$$E[S_n^2] = \frac{n-1}{n}\sigma_X^2$$
$$\lim_{n\to\infty} E[S_n^2] = \sigma_X^2$$

Clearly, because of the way it is defined, S_{n-1}^2 is an unbiased estimator of σ_X^2:

$$E[S_{n-1}^2] = \sigma_X^2$$

Given that S_{n-1}^2 is unbiased and S_n^2 is biased, one might wonder why we have defined both sample variances. The answer is given by the next property.

Efficiency

The *efficiency* of an estimator $\hat{\Theta}$ is measured by its variance $Var[\hat{\Theta}]$. In general, the smaller the variance of an estimator the better. The relative efficiency of two estimators, $\hat{\Theta}_1$ and $\hat{\Theta}_2$, is the ratio of their variances. Relative efficiency is used for comparing the spreads of the distributions of the two estimators.

To illustrate the concept of efficiency, consider the two sample variances, S_n^2 and S_{n-1}^2. From their definitions, we have:

$$Var[S_{n-1}^2] = Var[\frac{n}{n-1} S_n^2]$$

$$Var[S_{n-1}^2] = \frac{n^2}{(n-1)^2} Var[S_n^2]$$

Thus, $Var[S_{n-1}^2] > Var[S_n^2]$, and the estimator S_n^2 has less variance than S_{n-1}^2 for any sample size n and any value of $Var[S_n^2]$. This explains the need for both estimators. Even though S_n^2 is biased, it is more efficient than S_{n-1}^2. Note that, in general, it is not easy to compute the actual variance of S_n^2 or S_{n-1}^2. An exception is when X is normally distributed, in which case we have:

$$Var[S_n^2] = \frac{2(n-1)}{n^2} \sigma_X^4$$

$$Var[S_{n-1}^2] = \frac{2}{n-1} \sigma_X^4$$

Sufficiency

An estimator $\hat{\Theta}$ is called *sufficient* if it contains all the available information relevant to the estimation of a parameter θ, that may be found in the sample $X_1, X_2, \ldots, X_n$. To understand sufficiency we need to appreciate that in point estimation we take n random variables, X_i, and combine them into a single function, $\hat{\Theta}$, that yields a single estimate, $\hat{\theta}$. Sufficiency requires that we do no lose any of the information about θ that the n individual outcomes may contain when we combine them into a single function. The mathematical formulation of this property is beyond this discussion.

Consistency

An estimator $\hat{\Theta}$ is called consistent if it stochastically converges to the true parameter θ:

$$\lim_{n\to\infty} P[|\hat{\Theta}-\theta|>\varepsilon]=0 \quad \text{for any } \varepsilon>0$$

If $\hat{\Theta}$ is an unbiased estimator of θ:

$$E[\hat{\Theta}]=\theta$$

then $\hat{\Theta}$ is consistent if:

$$\lim_{n\to\infty} Var[\hat{\Theta}]=0$$

Methods for Finding Good Estimators

We now know what estimators are and some of the properties that they should have. As a result, if we are given a set of possible statistics that may be used as estimators of a particular parameter, we should be able to choose the one that will give the best results. This, however, is a cumbersome process. What we need now is a direct method for finding good estimators for the parameters of a PDF, assuming of course, that we know the general form of this PDF. This will spare us the trouble of adopting a blind search process.

We shall examine two methods for finding good estimators. The first is called the *Method of Moments*, or *MOM*. The second is the *Method of Maximum Likelihood*, or *MML*. An estimator suggested by *MML* is called the *Maximum Likelihood Estimator*, or *MLE*.

Neither of these two methods is guaranteed to give us the "best" possible estimator for every possible parameter of interest. In general, however, both methods produce acceptable results that can be used as a starting point for further refinement.

Method of Moments — MOM

The *Method of Moments* is an easy way to find good estimators. It is so easy that many newcomers to the field of statistics try to make it harder than it really is. The idea is as follows.

We are interested in finding good estimators for the parameters of a certain PDF. In general, the *moments* of this PDF are functions of the *parameters* we want to estimate. For example, the expected value (i.e., the mean) of the exponential PDF is a function of its parameter λ:

$$m_X = \frac{1}{\lambda}$$

For some other PDFs, like the Normal, the parameters may be the same as the moments.

For a PDF with h unknown parameters θ_j $(j = 1, \ldots, h)$, the Method of Moments suggests the following procedure:

1. Take the first h moments of the PDF and solve for the h unknown parameters. The result should be h equations that have each of the unknown parameters θ_j on the left-hand side and some function of the first h moments of the PDF on the right-hand side.

For example, for the Exponential PDF ($h = 1$) we have:

$$\lambda = \frac{1}{m_X}$$

As another example, for the Normal PDF ($h = 2$) we have:

$$m_X = m_X$$
$$\sigma_X^2 = E[X^2] - m_X^2$$

Notice that for the Normal PDF the two unknown *parameters*, m_X and σ_X^2, are shown on the left-hand side, whereas on the right-hand side we have functions of the first two PDF *moments*.

2. Substitute each unknown *moment* of the PDF on the right-hand side with the corresponding sample moment (which can be calculated with data):

$E[X]$	is substituted by	$\overline{X}$
$E[X^2]$	is substituted by	$\overline{X}_{(2)}$
$E[X^k]$	is substituted by	$\overline{X}_{(k)}$

3. Put a hat " ˆ " on the unknown parameter on the left-hand side to indicate that it is now an *estimator* of the parameter under the hat. The MOM is now finished.

For example, for the Exponential PDF ($h = 1$), the MOM estimator for parameter λ is:

$$\hat{\lambda} = \frac{1}{\overline{X}}$$

And for the Normal PDF ($h = 2$), the two MOM estimators for parameters m_X and σ_X^2 are:

$$\begin{aligned} \hat{m}_X &= \overline{X} \\ \hat{\sigma}_X^2 &= \overline{X}_{(2)} - \overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2 \\ &= S_n^2 \end{aligned}$$

Clearly, the MOM is a simple common-sense method that uses the *sample moments* to estimate the unknown *moments* of the PDF, and thus estimate the parameters of the PDF. Its main advantage is simplicity. Sometimes, MOM proposes very good estimators, like in the above two cases, but not always. As a result, the goodness of the proposed MOM estimators should be tested using the definitions of the desirable estimator properties.

Method of Maximum Likelihood — MML

The *Method of Maximum Likelihood* is more sophisticated than the Method of Moments, and usually gives good estimators. To facilitate its discussion, let us introduce the following concise notation:

The set of h unknown PDF parameters:	$\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_h)$
The sample as a set of n random variables:	$\underline{X} = (X_1, X_2, \ldots, X_n)$
The outcome of a sample of size n:	$\underline{x} = (x_1, x_2, \ldots, x_n)$
The Likelihood Function:	$L(\underline{\theta} \mid \underline{X}) = f_{\underline{X}\mid\underline{\theta}}(\underline{x}, \underline{\theta})$

The likelihood function $L(\underline{\theta} \mid \underline{X})$ is the joint PDF of the n random variables $\underline{X}$, given that the h parameters $\underline{\theta}$ are known. Loosely speaking, it gives the probability that the vector $\underline{x}$ will be the realization of the sampling process, given that we know $\underline{\theta}$. The basic idea of the method can be described as follows.

After we observe the sample, $L(\underline{\theta} \mid \underline{X})$ will no longer be a function of the random variables, $\underline{X}$. The random variables $\underline{X}$ will be substituted by the sample outcome $\underline{x}$ which is a set of numbers and not variables. As a result, $L(\underline{\theta} \mid \underline{X})$ will only be a function of the h unknown parameters $\underline{\theta}$. Since $L(\underline{\theta} \mid \underline{X})$ is the *likelihood* (*i.e.*, the probability) of observing this *particular* outcome $\underline{x}$, we should choose $\underline{\theta}$ such that this likelihood is maximized. I.e., we should choose $\underline{\theta}$ so that the observed outcome $\underline{x}$ (which is a set of numbers) is the *mode* (peak) of the joint density $f_{\underline{X}\mid\underline{\theta}}(\underline{x}, \underline{\theta})$.

The following procedure illustrates how to find MLE estimators:

1. Write the likelihood function $L(\underline{\theta} \mid \underline{X})$:

$$L(\underline{\theta} \mid \underline{X}) = f_{X_1\mid\underline{\theta}}(x_1, \underline{\theta}) \cdot f_{X_2\mid\underline{\theta}}(x_2, \underline{\theta}) \cdots f_{X_n\mid\underline{\theta}}(x_n, \underline{\theta}) = \prod_{i=1}^{n} f_{X\mid\underline{\theta}}(x_i, \underline{\theta})$$

Notice that because the random variables X_i can have n different values $x_1, x_2, \ldots, x_n$:

$$L(\underline{\theta} \mid \underline{X}) = \prod_{i=1}^{n} f_{X\mid\underline{\theta}}(x_i, \underline{\theta}) \neq [f_{X\mid\underline{\theta}}(x, \underline{\theta})]^n$$

2. Find the parameters $\hat{\theta}_j$ as functions of the x_i so that $L(\underline{\theta} \mid \underline{X})$ is maximized. To do so set:

$$\frac{L(\underline{\theta} \mid \underline{X})}{\partial \theta_j}\Big|_{\underline{\hat{\theta}}} = 0 \quad j = 1, 2, \ldots, h$$

Notice that this maximization must be done subject to any constraints on θ_j that may exist. For example, $\theta_j > 0$.

3. Solve the resulting h equations for the unknown $\hat{\theta}_j$ which become the MLE estimators $\hat{\Theta}_j$. Notice that the sample values x_i for the random variables $\underline{X}$ are treated as known (because they will eventually be observed). This is the reason behind the notation $L(\underline{\theta}\,|\,\underline{X})$ as opposed to the more natural $L(\underline{X}\,|\,\underline{\theta})$.

Note: Frequently, it is much easier to maximize the natural log of $L(\underline{\theta}\,|\,\underline{X})$ (*i.e.*, $\ln L(\underline{\theta}\,|\,\underline{X})$) rather than $L(\underline{\theta}\,|\,\underline{X})$. Maximizing either function yields the same MLE estimators $\hat{\theta}_j$ because $\ln L(\underline{\theta}\,|\,\underline{X})$ is a monotonically increasing function of $L(\underline{\theta}\,|\,\underline{X})$. Notice that $\ln L(\underline{\theta}\,|\,\underline{X})$ always exists because the function $L(\underline{\theta}\,|\,\underline{X})$ represents *probability* and thus it is never negative.

Example: MLE for the Normal Distribution

To illustrate *MML*, let us find the maximum likelihood estimators (*MLE*) for the two parameters ($h = 2$) of the Normal distribution, $\underline{\theta} = (m_X, \sigma_X^2)$.

We begin by writing the likelihood function:

$$L(\underline{\theta}\,|\,\underline{X}) = \frac{1}{\sigma_X\sqrt{2\pi}} exp\{-\frac{1}{2}(\frac{x_1 - m_X}{\sigma_X})^2\} \cdot \frac{1}{\sigma_X\sqrt{2\pi}} exp\{-\frac{1}{2}(\frac{x_2 - m_X}{\sigma_X})^2\} \cdots \frac{1}{\sigma_X\sqrt{2\pi}} exp\{-\frac{1}{2}(\frac{x_n - m_X}{\sigma_X})^2\}$$
$$= (\frac{1}{\sigma_X\sqrt{2\pi}})^n exp\{-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i - m_X}{\sigma_X})^2\}$$

We can now take the natural log of the likelihood function:

$$\ln L(\underline{\theta}\,|\,\underline{X}) = -\frac{n}{2}\ln(2\pi\sigma_X^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - m_X)^2}{\sigma_X^2}$$

Setting

$$\frac{\partial \ln L(\underline{\theta}\,|\,\underline{X})}{\partial m_X}\Big|_{\hat{m}_X, \hat{\sigma}_X^2} = 0$$

gives

$$-\frac{1}{2}\sum_{i=1}^{n} 2(\frac{x_i - \hat{m}_X}{\hat{\sigma}_X})(-\frac{1}{\hat{\sigma}_X}) = 0$$

$$\frac{1}{\hat{\sigma}_X^2}\sum_{i=1}^{n}(x_i-\hat{m}_X)=0$$

$$\sum_{i=1}^{n}(x_i-\hat{m}_X)=0$$

$$(\sum_{i=1}^{n}x_i)-n\hat{m}_X=0$$

$$\hat{m}_X=\overline{X}$$

Similarly, setting

$$\frac{\partial \ln L(\underline{\theta}\,|\,\underline{X})}{\partial \sigma_X^2}\Big|_{\hat{m}_X,\hat{\sigma}_X^2}=0$$

gives

$$-\frac{n}{2}\frac{1}{\hat{\sigma}_X^2}+\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{m}_X)^2\frac{1}{\hat{\sigma}_X^4}=0$$

$$-n\hat{\sigma}_X^2+\sum_{i=1}^{n}(x_i-\hat{m}_X)^2=0$$

$$\hat{\sigma}_X^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})^2$$

$$\hat{\sigma}_X^2=S_n^2$$

Clearly, the *Method of Moments* and the *Method of Maximum Likelihood* propose the same estimators for the parameters of the Normal distribution:

$$\hat{m}_X=\overline{X}$$
$$\hat{\sigma}_X^2=S_n^2$$

However, this is not always the case.

Numerical Example

As a numerical example, suppose that we are interested in $X\sim N(m_X,\sigma_X^2)$ with unknown parameters m_X and σ_X^2. Experimentation produces the following sample of size 20:

890	989	954	1026	985	970	950	1015	1044	1034
996	1001	979	1055	969	1030	1040	911	945	947

The maximum likelihood estimates for m_X and σ_X^2 are:

$$\hat{m}_X=\overline{x}=\frac{1}{20}\sum_{i=1}^{20}x_i=986.5$$

$$\overline{X}_{(2)} = \frac{1}{20}\sum_{i=1}^{20} x_i^2 = 975{,}132.7$$

$$\hat{\sigma}_X^2 = S_n^2 = \overline{X}_{(2)} - \overline{x}^2 = 975{,}132.7 - 986.5^2 = 1950.45$$

$$\hat{\sigma}_X = \sqrt{S_n^2} = 44$$

Invariance and other Properties of Maximum Likelihood Estimators

The *invariance property* of maximum likelihood estimators is an important property that allows us to find easily the maximum likelihood estimator for any quantity $g(\theta_1, \theta_2, \ldots, \theta_m)$ that is a function of the parameters of a PDF.

The invariance property can be stated as follows:

> If $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_m$, are maximum likelihood estimators of $\theta_1, \theta_2, \ldots, \theta_m$,
>
> then $g(\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_m)$ is a maximum likelihood estimator of $g(\theta_1, \theta_2, \ldots, \theta_m)$.

As an example, let us find the maximum likelihood estimator of $E[X^2]$ for the Normal distribution. The maximum likelihood estimators for the Normal distribution parameters are $\hat{m}_X = \overline{X}$ and $\hat{\sigma}_X^2 = S_n^2$. The unknown $E[X^2]$ is a function of m_X and σ_X^2 given by:

$$E[X^2] = \sigma_X^2 + m_X^2$$

Thus, using the invariance property, the maximum likelihood estimator for $E[X^2]$ is:

$$\begin{aligned} \widehat{E[x^2]} &= \hat{\sigma}_X^2 + \hat{m}_X^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ &= \overline{X}_{(2)} \end{aligned}$$

In addition to the invariance property, maximum likelihood estimators also have the following properties (at least asymptotically) under very general conditions.

- Asymptotically unbiased
- Asymptotically efficient
- Consistent

Notice, however, that because the above properties are asymptotic (large *n*), it is possible that better estimators maybe available when the sample size *n* is small.

Simulation Output Data and Stochastic Processes

A *stochastic process* is a collection of "similar" random variables, ordered over time, that are defined on a *common sample space*, called the *state space*. A *stochastic process* can be:

- *Discrete-time* stochastic process: $X_1, X_2, \ldots$
- *Continuous-time* stochastic process: $\{X(t), t \geq 0\}$

Stochastic processes are often used to describe s*imulation output data*, but with simplifying assumptions to make analytical results possible. One such assumption is that the random process is *covariance-stationary.*

A discrete-time stochastic process, $X_1, X_2, \ldots$, is said to be *covariance-stationary* if all X_i have the same *mean* and the same *variance*, which do not depend on i (i.e., they are *stationary over time*), and the covariance of any pair (X_i, X_{i+j}) depends only on their *relative* time separation j and not on i (i.e., the covariance does not depend on *absolute time*):

$$\mu_i = \mu \quad (= m_X) \quad \text{for } i = 1, 2, \ldots$$
$$\sigma_i = \sigma \quad (= \sigma_X) \quad \text{for } i = 1, 2, \ldots$$
$$C_{i,i+j} = C_j \quad (= Cov[X_i, X_{i+j}]) \quad \text{is independent of } i \text{ for } j = 1, 2, \ldots$$
$$\rho_j = \frac{C_{i,i+j}}{\sigma_i \sigma_{i+j}} = \frac{C_j}{\sigma^2} = \frac{C_j}{C_0} \quad \text{is independent of } i \text{ for } j = 1, 2, \ldots$$

Another typical assumption is that simulation output results, $X_i, (i = 1, 2, \ldots, n)$, are independent. However, this may be inappropriate and may lead to serious underestimates of true variability because in most cases in practice simulation output results tend to be *correlated*.

Fortunately, the average $\overline{X}$ remains an unbiased estimator of $\mu = m_X$ even if the data are correlated. However, the sample variance S^2_{n-1} is no longer an unbiased estimator of $\sigma^2 = \sigma^2_X$. Instead, for a covariance stationary stochastic process we have:

$$\frac{E[S^2_{n-1}]}{\sigma^2} = 1 - \frac{2}{n-1} \sum_{j=1}^{n-1} (1 - \frac{j}{n}) \rho_j$$

Thus, for the common case where the correlation coefficients ρ_j are positive, the above right-hand side implies that $E[S^2_{n-1}] < \sigma^2$. Thus, S^2_{n-1} *underestimates* the true variance σ^2.

Because simulation data are often correlated, at least some of the formulas from classical statistics that assume a *random sample*, *i.e,* that the $X_1, X_2, \ldots, X_n$, are independent, do not apply directly. Even so, not everything is lost. Often, it is possible to group simulation output data into new observations, to which the formulas based on IID observations *can* be applied. Thus, with due care, we can use the formulas for IID observations to analyze simulation results indirectly.

Central Limit Theorem

The Central Limit Theorem states that *under very general conditions* the sum of random variables tends to follow a Normal distribution. These *conditions* are not very strict. They require (a) that we sum up *enough* random variables, (b) that they are *not strongly dependent* on each other, and (c) that they do not include *dominant* random variables.

How many random variables are *enough* depends on the shape of their distributions. The more symmetrical the distributions, the fewer variables are needed. For example, if the random variables are IID and all follow a uniform distribution, as few as four variables may be enough. The random variables being summed do not need to be independent as long as they are not too strongly dependent on each other. Also, the random variables being summed should be comparable in magnitude. Clearly, the central limit theorem cannot apply if the magnitude of one of the random variables being summed is much larger than the rest. In this case, the distribution of the sum will essentially be the same as the distribution of the dominant random variable.

Estimating the Sample Average

Moments of the Sample Average $\bar{X}$

If $X_1, X_2, \ldots, X_n$, are IID then:

$$E[\bar{X}] = m_{\bar{X}} = m_X = \mu$$
$$Var[\bar{X}] = \sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n} = \frac{\sigma^2}{n}$$

If $X_1, X_2, \ldots, X_n$, follow a covariance-stationary process, then:

$$E[\bar{X}] = m_{\bar{X}} = m_X = \mu$$
$$Var[\bar{X}] = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}[1 + 2\sum_{j=1}^{n-1}(1 - \frac{j}{n})\rho_j]$$

The formulas for a covariance stationary process are included here as a reminder that the variance of the sample average is not always σ^2_X / n and cannot always be estimated by S^2_{n-1} / n (unless the observations are close to IID).

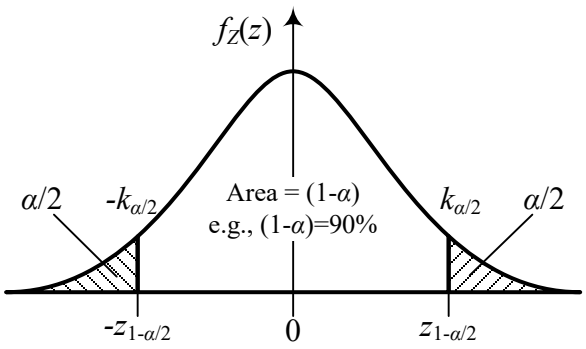
Distribution of the Sample Average $\bar{X}$

For many practical cases where $X_1, X_2, \ldots, X_n$, are close to IID, the Central Limit Theorem does apply, and it is reasonable to state that:

$$\bar{X} \sim N(m_{\bar{X}}, \sigma_{\bar{X}}) = N(\mu, \sqrt{\frac{\sigma^2}{n}})$$

Equivalently, the standardized sample average follows the standard Normal distribution:

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}} \sim N(0,1)$$



Let $z_{1-\alpha/2} = k_{\alpha/2}$ be the $(1-\alpha/2)$ percentile of X.[1] We can now make probability statements of the type:

$$P[-z_{1-\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}] = 1-\alpha$$

$$P[-k_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le k_{\alpha/2}] = 1-\alpha$$

Each of these probability expressions "can be turned inside out" to construct a $(1-\alpha)$ *confidence interval* for the true mean μ.

$$P[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}] = 1-\alpha$$

$$P[\bar{X} - k_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + k_{\alpha/2}\frac{\sigma}{\sqrt{n}}] = 1-\alpha$$

The above expressions for a confidence interval, however, require that we know the true variance of the distribution σ^2. Fortunately, because for large samples the estimator S_{n-1}^2 converges to σ^2, we can substitute σ^2 with S_{n-1}^2 and still write:

$$P[\bar{X} - z_{1-\alpha/2}\frac{S_{n-1}}{\sqrt{n}} \le \mu \le \bar{X} + z_{1-\alpha/2}\frac{S_{n-1}}{\sqrt{n}}] = 1-\alpha$$

$$P[\bar{X} - k_{\alpha/2}\frac{S_{n-1}}{\sqrt{n}} \le \mu \le \bar{X} + k_{\alpha/2}\frac{S_{n-1}}{\sqrt{n}}] = 1-\alpha$$

[1] Some textbooks and computer programs indicate these percentiles using the notation $z_{1-\alpha/2}$ (whose subscript indicates the area to the left of the value), whereas other textbooks use the notation $k_{\alpha/2}$ (whose subscript indicates the area to the right of the value). For completeness, we show formulas using both notations.

The Student t Distribution and the Sample Average $\bar{X}$

If $X_1, X_2, \ldots, X_n$, are *Normal* random variables, then the random variable

$$T = \frac{\bar{X} - \mu}{\sqrt{S_{n-1}^2 / n}}$$

follows a Student t distribution with $n-1$ degrees of freedom.

Notice the functional similarity between the definitions of $Z \sim N(0,1)$ and $T \sim$ Student with $n-1$ degrees of freedom. The only difference is that Z uses σ^2 whereas T uses S_{n-1}^2.

Let $t_{\alpha/2,n-1}$ be the $(1-\alpha/2)$ percentile of the Student distribution with $n-1$ degrees of freedom.[2]

An exact $1-\alpha$ confidence interval is given by:

$$P[\bar{X} - t_{\alpha/2,n-1}\sqrt{\frac{S_{n-1}^2}{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}\sqrt{\frac{S_{n-1}^2}{n}}] = 1-\alpha$$

or

$$P[\bar{X} - t_{\alpha/2,n-1}\frac{S_{n-1}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}\frac{S_{n-1}}{\sqrt{n}}] = 1-\alpha$$

Just like the standard Normal distribution, the Student t distribution is symmetrical around zero but is less peaked and has longer tails. Consequently, it produces wider confidence intervals for the true mean μ. As the sample size n increases, so do the degrees of freedom $n-1$ and the more the Student t distribution approaches the standard Normal distribution.

Given modern computing capabilities, in most practical cases it is best to construct confidence intervals for the true mean μ using the Student distribution approximation. If the sample size n is indeed very large, the Student distribution will converge to the standard Normal distribution automatically.

---

[2] The notation for this percentile adopted here is $t_{\alpha/2,n-1}$ where the subscript $\alpha/2$ indicates the area to the right of the value (similar to $k_{\alpha/2}$). Some sources, use the same notation but reverse the order of the subscripts and write $t_{n-1,\alpha/2}$. And still other sources use the notation $t_{1-\alpha/2,n-1}$ (or $t_{n-1,1-\alpha/2}$) where the subscript $1-\alpha/2$ indicates the area to the left of the value (similar to $z_{1-\alpha/2}$).

Confidence Interval for the Expected Value in Excel

Let X be a Normal random variable with unknown true mean μ and let $X_1, X_2, \ldots, X_n$ be IID sample observations of X. Then, the $(1-\alpha)$ confidence interval for μ is:

$$\overline{X} \pm \frac{S_{n-1}}{\sqrt{n}} t_{\alpha/2,n-1}$$

where $t_{\alpha/2,n-1}$ is the $1-\alpha/2$ percentile of the Student t distribution with $n-1$ degrees of freedom.

The above is true if X is normally distributed, or when the sample size is sufficiently large, say $n \geq 30$, so that the central limit theorem applies and $\overline{X}$ is normally distributed.

The calculation of confidence intervals in Excel is easy but not very well documented. The Excel built-in functions needed to calculate the quantities above are:

$$\overline{X} = \text{AVERAGE(range)}$$
$$S_{n-1} = \text{STDEV(range)}$$
$$t_{\alpha/2,n-1} = \text{TINV}(\alpha, n-1)$$

To see how the TINV function works, let us assume that we have a sample of size $n = 50$ and we would like to construct a 95% confidence interval (*i.e.,* $1-\alpha = 95\%$). In this case, $\alpha/2 = 0.025$ and $n-1 = 49$, so the associated t-statistic value is

$$t_{(0.025,49)} = \text{TINV}(0.05, 49) = 2.009575199\,.$$

A related Excel function is TDIST which gives the *complement* of the CDF of the Student distribution. If T follows a Student t distribution with $n-1$ degrees of freedom, then the *complement* of the CDF of T, *i.e.,* $G_T(t) = 1 - F_T(t) = P[T > t]$, is given by:

TDIST(t, degrees_freedom, tails)

If tails = 1, then TDIST $= P[T \geq t]$

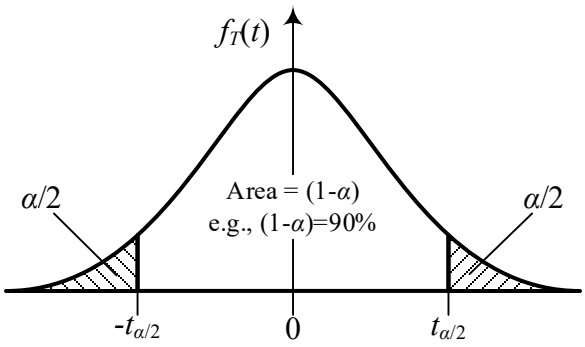
If tails = 2, then TDIST = $P[(T < -t) \cup (t < T)] = P[|T| > t]$

The function TDIST is not defined for negative values, $-t < 0$, so we must use symmetry:

$$\text{TDIST}(-t, df, 1) = 1 - \text{TDIST}(t, df, 1) = P[T > -t]$$
$$\text{TDIST}(-t, df, 2) = \text{TDIST}(t, df, 2) = P[|T| > t]$$

Excel Functions for the Student-t Distribution

The functions provided by Excel to evaluate the Student t distribution are shown below in relation to the probability density function of the Student t distribution, $f_T(t)$.



New Student t Functions	Old Student t Functions
1-α/2 = T.DIST ($t_{\alpha/2}$, n, *cum*=1)	
α/2 = T.DIST.RT ($t_{\alpha/2}$, n)	α/2 = TDIST ($t_{\alpha/2}$, n, *tails*=1)
α = T.DIST.2T ($t_{\alpha/2}$, n)	α = TDIST ($t_{\alpha/2}$, n, *tails*=2)
$t_{\alpha/2}$ = T.INV.2T (α, n)	$t_{\alpha/2}$ = TINV (α, n)
$t_{\alpha/2}$ = T.INV (1-α/2, n)	
$t_{\alpha/2,n-1} \cdot S_{n-1} / \sqrt{n}$ = CONFIDENCE.T(α, S_{n-1}, n)	

Experimental Design — Required Sample Size

As we have seen above, the Central Limit Theorem implies that for a large-enough sample of IID observations, $X_1, X_2, \ldots, X_n$, the distribution of the sample average $\bar{X}$ can be approximated by the Normal distribution:

$$\bar{X} \sim N(m_{\bar{X}}, \sigma_{\bar{X}}) = N(\mu, \sqrt{\frac{\sigma^2}{n}})$$

Equivalently, the standardized sample average follows the standard Normal distribution:

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}} \sim N(0,1)$$

On the basis of these observations we were able to construct a $1-\alpha$ confidence interval for the unknown mean μ:

$$P[\bar{X} - k_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + k_{\alpha/2}\frac{\sigma}{\sqrt{n}}] = 1-\alpha$$

The above confidence interval is centered on $\bar{X}$ and has a halfwidth D equal to:

$$D = k_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

We now turn our attention to a very important question. **What sample size n do we need to estimate the true mean μ and be accurate within an amount D with probability $1-\alpha$?**

The answer is obtained by rearranging the above expression for D:

$$n = k_{\alpha/2}^2\frac{\sigma^2}{D^2}$$

This expression for n requires the true variance of the distribution σ^2 which is often unknown. Fortunately, for large samples we can substitute σ^2 with S_{n-1}^2 and write the approximation:

$$n = k_{\alpha/2}^2\frac{S_{n-1}^2}{D^2}$$

Note that the above approximation uses $k_{\alpha/2}^2$ which is the $(1-\alpha/2)$ percentile of the standard Normal distribution. One could argue that since we are using S_{n-1}^2 and not σ^2, we should have used $t_{\alpha/2,n-1}^2$, the $(1-\alpha/2)$ percentile of the Student t distribution with $n-1$ degrees of freedom. However, this is not practical because the evaluation of $t_{\alpha/2,n-1}^2$ requires the unknown sample size n to establish the degrees of freedom. Also, the approximation given by $k_{\alpha/2}^2$ is sufficient for most practical cases.

Another practical difficulty is the need to calculate S_{n-1}^2 to estimate σ^2 *before* we have determined the required sample size n. An obvious approach is to perform experiments incrementally and stop when the value of S_{n-1}^2 appears to stabilize enough. We can then use the current value of S_{n-1}^2 to obtain an estimate of the required sample size n that will achieve the desired level of precision in predicting μ.