# Chapter 3

## *Sampling Concepts*

### 3.1 Introduction

In this chapter, we cover the concepts associated with random sampling and the sampling distribution of statistics. These notions are fundamental to computational statistics and are needed to understand the topics covered in the rest of the book. As with Chapter 2, those readers who have a basic understanding of these ideas may safely move on to more advanced topics.

In Section 3.2, we discuss the terminology and concepts associated with random sampling and sampling distributions. Section 3.3 contains a brief discussion of the Central Limit Theorem. In Section 3.4, we describe some methods for deriving estimators (maximum likelihood and the method of moments) and introduce criteria for evaluating their performance. Section 3.5 covers the empirical distribution function and how it is used to estimate quantiles. Finally, we conclude with a section on the MATLAB functions that are available for calculating the statistics described in this chapter and a section on further readings.

### 3.2 Sampling Terminology and Concepts

In Chapter 2, we introduced the idea of a random experiment. We typically perform an experiment where we collect data that will provide information on the phenomena of interest. Using these data, we draw conclusions that are usually beyond the scope of our particular experiment. The researcher generalizes from that experiment to the class of all similar experiments. This is the heart of inferential statistics. The problem with this sort of generalization is that we cannot be absolutely certain about our conclusions. However, by

using statistical techniques, we can measure and manage the degree of uncertainty in our results.

***Inferential statistics*** is a collection of techniques and methods that enable researchers to observe a subset of the objects of interest and using the information obtained from these observations make statements or inferences about the entire population of objects. Some of these methods include the estimation of population parameters, statistical hypothesis testing, and probability density estimation.

The ***target population*** is defined as the entire collection of objects or individuals about which we need some information. The target population must be well defined in terms of what constitutes membership in the population (e.g., income level, geographic area, etc.) and what characteristics of the population we are measuring (e.g., height, IQ, number of failures, etc.).

The following are some examples of populations, where we refer back to those described at the beginning of Chapter 2.

- For the piston ring example, our population is all piston rings contained in the legs of steam-driven compressors. We would be observing the time to failure for each piston ring.
- In the glucose example, our population might be all pregnant women, and we would be measuring the glucose levels.
- For cement manufacturing, our population would be batches of cement, where we measure the tensile strength and the number of days the cement is cured.
- In the software engineering example, our population consists of all executions of a particular command and control software system, and we observe the failure time of the system in seconds.

In most cases, it is impossible or unrealistic to observe the entire population. For example, some populations have members that do not exist yet (e.g., future batches of cement) or the population is too large (e.g., all pregnant women). So researchers measure only a part of the target population, called a ***sample.*** If we are going to make inferences about the population using the information obtained from a sample, then it is important that the sample be representative of the population. This can usually be accomplished by selecting a ***simple random sample***, where all possible samples are equally likely to be selected.

A random sample of size $n$ is said to be ***independent and identically distributed*** (iid) when the random variables $X_1, X_2, \ldots, X_n$ each have a common probability density (mass) function given by $f(x)$. Additionally, when they are both independent and identically distributed (iid), the joint probability density (mass) function is given by

$$f(x_1, \ldots, x_n) = f(x_1) \times \ldots \times f(x_n),$$

which is simply the product of the individual densities (or mass functions) evaluated at each sample point.

There are two types of simple random sampling: sampling with replacement and sampling without replacement. When we sample with replacement, we select an object, observe the characteristic we are interested in, and return the object to the population. In this case, an object can be selected for the sample more than once. When the sampling is done without replacement, objects can be selected at most one time. These concepts will be used in Chapters 6 and 7 where the bootstrap and other resampling methods are discussed.

Alternative sampling methods exist. In some situations, these methods are more practical and offer better random samples than simple random sampling. One such method, called ***stratified random sampling***, divides the population into levels, and then a simple random sample is taken from each level. Usually, the sampling is done in such a way that the number sampled from each level is proportional to the number of objects of that level that are in the population. Other sampling methods include cluster sampling and systematic random sampling. For more information on these and others, see the book by Levy and Lemeshow [1999].

Sometimes the goal of inferential statistics is to use the sample to estimate or make some statements about a population parameter. Recall from Chapter 2 that a ***parameter*** is a descriptive measure for a population or a distribution of random variables. For example, population parameters that might be of interest include the mean ($\mu$), the standard deviation ($\sigma$), quantiles, proportions, correlation coefficients, etc.

A ***statistic*** is a function of the observed random variables obtained in a random sample and does not contain any unknown population parameters. Often the statistic is used for the following purposes:

- as a point estimate for a population parameter,
- to obtain a confidence interval estimate for a parameter, or
- as a test statistic in hypothesis testing.

Before we discuss some of the common methods for deriving statistics, we present some of the statistics that will be encountered in the remainder of the text. In most cases, we assume that we have a random sample, $X_1, \ldots, X_n$, of independent, identically (iid) distributed random variables.

### Sample Mean and Sample Variance

A familiar statistic is the ***sample mean*** given by

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i . \tag{3.1}$$

To calculate this in MATLAB, one can use the function called **mean**. If the argument to this function is a matrix, then it provides a vector of means, each one corresponding to the mean of a column. One can find the mean along any dimension (**dim**) of multi-dimensional arrays using the syntax: **mean(x,dim)**.

Another statistic that we will see again is the ***sample variance***, calculated from

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n(n-1)}\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right). \tag{3.2}$$

The sample standard deviation is given by the square root of the variance (Equation 3.2) and is denoted by $S$. These statistics can be calculated in MATLAB using the functions **std(x)** and **var(x)**, where **x** is an array containing the sample values. As with the function **mean**, these can have matrices or multi-dimensional arrays as input arguments.

### Sample Moments

The sample moments can be used to estimate the population moments described in Chapter 2. The ***r-th sample moment*** about zero is given by

$$M'_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r . \tag{3.3}$$

Note that the sample mean is obtained when $r = 1$. The $r$-th sample moments about the sample mean are statistics that estimate the population central moments and can be found using the following

$$M_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^r . \tag{3.4}$$

We can use Equation 3.4 to obtain estimates for the coefficient of skewness $\gamma_1$ and the coefficient of kurtosis $\gamma_2$. Recall that these are given by

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}, \tag{3.5}$$

and

$$\gamma_2 = \frac{\mu_4}{\mu_2^2}. \tag{3.6}$$

Substituting the sample moments for the population moments in Equations 3.5 and 3.6, we have

$$\hat{\gamma}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{3/2}}, \tag{3.7}$$

and

$$\hat{\gamma}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{2}}. \tag{3.8}$$

We are using the 'hat' notation to denote an estimate. Thus, $\hat{\gamma}_1$ is an estimate for $\gamma_1$. The following example shows how to use MATLAB to obtain the sample coefficient of skewness and sample coefficient of kurtosis.



### Example 3.1

In this example, we will generate a random sample that is uniformly distributed over the interval (0, 1). We would expect this sample to have a coefficient of skewness close to zero because it is a symmetric distribution. We would expect the kurtosis to be different from 3, because the random sample is not generated from a normal distribution.

```
% Generate a random sample from the uniform
% distribution.
n = 200;
x = rand(1,200);
% Find the mean of the sample.
```

```
mu = mean(x);
% Find the numerator and denominator for gamma_1.
num = (1/n)*sum((x-mu).^3);
den = (1/n)*sum((x-mu).^2);
gam1 = num/den^(3/2);
```

This results in a coefficient of skewness of `gam1 = -0.0542`, which is not too far from zero. Now we find the kurtosis using the following MATLAB commands:

```
% Find the kurtosis.
num = (1/n)*sum((x-mu).^4);
den = (1/n)*sum((x-mu).^2);
gam2 = num/den^2;
```

This gives a kurtosis of `gam2 = 1.8766`, which is not close to 3, as expected. ❑

We note that these statistics might not be the best to use in terms of bias (see Section 3.4). However, they will prove to be useful as examples in Chapters 6 and 7, where we look at bootstrap methods for estimating the bias in a statistic. The MATLAB Statistics Toolbox function called **`skewness`** returns the coefficient of skewness for a random sample. The function **`kurtosis`** calculates the sample coefficient of kurtosis (*not* the coefficient of excess kurtosis).

### Covariance

In the definitions given below (Equations 3.9 and 3.10), we assume that all expectations exist. The ***covariance*** of two random variables $X$ and $Y$, with joint probability density function $f(x, y)$, is defined as

$$Cov(X, Y) = \sigma_{X,Y} = E[(X - \mu_X)(Y - \mu_Y)]. \tag{3.9}$$

The ***correlation coefficient*** of $X$ and Y is given by

$$Corr(X, Y) = \rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}, \tag{3.10}$$

where $\sigma_X > 0$ and $\sigma_Y > 0$.

The correlation is a measure of the linear relationship between two random variables. If the joint distribution of two variables has a correlation coefficient, then $-1 \leq \rho_{X,Y} \leq 1$. When $\rho_{X,Y} = 1$, then $X$ and $Y$ are perfectly positively correlated. This means that the possible values for $X$ and $Y$ lie on a line with positive slope. On the other hand, when $\rho_{X,Y} = -1$, then the situation is the opposite: $X$ and $Y$ are perfectly negatively correlated. If $X$ and $Y$ are

independent, then $\rho_{X,Y} = 0$. Note that the converse of this statement does not necessarily hold.

There are statistics that can be used to estimate these quantities. Let's say we have a random sample of size $n$ denoted as $(X_1, Y_1), \ldots, (X_n, Y_n)$. The sample covariance is typically calculated using the following statistic

$$\hat{\sigma}_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \quad . \tag{3.11}$$

This is the definition used in the MATLAB function **cov**. In some instances, the empirical covariance is used [Efron and Tibshirani, 1993]. This is similar to Equation 3.11, except that we divide by $n$ instead of $n-1$. The sample correlation coefficient for two variables is given by

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{1/2}\left(\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right)^{1/2}} \quad . \tag{3.12}$$

In the next example, we investigate the commands available in MATLAB that return the statistics given in Equations 3.11 and 3.12. It should be noted that the quantity in Equation 3.12 is also bounded below by $-1$ and above by 1.

### Example 3.2

In this example, we show how to use the MATLAB **cov** function to find the covariance between two variables and the **corrcoef** function to find the correlation coefficient. Both of these functions are available in the standard MATLAB language. We use the **cement** data [Hand, et al., 1994], which were analyzed by Hald [1952], to illustrate the basic syntax of these functions. The relationship between the two variables is nonlinear, so Hald looked at the log of the tensile strength as a function of the reciprocal of the drying time. When the **cement** data are loaded, we get a vector **x** representing the drying times and a vector **y** that contains the tensile strength. A scatterplot of the transformed data is shown in Figure 3.1.

```
% First load the data.
load cement
% Now get the transformations.
xr = 1./x;
logy = log(y);
% Now get a scatterplot of the data to see if
% the relationship is linear.
```

```
plot(xr,logy,'x')
axis([0 1.1 2.4 4])
xlabel('Reciprocal of Drying Time')
ylabel('Log of Tensile Strength')
```

We now show how to get the covariance matrix and the correlation coefficient for these two variables.

```
% Now get the covariance and
% the correlation coefficient.
cmat = cov(xr,logy);
cormat = corrcoef(xr,logy);
```

The results are:

```
cmat =
    0.1020   -0.1169
   -0.1169    0.1393
cormat =
   1.0000   -0.9803
   -0.9803    1.0000
```

Note that the sample correlation coefficient (Equation 3.12) is given by the off-diagonal element of **cormat**, $\hat{\rho} = -0.9803$ . We see that the variables are negatively correlated, which is what we expect from Figure 3.1 (the log of the tensile strength decreases with increasing reciprocal of drying time).
❑

## 3.3 Sampling Distributions

It was stated in the previous section that we sometimes use a statistic calculated from a random sample as a point estimate of a population parameter. For example, we might use $\overline{X}$ to estimate $\mu$ or use $S$ to estimate $\sigma$. Since we are using a sample and not observing the entire population, there will be some error in our estimate. In other words, it is unlikely that the statistic will equal the parameter. To manage the uncertainty and error in our estimate, we must know the sampling distribution for the statistic. The ***sampling distribution*** is the underlying probability distribution for a statistic. To understand the remainder of the text, it is important to remember that *a statistic is a random variable*.

The sampling distributions for many common statistics are known. For example, if our random variable is from the normal distribution, then we know how the sample mean is distributed. Once we know the sampling distribution of our statistic, we can perform statistical hypothesis tests and calculate confidence intervals. If we do not know the distribution of our statistic,

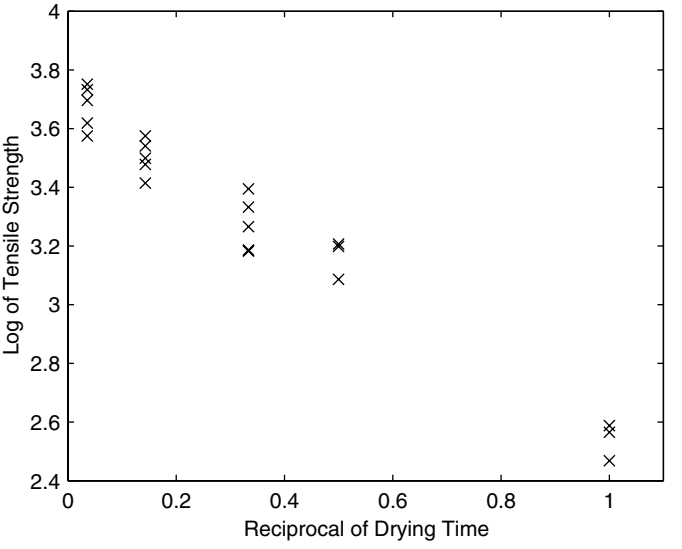


FIGURE 3.1
This scatterplot shows the observed drying times and corresponding tensile strength of the cement. Since the relationship is nonlinear, the variables are transformed as shown here. A linear relationship seems to be a reasonable model for these data.



then we must use Monte Carlo simulation techniques or bootstrap methods to estimate the sampling distribution (see Chapter 6).

To illustrate the concept of a sampling distribution, we discuss the sampling distribution for $\bar{X}$, where the random variable $X$ follows a distribution given by the probability density function $f(x)$. It turns out that the distribution for the sample mean can be found using the Central Limit Theorem.

***CENTRAL LIMIT THEOREM***

*Let $f(x)$ represent a probability density with finite variance $\sigma^2$ and mean $\mu$. Also, let $\bar{X}$ be the sample mean for a random sample of size n drawn from this distribution. For large n, the distribution of $\bar{X}$ is approximately normally distributed with mean $\mu$ and variance given by $\sigma^2/n$.*

❑

The Central Limit Theorem states that as the sample size gets large, the distribution of the sample mean approaches the normal distribution regardless of how the random variable $X$ is distributed. However, if we are sampling from a normal population, then the distribution of the sample mean is exactly normally distributed with mean $\mu$ and variance $\sigma^2/n$.

This information is important, because we can use it to determine how much error there is in using $\overline{X}$ as an estimate of the population mean $\mu$. We can also perform statistical hypothesis tests using $\overline{X}$ as a test statistic and can calculate confidence intervals for $\mu$. In this book, we are mainly concerned with computational (rather than theoretical) methods for finding sampling distributions of statistics (e.g., Monte Carlo simulation or resampling). The sampling distribution of $\overline{X}$ is used to illustrate the concepts covered in remaining chapters.

## 3.4 Parameter Estimation

One of the first tasks a statistician or an engineer undertakes when faced with data is to try to summarize or describe the data in some manner. Some of the statistics (sample mean, sample variance, coefficient of skewness, etc.) we covered in Section 3.2 can be used as descriptive measures for our sample. In this section, we look at methods to derive and to evaluate estimates of population parameters.

There are several methods available for obtaining parameter estimates. These include the method of moments, maximum likelihood estimation, Bayes estimators, minimax estimation, Pitman estimators, interval estimates, robust estimation, and many others. In this book, we discuss the maximum likelihood method and the method of moments for deriving estimates for population parameters. These somewhat classical techniques are included as illustrative examples only and are not meant to reflect the state of the art in this area. Many useful (and computationally intensive!) methods are not covered here, but references are provided in Section 3.7. However, we do present some alternative methods for calculating interval estimates using Monte Carlo simulation and resampling methods (see Chapters 6 and 7).

Recall that a sample is drawn from a population that is distributed according to some function whose characteristics are governed by certain parameters. For example, our sample might come from a population that is normally distributed with parameters $\mu$ and $\sigma^2$. Or, it might be from a population that is exponentially distributed with parameter $\lambda$. The goal is to use the sample to estimate the corresponding population parameters. If the sample is representative of the population, then a function of the sample should provide a useful estimate of the parameters.

Before we undertake our discussion of maximum likelihood, we need to define what an estimator is. Typically, population parameters can take on values from a subset of the real line. For example, the population mean can be any real number, $-\infty < \mu < \infty$, and the population standard deviation can be any positive real number, $\sigma > 0$. The set of all possible values for a parameter $\theta$ is called the ***parameter space***. The ***data space*** is defined as the set of all possible values of the random sample of size $n$. The estimate is calculated from

the sample data as a function of the random sample. An ***estimator*** is a function or mapping from the data space to the parameter space and is denoted as

$$T = t(X_1, \ldots, X_n). \tag{3.13}$$

Since an estimator is calculated using the sample alone, it is a statistic. Furthermore, if we have a random sample, then an estimator is also a random variable. This means that the value of the estimator varies from one sample to another based on its sampling distribution. In order to assess the usefulness of our estimator, we need to have some criteria to measure the performance. We discuss four criteria used to assess estimators: bias, mean squared error, efficiency, and standard error. In this discussion, we only present the definitional aspects of these criteria.



### Bias

The bias in an estimator gives a measure of how much error we have, on average, in our estimate when we use $T$ to estimate our parameter $\theta$. The ***bias*** is defined as

$$\text{bias}(T) = E[T] - \theta. \tag{3.14}$$

If the estimator is unbiased, then the expected value of our estimator equals the true parameter value, so $E[T] = \theta$.

To determine the expected value in Equation 3.14, we must know the distribution of the statistic $T$. In these situations, the bias can be determined analytically. When the distribution of the statistic is not known, then we can use methods such as the jackknife and the bootstrap (see Chapters 6 and 7) to estimate the bias of $T$.

### Mean Squared Error

Let $\theta$ denote the parameter we are estimating and $T$ denote our estimate, then the ***mean squared error*** (MSE) of the estimator is defined as

$$\text{MSE}(T) = E[(T - \theta)^2]. \tag{3.15}$$

Thus, the MSE is the expected value of the squared error. We can write this in more useful quantities such as the bias and variance of $T$. (The reader will see this again in Chapter 8 in the context of probability density estimation.) If we expand the expected value on the right hand side of Equation 3.15, then we have

$$\text{MSE}(T) = E[(T^2 - 2T\theta + \theta^2)] = E[T^2] - 2\theta E[T] + \theta^2. \quad (3.16)$$

By adding and subtracting $(E[T])^2$ to the right hand side of Equation 3.16, we have the following

$$\text{MSE}(T) = E[T^2] - (E[T])^2 + (E[T])^2 - 2\theta E[T] + \theta^2. \quad (3.17)$$

The first two terms of Equation 3.17 are the variance of $T$, and the last three terms equal the squared bias of our estimator. Thus, we can write the mean squared error as

$$\begin{aligned} \text{MSE}(T) &= E[T^2] - (E[T])^2 + (E[T] - \theta)^2 \\ &= V(T) + [\text{bias}(T)]^2. \end{aligned} \quad (3.18)$$

Since the mean squared error is based on the variance and the squared bias, the error will be small when the variance and the bias are both small. When $T$ is unbiased, then the mean squared error is equal to the variance only. The concepts of bias and variance are important for assessing the performance of any estimator.

### Relative Efficiency

Another measure we can use to compare estimators is called efficiency, which is defined using the MSE. For example, suppose we have two estimators $T_1 = t_1(X_1, \ldots, X_n)$ and $T_2 = t_2(X_1, \ldots, X_n)$ for the same parameter. If the MSE of one estimator is less than the other (e.g., $\text{MSE}(T_1) < \text{MSE}(T_2)$), then $T_1$ is said to be more efficient than $T_2$.

The ***relative efficiency*** of $T_1$ to $T_2$ is given by

$$eff(T_1, T_2) = \frac{\text{MSE}(T_2)}{\text{MSE}(T_1)}. \quad (3.19)$$

If this ratio is greater than one, then $T_1$ is a more efficient estimator of the parameter.

### Standard Error

We can get a measure of the precision of our estimator by calculating the standard error. The ***standard error*** of an estimator (or a statistic) is defined as the standard deviation of its sampling distribution:

$$SE(T) = \sqrt{V(T)} = \sigma_T.$$

To illustrate this concept, let's use the sample mean as an example. We know that the variance of the estimator is

$$V(\overline{X}) = \frac{1}{n}\sigma^2 ,$$

for large $n$. So, the standard error is given by

$$SE(\overline{X}) = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} . \tag{3.20}$$

If the standard deviation $\sigma$ for the underlying population is unknown, then we can substitute an estimate for the parameter. In this case, we call it the estimated standard error:

$$\hat{SE}(\overline{X}) = \hat{\sigma}_{\overline{X}} = \frac{S}{\sqrt{n}} . \tag{3.21}$$

Note that the estimate in Equation 3.21 is also a random variable and has a probability distribution associated with it.

If the bias in an estimator is small, then the variance of the estimator is approximately equal to the MSE, $V(T) \approx \text{MSE}(T)$. Thus, we can also use the square root of the MSE as an estimate of the standard error.

### Maximum Likelihood Estimation

A ***maximum likelihood estimator*** is that value of the parameter (or parameters) that maximizes the likelihood function of the sample. The ***likelihood function*** of a random sample of size $n$ from density (mass) function $f(x;\theta)$ is the joint probability density (mass) function, denoted by

$$L(\theta;x_1, \ldots, x_n) = f(x_1, \ldots, x_n;\theta) . \tag{3.22}$$

Equation 3.22 provides the likelihood that the random variables take on a particular value $x_1, \ldots, x_n$. Note that the likelihood function $L$ is a function of the unknown parameter $\theta$, and that we allow $\theta$ to represent a vector of parameters.

If we have a random sample (independent, identically distributed random variables), then we can write the likelihood function as

$$L(\theta) = L(\theta;x_1, \ldots, x_n) = f(x_1;\theta) \times \ldots \times f(x_n;\theta) , \tag{3.23}$$

which is the product of the individual density functions evaluated at each $x_i$ or sample point.

In most cases, to find the value $\hat{\theta}$ that maximizes the likelihood function, we take the derivative of $L$, set it equal to 0 and solve for $\theta$. Thus, we solve the following likelihood equation

$$\frac{d}{d\theta}L(\theta) = 0. \tag{3.24}$$

It can be shown that the likelihood function, $L(\theta)$, and logarithm of the likelihood function, $\ln L(\theta)$, have their maxima at the same value of $\theta$. It is sometimes easier to find the maximum of $\ln L(\theta)$, especially when working with an exponential function. However, keep in mind that a solution to the above equation does not imply that it is a maximum; it could be a minimum. It is important to ensure this is the case before using the result as a maximum likelihood estimator.

When a distribution has more than one parameter, then the likelihood function is a function of all parameters that pertain to the distribution. In these situations, the maximum likelihood estimates are obtained by taking the partial derivatives of the likelihood function (or $\ln L(\theta)$), setting them all equal to zero, and solving the system of equations. The resulting estimators are called the joint maximum likelihood estimators. We see an example of this below, where we derive the maximum likelihood estimators for $\mu$ and $\sigma^2$ for the normal distribution.

**Example 3.3**

In this example, we derive the maximum likelihood estimators for the parameters of the normal distribution. We start off with the likelihood function for a random sample of size $n$ given by

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Since this has the exponential function in it, we will take the logarithm to obtain

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

This simplifies to

$$\ln[L(\theta)] = -\frac{n}{2}\ln[2\pi] - \frac{n}{2}\ln[\sigma^2] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2, \tag{3.25}$$

with $\sigma > 0$ and $-\infty < \mu < \infty$. The next step is to take the partial derivative of Equation 3.25 with respect to $\mu$ and $\sigma^2$. These derivatives are

$$\frac{\partial}{\partial\mu}\ln L = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu), \tag{3.26}$$

and

$$\frac{\partial}{\partial\sigma^2}\ln L = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{3.27}$$

We then set Equations 3.26 and 3.27 equal to zero and solve for $\mu$ and $\sigma^2$. Solving the first equation for $\mu$, we get the familiar sample mean for the estimator.

$$\begin{aligned}
\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) &= 0, \\
\sum_{i=1}^{n} x_i &= n\mu, \\
\hat{\mu} = \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i.
\end{aligned}$$

Substituting $\hat{\mu} = \bar{x}$ into Equation 3.27, setting it equal to zero, and solving for the variance, we get

$$\begin{aligned}
-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= 0 \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.
\end{aligned} \tag{3.28}$$

These are the sample moments about the sample mean, and it can be verified that these solutions jointly maximize the likelihood function [Lindgren, 1993].
❑

We know that the $E[\bar{X}] = \mu$ [Mood, Graybill and Boes, 1974], so the sample mean is an unbiased estimator for the population mean. However, that is not the case for the maximum likelihood estimate for the variance. It can be shown [Hogg and Craig, 1978] that

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

so we know (from Equation 3.14) that the maximum likelihood estimate, $\hat{\sigma}^2$, for the variance is biased. If we want to obtain an unbiased estimator for the variance, we simply multiply our maximum likelihood estimator by $n/(n-1)$. This yields the familiar statistic for the sample variance given by

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

### Method of Moments

In some cases, it is difficult finding the maximum of the likelihood function. For example, the gamma distribution has the unknown parameter $t$ that is used in the gamma function, $\Gamma(t)$. This makes it hard to take derivatives and solve the equations for the unknown parameters. The method of moments is one way to approach this problem.

In general, we write the unknown population parameters in terms of the population moments. We then replace the population moments with the corresponding sample moments. We illustrate these concepts in the next example, where we find estimates for the parameters of the gamma distribution.

### Example 3.4

The gamma distribution has two parameters, $t$ and $\lambda$. Recall that the mean and variance are given by $t/\lambda$ and $t/\lambda^2$, respectively. Writing these in terms of the population moments, we have

$$E[X] = \frac{t}{\lambda}, \tag{3.29}$$

and

$$V(X) = E[X^2] - (E[X])^2 = \frac{t}{\lambda^2}. \tag{3.30}$$

The next step is to solve Equations 3.29 and 3.30 for $t$ and $\lambda$. From Equation 3.29, we have $t = \lambda E[X]$, and substituting this in the second equation yields

$$E[X^2] - (E[X])^2 = \frac{\lambda E[X]}{\lambda^2}. \tag{3.31}$$

Rearranging Equation 3.31 gives the following expression for $\lambda$

$$\lambda = \frac{E[X]}{E[X^2] - (E[X])^2}. \tag{3.32}$$

We can now obtain the parameter $t$ in terms of the population moments (substitute Equation 3.32 for $\lambda$ in Equation 3.29) as

$$t = \frac{(E[X])^2}{E[X^2] - (E[X])^2}. \tag{3.33}$$

To get our estimates, we substitute the sample moments for $E[X]$ and $E[X^2]$ in Equations 3.32 and 3.33. This yields

$$\hat{t} = \frac{\overline{X}^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}, \tag{3.34}$$

and

$$\hat{\lambda} = \frac{\overline{X}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}. \tag{3.35}$$

❑

In Table 3.1, we provide some suggested point estimates for several of the distributions covered in Chapter 2. This table also contains the names of functions to calculate the estimators. In Section 3.6, we discuss the MATLAB code available in the Statistics Toolbox for calculating maximum likelihood estimates of distribution parameters. The reader is cautioned that the estimators

discussed in this chapter are not necessarily the best in terms of bias, variance, etc.

TABLE 3.1

Suggested Point Estimators for Parameters

| Distribution | Suggested Estimator | MATLAB Function |
|---|---|---|
| Binomial<br>Note: $X$ is the number of successes in $n$ trials | $\hat{p} = \frac{X}{n}$ | `csbinpar` |
| Exponential | $\hat{\lambda} = 1/\overline{X}$ | `csexpar` |
| Gamma | $\hat{t} = \overline{X}^2 / \left( \frac{1}{n} \sum X_i^2 - \overline{X}^2 \right)$<br>$\hat{\lambda} = \overline{X} / \left( \frac{1}{n} \sum X_i^2 - \overline{X}^2 \right)$ | `csgampar` |
| Normal | $\hat{\mu} = \overline{X}$<br>$\hat{\sigma}^2 = S^2$ | `mean`<br>`var` |
| Multivariate Normal | $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$<br>$\hat{\Sigma}_{ij} = \frac{n \sum_{k=1}^{n} X_{ik} X_{jk} - \sum_{k=1}^{n} X_{ik} \sum_{k=1}^{n} X_{jk}}{n(n-1)}$ | `mean`<br>`cov` |
| Poisson | $\hat{\lambda} = \overline{X}$ | `cspoipar` |



## 3.5 Empirical Distribution Function

Recall from Chapter 2 that the cumulative distribution function is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \tag{3.36}$$

for a continuous random variable and by

$$F(a) = \sum_{x_i \le a} f(x_i) \tag{3.37}$$

for a discrete random variable. In this section, we examine the sample analog of the cumulative distribution function called the ***empirical distribution function***. When it is not suitable to assume a distribution for the random variable, then we can use the empirical distribution function as an estimate of the underlying distribution. One can call this a ***nonparametric*** estimate of the distribution function, because we are not assuming a specific parametric form for the distribution that generates the random phenomena. In a ***parametric*** setting, we would assume a particular distribution generated the sample and estimate the cumulative distribution function by estimating the appropriate parameters.

The empirical distribution function is based on the ***order statistics***. The order statistics for a sample are obtained by putting the data in ascending order. Thus, for a random sample of size $n$, the order statistics are defined as

$$X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)},$$

with $X_{(i)}$ denoting the $i$-th order statistic. The order statistics for a random sample can be calculated easily in MATLAB using the `sort` function.

The empirical distribution function $\hat{F}_n(x)$ is defined as the number of data points less than or equal to $x$ ($\#(X_i \le x)$) divided by the sample size $n$. It can be expressed in terms of the order statistics as follows

$$\hat{F}_n(x) = \begin{cases} 0; & x < X_{(1)} \\ j/n; & X_{(j)} \le x < X_{(j+1)} \\ 1; & x \ge X_{(n)}. \end{cases} \tag{3.38}$$

Figure 3.2 illustrates these concepts. We show the empirical cumulative distribution function for a standard normal and include the theoretical distribution function to verify the results. In the following section, we describe a descriptive measure for a population called a quantile, along with its corresponding estimate. Quantiles are introduced here, because they are based on the cumulative distribution function.

## Quantiles

Quantiles have a fundamental role in statistics. For example, they can be used as a measure of central tendency and dispersion, they provide the critical val-

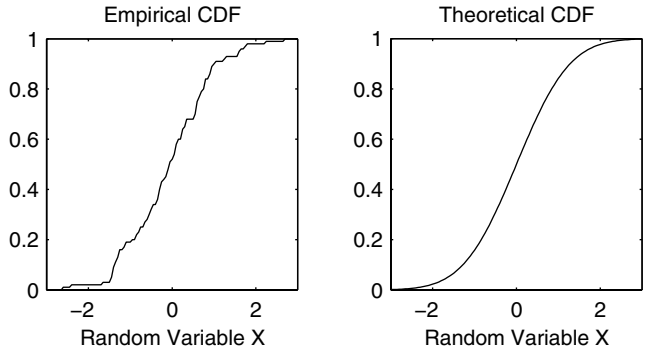


FIGURE 3.2
This shows the theoretical and empirical distribution functions for a standard normal distribution.



ues in hypothesis testing (see Chapter 6), and they are used in exploratory data analysis for assessing distributions (see Chapter 5).

The ***quantile*** $q_p$ of a random variable (or equivalently of its distribution) is defined as the smallest number $q$ such that the cumulative distribution function is greater than or equal to some $p$, where $0 < p < 1$. This can be calculated for a continuous random variable with density function $f(x)$ by solving

$$p = \int_{-\infty}^{q_p} f(x)dx \tag{3.39}$$

for $q_p$, or by using the inverse of the cumulative distribution function,

$$q_p = F^{-1}(p). \tag{3.40}$$

Stating this another way, the $p$-th quantile of a random variable $X$ is the value $q_p$ such that

$$F(q_p) = P(X \leq q_p) = p \tag{3.41}$$

for $0 < p < 1$.

Some well known examples of quantiles are the ***quartiles***. These are denoted by $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$. In essence, these divide the distribution into four equal (in terms of probability or area under the curve) segments. The second quartile is also called the ***median*** and satisfies

$$0.5 = \int_{-\infty}^{q_{0.5}} f(x)dx . \tag{3.42}$$

We can get a measure of the dispersion of the random variable by looking at the ***interquartile range*** (IQR) given by

$$\text{IQR} = q_{0.75} - q_{0.25} . \tag{3.43}$$

One way to obtain an estimate of the quantiles is based on the empirical distribution function. If we let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics for a random sample of size $n$, then $X_{(j)}$ is an estimate of the $(j - 0.5)/n$ quantile [Banks, 2001; Cleveland, 1993]:

$$X_{(j)} \approx F^{-1}\left(\frac{j - 0.5}{n}\right). \tag{3.44}$$

We are not limited to a value of 0.5 in Equation 3.44. In general, we can estimate the $p$-th quantile using the following

$$\hat{q}_p = X_{(j)}; \qquad \frac{j-1}{n} < p \leq \frac{j}{n}; \qquad j = 1, \ldots, n . \tag{3.45}$$

As already stated, Equation 3.45 is not the only way to estimate quantiles. For more information on other methods, see Kotz and Johnson [Vol. 7, 1986]. The analyst should exercise caution when calculating quartiles (or other quantiles) using computer packages. Statistical software packages define them differently [Frigge, Hoaglin, and Iglewicz, 1989], so these statistics might vary depending on the formulas that are used.

### EXAMPLE 3.5

In this example, we will show one way to determine the sample quartiles. The second sample quartile $\hat{q}_{0.5}$ is the sample median of the data set. We can calculate this using the function **median**. We could calculate the first quartile $\hat{q}_{0.25}$ as the median of the ordered data that are at the median or below. The third quartile $\hat{q}_{0.75}$ would be calculated as the median of the data that are at $\hat{q}_{0.5}$ or above. The following MATLAB code illustrates these concepts.

```
% Generate the random sample and sort.
x = sort(rand(1,100));
% Find the median of the lower half - first quartile.
q1 = median(x(1:50));
% Find the median.
q2 = median(x);
```

```
% Find the median of the upper half - third quartile.
q3 = median(x(51:100));
```

The quartiles obtained from this random sample are:

```
q1 = 0.29, q2 = 0.53, q3 = 0.79
```

The theoretical quartiles for the uniform distribution are $q_{0.25} = 0.25$, $q_{0.5} = 0.5$, and $q_{0.75} = 0.75$. So we see that the estimates seem reasonable.
❑

Equation 3.44 provides one way to estimate the quantiles from a random sample. In some situations, we might need to determine an estimate of a quantile that does not correspond to $(j - 0.5)/n$. For instance, this is the case when we are constructing q-q plots (see Chapter 5), and the sample sizes differ. We can use interpolation to find estimates of quantiles that are not represented by Equation 3.44.

### Example 3.6

The MATLAB function **interp1** (in the standard package) returns the interpolated value $Y_I$ at a given $X_I$, based on some observed values $X_{obs}$ and $Y_{obs}$. The general syntax is

```
yint = interp1(xobs, yobs, xint);
```

In our case, the argument of $F^{-1}$ in Equation 3.44 represents the observed values $X_{obs}$, and the order statistics $X_{(j)}$ correspond to the $Y_{obs}$. The MATLAB code for this procedure is shown below.

```
% First generate some standard normal data.
x = randn(500,1);
% Now get the order statistics. These will serve
% as the observed values for the ordinate (Y_obs).
xs = sort(x);
% Now get the observed values for the abscissa (X_obs).
n=length(x);
phat = ((1:n)-0.5)/n;
% We want to get the quartiles.
p = [0.25, 0.5, 0.75];
% The following provides the estimates of the quartiles
% using linear interpolation.
qhat = interp1(phat,xs,p);
```

The resulting estimates are

```
qhat = -0.6928    0.0574    0.6453.
```

The reader is asked to explore this further in the exercises.
❑

## 3.6 MATLAB Code

The MATLAB Statistics Toolbox has functions for calculating the maximum likelihood estimates for most of the common distributions, including the gamma and the Weibull distributions. It is important to remember that the parameters estimated for some of the distributions (e.g., exponential and gamma) are different from those defined in Chapters 2 and 3. We refer the reader to Appendix E for a complete list of the functions appropriate to this chapter. Table 3.2 provides a partial list of MATLAB functions for calculating statistics.We also provide some functions for statistics with the Computational Statistics Toolbox. These are summarized in Table 3.3.

TABLE 3.2

List of MATLAB functions for calculating statistics

| Purpose | MATLAB Function |
|---|---|
| These functions are available in the standard MATLAB package. | `mean`<br>`var`<br>`std`<br>`cov`<br>`median`<br>`corrcoef`<br>`max, min`<br>`sort` |
| These functions for calculating descriptive statistics are available in the MATLAB Statistics Toolbox. | `harmmean`<br>`iqr`<br>`kurtosis`<br>`mad`<br>`moment`<br>`prctile`<br>`range`<br>`skewness`<br>`trimmean` |
| These MATLAB Statistics Toolbox functions provide the maximum likelihood estimates for distributions. | `betafit`<br>`binofit`<br>`expfit`<br>`gamfit`<br>`normfit`<br>`poissfit`<br>`weibfit`<br>`unifit`<br>`mle` |

TABLE 3.3

List of Functions from Chapter 3 Included in the Computational Statistics Toolbox

| Purpose | MATLAB Function |
|---|---|
| These functions are used to obtain parameter estimates for a distribution. | `csbinpar`<br>`csexpar`<br>`csgampar`<br>`cspoipar`<br>`csunipar` |
| These functions return the quantiles. | `csbinoq`<br>`csexpoq`<br>`csunifq`<br>`csweibq`<br>`csnormq`<br>`csquantiles` |
| Other descriptive statistics | `csmomentc`<br>`cskewness`<br>`cskurtosis`<br>`csmoment`<br>`csecdf` |



## 3.7 Further Reading

Many books discuss sampling distributions and parameter estimation. These topics are covered at an undergraduate level in most introductory statistics books for engineers or non-statisticians. For the advanced undergraduate and beginning graduate student, we recommend the text on mathematical statistics by Hogg and Craig [1978]. Another excellent introductory book on mathematical statistics that contains many applications and examples is written by Mood, Graybill and Boes [1974]. Other texts at this same level include Bain and Engelhardt [1992], Bickel and Doksum [2001], and Lindgren [1993]. For the reader interested in the theory of point estimation on a more advanced graduate level, the book by Lehmann and Casella [1998] and Lehmann [1994] are classics.

Most of the texts already mentioned include descriptions of other methods (Bayes methods, minimax methods, Pitman estimators, etc.) for estimating parameters. For an introduction to robust estimation methods, see the books by Wilcox [1997], Launer and Wilkinson [1979], Huber [1981], or Rousseeuw and Leroy [1987] or see the survey paper by Hogg [1974]. Finally, the text by

Keating, Mason and Sen [1993] provides an introduction to Pitman's measure of closeness as a way to assess the performance of competing estimators.

## Exercises

3.1. Generate 500 random samples from the standard normal distribution for sample sizes of $n$ = 2, 15, and 45. At each sample size, calculate the sample mean for all 500 samples. How are the means distributed as $n$ gets large? Look at a histogram of the sample means to help answer this question. What is the mean and variance of the sample means for each $n$? Is this what you would expect from the Central Limit Theorem? Here is some MATLAB code to get you started.

For each $n$:

```
% Generate 500 random samples of size n:
x = randn(n, 500);
% Get the mean of each sample:
xbar = mean(x);
% Do a histogram with superimposed normal density.
% This function is in the MATLAB Statistics Toolbox.
% If you do not have this, then just use the
% function hist instead of histfit.
histfit(xbar);
```

3.2. Repeat problem 3.1 for random samples drawn from a uniform distribution. Use the MATLAB function **rand** to get the samples.

3.3. We have two unbiased estimators $T_1$ and $T_2$ of the parameter $\theta$. The variances of the estimators are given by $V(T_2) = 8$ and $V(T_1) = 4$. What is the MSE of the estimators? Which estimator is better and why? What is the relative efficiency of the two estimators?

3.4. Repeat Example 3.1 using different sample sizes. What happens to the coefficient of skewness and kurtosis as the sample size gets large?

3.5. Repeat Example 3.1 using samples generated from a standard normal distribution. You can use the MATLAB function **randn** to generate your samples. What happens to the coefficient of skewness and kurtosis as the sample size gets large?

3.6. Generate a random sample that is uniformly distributed over the interval $(0, 1)$. Plot the empirical distribution function over the interval (-0.5, 1.5). There is also a function in the Statistics Toolbox called **cdfplot** that will do this.

3.7. Generate a random sample of size 100 from a normal distribution with mean 10 and variance of 2 (use **randn(1,100)*sqrt(2)+10**). Plot the empirical cumulative distribution function. What is the value of the empirical distribution function evaluated at a point less than

the smallest observation in your random sample? What is the value of the empirical cumulative distribution function evaluated at a point that is greater than the largest observation in your random sample?

3.8. Generate a random sample of size 100 from a normal distribution. What are the estimated quartiles?

3.9. Generate a random sample of size 100 from a uniform distribution (use the MATLAB function **rand** to generate the samples). What are the sample quantiles for $p = 0.33, 0.40, 0.63, 0.90$? Is this what you would expect from theory?

3.10. Write a MATLAB function that will return the sample quartiles based on the general definition given for sample quantiles (Equation 3.44).

3.11. Repeat Examples 3.5 and 3.6 for larger sample sizes. Do your estimates for the quartiles get closer to the theoretical values?

3.12. Derive the median for an exponential random variable.

3.13. Calculate the quartiles for the exponential distribution.

3.14. Compare the values obtained for the estimated quartiles in Example 3.6 with the theoretical quantities. You can find the theoretical quantities using **norminv**. Increase the sample size to $n = 1000$. Does your estimate get better?

3.15. Another measure of skewness, called the ***quartile coefficient of skewness***, for a sample is given by

$$\hat{\gamma}_{1_q} = \frac{\hat{q}_{0.75} - 2\hat{q}_{0.5} + \hat{q}_{0.25}}{\hat{q}_{0.75} - \hat{q}_{0.25}}.$$

Write a MATLAB function that returns this statistic.

3.16. Investigate the bias in the maximum likelihood estimate of the variance that is given in Equation 3.28. Generate a random sample from the standard normal distribution. You can use the **randn** function that is available in the standard MATLAB package. Calculate $\hat{\sigma}^2$ using Equation 3.28 and record the value in a vector. Repeat this process (generate a random sample from the standard normal distribution, estimate the variance, save the value) many times. Once you are done with this procedure, you should have many estimates for the variance. Take the mean of these estimates to get an estimate of the expected value of $\hat{\sigma}^2$. How does this compare with the known value of $\sigma^2 = 1$? Does this indicate that the maximum likelihood estimate for the variance is biased? What is the estimated bias from this procedure?