The purpose of an exploration of data may be rather limited and ad hoc, or the purpose may be more general, perhaps to gain understanding of some natural phenomenon. The questions addressed may be somewhat open-ended. The process of understanding often begins with general questions about the structure of the data. At any stage of the analysis, our understanding is facilitated by means of a model.

A model is a description that embodies our current understanding of a phenomenon. In an operational sense, we can formulate a model either as a description of a data-generating process, or as a prescription for processing data. The model is often expressed as a set of equations that relate data elements to each other. It may include probability distributions for the data elements. If any of the data elements are considered to be realizations of random variables, the model is a stochastic model.

A model should not limit our analysis; rather, the model should be able to evolve. The process of understanding involves successive refinements of the model. The refinements proceed from vague models to more specific ones. An exploratory data analysis may begin by mining the data to identify interesting properties. These properties generally raise questions that are to be explored further.

A family of models may have a common form within which the members of the family are distinguished by values of parameters. For example, the family of normal probability distributions has a single form of a probability density function that has two parameters. If this form of model is chosen to represent the properties of a dataset, we may seek confidence intervals for values of the two parameters or perform statistical tests of hypothesized values of the parameters.

In models that are not as mathematically tractable as the normal probability model—and many realistic models are not—computationally intensive methods involving simulations, resamplings, and multiple views may be used to make inferences about the parameters of a model.
1.1 Discovering Structure: Data Structures and Structure in Data

The components of statistical datasets are “observations” and “variables”. In general, “data structures” are ways of organizing data to take advantage of the relationships among the variables constituting the dataset. Data structures may express hierarchical relationships, crossed relationships (as in “relational” databases), or more complicated aspects of the data (as in “object-oriented” databases).

In data analysis, “structure in the data” is of interest. Structure in the data includes such nonparametric features as modes, gaps, or clusters in the data, the symmetry of the data, and other general aspects of the shape of the data. Because many classical techniques of statistical analysis rely on an assumption of normality of the data, the most interesting structure in the data may be those aspects of the data that deviate most from normality.

Sometimes, it is possible to express the structure in the data in terms of mathematical models. Prior to doing this, graphical displays may be used to discover qualitative structure in the data. Patterns observed in the data may suggest explicit statements of the structure or of relationships among the variables in the dataset. The process of building models of relationships is an iterative one, and graphical displays are useful throughout the process. Graphs comparing data and the fitted models are used to refine the models.

Multiple Analyses and Multiple Views

Effective use of graphics often requires multiple views. For multivariate data, plots of individual variables or combinations of variables can be produced quickly and used to get a general idea of the properties of the data. The data should be inspected from various perspectives. Instead of a single histogram to depict the general shape of univariate data, for example, multiple histograms with different bin widths and different bin locations may provide more insight.

Sometimes, a few data points in a display can completely obscure interesting structure in the other data points. A zooming window to restrict the scope of the display and simultaneously restore the scale to an appropriate viewing size can reveal structure. A zooming window can be used with any graphics software whether the software supports it or not; zooming can be accomplished by deletion of the points in the dataset outside of the window.

Scaling the axes can also be used effectively to reveal structure. The relative scale is called the “aspect ratio”. In Figure 1.1, which is a plot of a bivariate dataset, we form a zooming window that deletes a single observation. The greater magnification and the changed aspect ratio clearly show a relationship between $X$ and $Y$ in a region close to the origin that may not hold for the full range of data. A simple statement of this relationship, however, would not extrapolate outside the window to the outlying point.
The use of a zooming window is not “deletion of outliers”; it is focusing in on a subset of the data and is done independently of whatever is believed about the data outside of the window.

One type of structure that may go undetected is that arising from the order in which the data were collected. For data that are recognized as a time series by the analyst, this is obviously not a problem, but often there is a time dependency in the data that is not recognized immediately. “Time” or “location” may not be an explicit variable on the dataset, even though it may be an important variable. The index of the observation within the dataset may be a surrogate variable for time, and characteristics of the data may vary as the index varies. Often it is useful to make plots in which one axis is the index number of the observations. More subtle time dependencies are those in which the values of the variables are not directly related to time, but relationships among variables are changing over time. The identification of such time dependencies is much more difficult, and often requires fitting a model and plotting residuals. Another strictly graphical way of observing changes in relationships over time is by using a sequence of graphical displays.

**Simple Plots May Reveal the Unexpected**

A simple plot of the data will often reveal structure or other characteristics of the data that numerical summaries do not.
An important property of data that is often easily seen in a graph is the unit of measurement. Data on continuous variables are often rounded or measured on a coarse grid. This may indicate other problems in the collection of the data. The horizontal lines in Figure 1.2 indicate that the data do not come from a continuous distribution. Whether we can use methods of data analysis that assume continuity depends on the coarseness of the grid or of the measurement; that is, on the extent to which the data are discrete or the extent to which they have been discretized.

![Figure 1.2. Discrete Data, Rounded Data, or Data Measured Imprecisely](image)

1.2 Modeling and Computational Inference

The process of building models involves successive refinements. The evolution of the models proceeds from vague, tentative models to more complete ones, and our understanding of the process being modeled grows in this process.

The usual statements about statistical methods regarding bias, variance, and so on are made in the context of a model. It is not possible to measure bias or variance of a procedure to select a model, except in the relatively simple case
of selection from some well-defined and simple set of possible models. Only within the context of rigid assumptions (a “metamodel”) can we do a precise statistical analysis of model selection. Even the simple cases of selection of variables in linear regression analysis under the usual assumptions about the distribution of residuals (and this is a highly idealized situation) present more problems to the analyst than are generally recognized.

**Probability Models**

Some of the simplest models used in statistics are probability models for random variables. For a random variable $X$, the model specifies the probability that $X$ is in a given set of real numbers (specifically, a Borel set, but we will not dwell on technical details here). This probability can be expressed in terms of the probability that the random variable is less than or equal to a given number; hence, the fundamental function in a probability model is the cumulative distribution function, or CDF, which for the random variable $X$ yields

$$\Pr(X \leq x)$$

for any real number $x$. We often denote the CDF of $X$ as $P_X(x)$ or $F_X(x)$.

It is clear that the CDF $P_X(x)$ is nondecreasing in $x$, that $P_X(x) \to 0$ as $x \to -\infty$, that $P_X(x) \to 1$ as $x \to \infty$, and that $P_X(x)$ is continuous from the right.

Another very important property of the CDF is that if the random variable $X$ has CDF $P_X(x)$, then

$$P_X(X) \sim U(0, 1),$$

(1.1)

where the symbol “$\sim$” means “is distributed as”, and $U(0, 1)$ represents the uniform distribution over the interval $(0, 1)$.

We call the derivative of a CDF $P_X$ the probability density function, or PDF, and we often denote it by the corresponding lower-case letter, $p_X$ or $f_X$. (If the CDF is not differentiable in the usual sense, the PDF is a special type of derivative, and may denote the probability of a single point.)

There are several probability models that are useful over various ranges of applications. We call a probability model a “distribution”. Two common ones are the normal model, which we denote as $N(\mu, \sigma^2)$, and the uniform model mentioned above. Notation of the form $N(\mu, \sigma^2)$ may denote a specific distribution (if $\mu$ and $\sigma^2$ are assumed fixed), or it may denote the family of distributions. The concept of families of distributions is pervasive in statistics; the simple tasks of statistical inference often involve inference about specific values of the parameters in a given (assumed) family of distributions.

Given any distribution, transformations of the random variable can be used to form various families of related distribution. For example, if $X$ is a random variable, and $Y = aX + b$, with $a \neq 0$, the family of distributions of all such $Y$ is called a location-scale family. Many common families of distributions, for example the normal family, are location-scale families.
Often in statistics, we suppose that we have a random sample from a particular distribution. A random sample is the same as a set of random variables that are independent and have the same distribution; that is, the random variables are independent and identically distributed, or iid. We often use notation of the form $X_1, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)$, for example, to indicate that the $n$ random variables are independent and identically normally distributed.

**Descriptive Statistics, Inferential Statistics, and Model Building**

We can distinguish statistical activities that involve:

- data collection;
- descriptions of a given dataset;
- inference within the context of a model or family of models; and
- model selection.

In any given application, it is likely that all of these activities will come into play. Sometimes (and often, ideally!), a statistician can specify how data are to be collected, either in surveys or in experiments. We will not be concerned with this aspect of the process in this text.

Once data are available, either from a survey or designed experiment, or just observational data, a statistical analysis begins by considering general descriptions of the dataset. These descriptions include ensemble characteristics, such as averages and spreads, and identification of extreme points. The descriptions are in the form of various summary statistics and graphical displays. The descriptive analyses may be computationally intensive for large datasets, especially if there are a large number of variables. The computationally intensive approach also involves multiple views of the data, including consideration of various transformations of the data. We discuss these methods in Chapters 5 and 7 and in Part II.

A stochastic model is often expressed as a probability density function or as a cumulative distribution function of a random variable. In a simple linear regression model with normal errors,

$$Y = \beta_0 + \beta_1 x + E,$$

for example, the model may be expressed by use of the probability density function for the random variable $E$. (Notice that $Y$ and $E$ are written in uppercase because they represent random variables.) The probability density function for $Y$ is

$$p(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(y-\beta_0-\beta_1 x)^2/(2\sigma^2)}.$$

In this model, $x$ is an observable covariate; $\sigma$, $\beta_0$, and $\beta_1$ are unobservable (and, generally, unknown) parameters; and 2 and $\pi$ are constants. Statistical inference about parameters includes estimation or tests of their values or
1.2 The Role of the ECDF in Inference

statements about their probability distributions based on observations of the elements of the model.

The elements of a stochastic model include observable random variables, observable covariates, unobservable parameters, and constants. Some random variables in the model may be considered to be “responses”. The covariates may be considered to affect the response; they may or may not be random variables. The parameters are variable within a class of models, but for a specific data model the parameters are constants. The parameters may be considered to be unobservable random variables, and in that sense, a specific data model is defined by a realization of the parameter random variable. In the model, written as

\[ Y = f(x; \beta) + E, \quad (1.4) \]

we identify a “systematic component”, \( f(x; \beta) \), and a “random component”, \( E \). The selection of an appropriate model may be very difficult, and almost always involves not only questions of how well the model corresponds to the observed data, but also the tractability of the model. The methods of computational statistics allow a much wider range of tractability than can be contemplated in mathematical statistics.

Statistical analyses generally are undertaken with the purpose of making a decision about a dataset, about a population from which a sample dataset is available, or in making a prediction about a future event. Much of the theory of statistics developed during the middle third of the twentieth century was concerned with formal inference; that is, use of a sample to make decisions about stochastic models based on probabilities that would result if a given model was indeed the data-generating process. The heuristic paradigm calls for rejection of a model if the probability is small that data arising from the model would be similar to the observed sample. This process can be quite tedious because of the wide range of models that should be explored and because some of the models may not yield mathematically tractable estimators or test statistics. Computationally intensive methods include exploration of a range of models, many of which may be mathematically intractable.

In a different approach employing the same paradigm, the statistical methods may involve direct simulation of the hypothesized data-generating process rather than formal computations of probabilities that would result under a given model of the data-generating process. We refer to this approach as computational inference. We discuss methods of computational inference in Chapters 2, 3, and 4. In a variation of computational inference, we may not even attempt to develop a model of the data-generating process; rather, we build decision rules directly from the data. This is often the approach in clustering and classification, which we discuss in Chapter 10. Computational inference is rooted in classical statistical inference. In subsequent sections of the current chapter, we discuss general techniques used in statistical inference.
1.3 The Role of the Empirical Cumulative Distribution Function

Methods of statistical inference are based on an assumption (often implicit) that a discrete uniform distribution with mass points at the observed values of a random sample is asymptotically the same as the distribution governing the data-generating process. Thus, the distribution function of this discrete uniform distribution is a model of the distribution function of the data-generating process.

For a given set of univariate data, \( y_1, \ldots, y_n \), the empirical cumulative distribution function, or ECDF, is

\[
P_n(y) = \frac{\# \{y_i, \text{s.t. } y_i \leq y\}}{n}.
\]  

(1.5)

The ECDF is the basic function used in many methods of computational inference.

Although the ECDF has similar definitions for univariate and multivariate random variables, it is most useful in the univariate case. An equivalent expression for univariate random variables, in terms of intervals on the real line, is

\[
P_n(y) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,y]}(y_i),
\]  

(1.6)

where \( I \) is the indicator function. (See page 413 for the definition and some of the properties of the indicator function. The measure \( dI_{(-\infty,a]}(x) \), which we use in equation (1.14) below, is particularly interesting.)

It is easy to see that the ECDF is pointwise unbiased for the CDF; that is, if the \( y_i \) are independent realizations of random variables \( Y_i \), each with CDF \( P(\cdot) \), for a given \( y \),

\[
E(P_n(y)) = E\left(\frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,y]}(Y_i)\right)
= \frac{1}{n} \sum_{i=1}^{n} E\left(I_{(-\infty,y]}(Y_i)\right)
= \text{Pr}(Y \leq y)
= P(y).
\]  

(1.7)

Similarly, we find

\[
V(P_n(y)) = P(y)(1 - P(y)) / n;
\]  

(1.8)

indeed, at a fixed point \( y \), \( nP_n(y) \) is a binomial random variable with parameters \( n \) and \( \pi = P(y) \). Because \( P_n \) is a function of the order statistics, which form a complete sufficient statistic for \( P \), there is no unbiased estimator of \( P(y) \) with smaller variance.

1.3 The Role of the ECDF in Inference

We also define the *empirical probability density function* (EPDF) as the derivative of the ECDF:

\[ p_n(y) = \frac{1}{n} \sum_{i=1}^{n} \delta(y - y_i), \]

where \( \delta \) is the Dirac delta function. The EPDF is just a series of spikes at points corresponding to the observed values. It is not as useful as the ECDF. It is, however, unbiased at any point for the probability density function at that point.

The ECDF and the EPDF can be used as estimators of the corresponding population functions, but there are better estimators (see Chapter 9).

### 1.3.1 Statistical Functions of the CDF and the ECDF

In many models of interest, a parameter can be expressed as a functional of the probability density function or of the cumulative distribution function of a random variable in the model. The mean of a distribution, for example, can be expressed as a functional \( \Theta \) of the CDF \( P \):

\[ \Theta(P) = \int_{\mathbb{R}^d} y \, dP(y). \]

A functional that defines a parameter is called a *statistical function*.

#### Estimation of Statistical Functions

A common task in statistics is to use a random sample to estimate the parameters of a probability distribution. If the statistic \( T \) from a random sample is used to estimate the parameter \( \theta \), we measure the performance of \( T \) by the magnitude of the bias,

\[ |E(T) - \theta|, \]

by the variance,

\[ V(T) = E \left( (T - E(T)) (T - E(T))^T \right), \]

by the mean squared error,

\[ E \left( (T - \theta)^T (T - \theta) \right), \]

and by other expected values of measures of the distance from \( T \) to \( \theta \). (These expressions above are for the scalar case, but similar expressions apply to vectors \( T \) and \( \theta \), in which case the bias is a vector, the variance is the variance-covariance matrix, and the mean squared error is a dot product and hence a scalar.)
If \( E(T) = \theta \), \( T \) is unbiased for \( \theta \). For sample size \( n \), if \( E(T) = \theta + O(n^{-1/2}) \), \( T \) is said to be first-order accurate for \( \theta \); if \( E(T) = \theta + O(n^{-1}) \), it is second-order accurate. (See page 414 for the definition of \( O(\cdot) \). Convergence of \( E(T) \) can also be expressed as a stochastic convergence of \( T \), in which case we use the notation \( O_{\mathbb{P}}(\cdot) \).)

The order of the mean squared error is an important characteristic of an estimator. For good estimators of location, the order of the mean squared error is typically \( O(n^{-1}) \). Good estimators of probability densities, however, typically have mean squared errors of at least order \( O(n^{-4/5}) \) (see Chapter 9).

**Estimation Using the ECDF**

There are many ways to construct an estimator and to make inferences about the population. In the univariate case especially, we often use data to make inferences about a parameter by applying the statistical function to the ECDF. An estimator of a parameter that is defined in this way is called a plug-in estimator. A plug-in estimator for a given parameter is the same functional of the ECDF as the parameter is of the CDF.

For the mean of the model, for example, we use the estimate that is the same functional of the ECDF as the population mean in equation (1.10),

\[
\Theta(P_n) = \int_{-\infty}^{\infty} y \, dP_n(y) = \int_{-\infty}^{\infty} y \, d\left(\frac{1}{n} \sum_{i=1}^{n} I(-\infty, y_i]\right) \\
= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} y \, dI(-\infty, y_i] \\
= \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}.
\]

The sample mean is thus a plug-in estimator of the population mean. Such an estimator is called a method of moments estimator. This is an important type of plug-in estimator. The method of moments results in estimates of the parameters \( E(Y^r) \) that are the corresponding sample moments.

Statistical properties of plug-in estimators are generally relatively easy to determine. In some cases, the statistical properties, such as expectation and variance, are optimal in some sense.

In addition to estimation based on the ECDF, other methods of computational statistics make use of the ECDF. In some cases, such as in bootstrap methods, the ECDF is a surrogate for the CDF. In other cases, such as Monte Carlo methods, an ECDF for an estimator is constructed by repeated sampling, and that ECDF is used to make inferences using the observed value of the estimator from the given sample.
Viewed as a statistical function, \( \Theta \) denotes a specific functional form. Any functional of the ECDF is a function of the data, so we may also use the notation \( \Theta(Y_1, \ldots, Y_n) \). Often, however, the notation is cleaner if we use another letter to denote the function of the data; for example, \( T(Y_1, \ldots, Y_n) \), even if it might be the case that
\[
T(Y_1, \ldots, Y_n) = \Theta(P_n).
\]

We will also often use the same letter that denotes the functional of the sample to represent the random variable computed from a random sample; that is, we may write
\[
T = T(Y_1, \ldots, Y_n).
\]

As usual, we will use \( t \) to denote a realization of the random variable \( T \).

Use of the ECDF in statistical inference does not require many assumptions about the distribution. Other methods discussed below are based on information or assumptions about the data-generating process.

### 1.3.2 Order Statistics

In a set of iid random variables \( X_1, \ldots, X_n \), it is often of interest to consider the ranked values \( X_{i_1} \leq \cdots \leq X_{i_n} \). These are called the order statistics and are denoted as \( X_{(1:n)}, \ldots, X_{(n:n)} \). For \( 1 \leq k \leq n \), we refer to \( X_{(k:n)} \) as the \( k \)th order statistic. We often use the simpler notation \( X_k \), assuming that \( n \) is some fixed and known value. Also, we sometimes drop the parentheses in the other representation, \( X_{k:n} \).

If the CDF of the \( n \) iid random variables is \( P(x) \) and the PDF is \( p(x) \), we can get the PDF of the \( k \)th order statistic by forming the joint density, and integrating out all variables except the \( k \)th order statistic. This yields
\[
p_{X_{(k:n)}}(x) = \binom{n}{k} (P(x))^{k-1} (1 - P(x))^{n-k} p(x).
\]

Clearly, the order statistics \( X_{(1:n)}, \ldots, X_{(n:n)} \) from an iid sample of random variables \( X_1, \ldots, X_n \) are neither independent nor identically distributed.

From equation (1.15), and the fact that if the random variable \( X \) has CDF \( P \), then \( P(X) \sim U(0,1) \) (expression (1.1) on page 9), we see that the distribution of the \( k \)th order statistic from a \( U(0,1) \) is the beta distribution with parameters \( k \) and \( n - k + 1 \); that is,
\[
X_{(k:n)} \sim \text{beta}(k, n - k + 1),
\]
if \( X_{(k:n)} \) is the \( k \)th order statistic in a sample of size \( n \) from a uniform\((0,1)\) distribution.

Another important fact about order statistics from a uniform distribution is that they have simple conditional distributions. Given the \((k+1)\)th order

*Elements of Computational Statistics, Second Edition ©2013 James E. Gentle*
Basic Computations in Statistical Inference

Statistic, $U_{k+1:n} = v$, the conditional joint distribution of $vU_{1:k}, \ldots vU_{k:k}$ is the same as the joint distribution of $U_{1:n}, \ldots U_{k:n}$; that is,

$$(vU_{1:k}, \ldots vU_{k:k}) \overset{d}{=} (U_{1:n}, \ldots U_{k:n}). \quad (1.17)$$

(See Exercise 1.7.)

**Quantiles**

For $\alpha \in (0, 1)$, the $\alpha$ quantile of the distribution with CDF $P$ is the value $x_{(\alpha)}$ such that $P(x_{(\alpha)}) = \alpha$ if such a value exists. (For a univariate random variable, this is a single point. For a $d$-variate random variable, it is a $(d-1)$-dimensional object that is generally nonunique.)

In a discrete distribution, for a given value of $\alpha$, there may be no value $x_{(\alpha)}$ such that $P(x_{(\alpha)}) = \alpha$. We can define a useful concept of “quantile”, however, that always exists. For a univariate probability distribution with CDF $P$, we define the function $P^{-1}$ on the open interval $(0, 1)$ as

$$P^{-1}(\alpha) = \inf\{x, \text{ s.t. } P(x) \geq \alpha\}. \quad (1.18)$$

We call $P^{-1}$ the quantile function. Notice that if $P$ is strictly increasing, the quantile function is the ordinary inverse of the cumulative distribution function. If $P$ is not strictly increasing, the quantile function can be interpreted as a generalized inverse of the cumulative distribution function. This definition is reasonable (at the expense of overloading the notation used for the ordinary inverse of a function) because, while a CDF may not be an invertible function, it is monotonic nondecreasing. Notice that for the univariate random variable $X$ with CDF $P$, if $x_{(\alpha)} = P^{-1}(\alpha)$, then $x_{(\alpha)}$ is the $\alpha$ quantile of $X$ as above.

In a discrete distribution, the quantile function is a step function, and the quantile is the same for values of $\alpha$ in an interval.

The quantile function, just as the CDF, fully determines a probability distribution.

It is clear that $x_{(\alpha)}$ is a functional of the CDF, say $\Xi_{\alpha}(P)$. The functional is very simple. It is

$$\Xi_{\alpha}(P) = P^{-1}(\alpha),$$

where $P^{-1}$ is the quantile function.

If $P(x) = \alpha$, we say the quantile-level of $x$ is $\alpha$. (We also sometimes use the term “quantile” by itself in a slightly different way: if $P(x) = \alpha$, we say the quantile of $x$ is $\alpha$.)

**Empirical Quantiles**

The quantile function $P_n^{-1}$ associated with an ECDF $P$ leads to the order statistics on which the ECDF is based. These quantiles do not correspond
symmetrically to the quantiles of the underlying population. For a quantile-
level $\alpha < 2/n$, $P_n^{-1}(\alpha) = x_{(i:n)}$, then for each increase in $\alpha$ of $1/n$, $P_n^{-1}$
increases to the next order statistic, until finally, it is $x_{(n:n)}$ for the single
limiting value $\alpha = 1$. This likely disconnect between the quantiles of the
ECDF and the quantiles of the underlying distribution leads us to consider
other definitions for the empirical quantile, or sample quantile.

For a given sample from a continuous distribution, the intervals between
successive order statistics are unbiased estimates of intervals of equal proba-
bility. For quantile-levels between $1/n$ and $1 - 1/n$, this leads to estimates of
quantiles that are between two specific order statistics; that is, for $\alpha$ between
$1/n$ and $1 - 1/n$, an unbiased estimate of the $\alpha$ quantile would be between
$x_{i:n}$ and $x_{i+1:n}$, where $i/n \leq \alpha \leq (i + 1)/n$. For $\alpha$ between $1/n$ and $1 - 1/n$,
we therefore define the empirical quantile as a convex linear combination of
two successive order statistics:

$$ q_i = \lambda x_{i:n} + (1 - \lambda)x_{i+1:n}, $$

(1.19)

for $0 \leq \lambda \leq 1$. We define the quantile-level of such a quantile as

$$ p_i = \frac{i - \iota}{n + \nu} $$

(1.20)

for some $\iota \in [0, \frac{1}{2}]$ and some $\nu \in [0, 1]$. This expression is based on a linear
interpolation of the ECDF between the order statistics. The values of $\iota$ and $\nu$
determine the value of $\lambda$ in equation (1.19).

We would like to choose values of $\iota$ and $\nu$ in expression (1.20) that make
the empirical quantiles of a random sample correspond closely to those of the
population depend on the distribution of the population, but of course, those
are generally unknown. A certain symmetry may be imposed by requiring
$\nu = 1 - 2\iota$. For a discrete distribution, it is generally best to take $\iota = \nu = 0$.
A common pair of choices in continuous distributions are $\iota = \frac{1}{2}$ and $\nu = 0$.
Another reasonable pair of choices are $\iota = \frac{3}{8}$ and $\nu = 1/4$. This corresponds
to values that match quantiles of a normal distribution.

Hyndman and Fan (1996) identify nine different versions of the expres-
sion (1.20) that are used in statistical computer packages. In some cases, the
software only uses one version; in other cases, the user is given a choice.

The R function `quantiles` allows the user to choose among nine types,
and for the chosen type returns the empirical quantiles and the quantile-levels.
The R function `ppoints` generates values of $p_i$ as in equation (1.20). For $n$
less that 11, `ppoints` uses the values $\iota = \frac{3}{8}$ and $\nu = 1/4$, and for larger values
of $n$, it uses $\iota = \frac{1}{2}$ and $\nu = 0$.

Empirical quantiles are used in Monte Carlo inference, in nonparametric
inference, and in graphical displays for comparing a sample with a standard
distribution or with another sample.

Empirical quantiles can also be used as estimators of the population quan-
tiles, but there are other estimators for quantiles of a continuous distribution.
Some, such as the Kaigh-Lachenbruch estimator and the Harrell-Davis estimator, use a weighted combination of multiple data points instead of just a single one, as in the simple estimators above. (See Kaigh and Lachenbruch (1982) and Harrell and Davis (1982) and also see Exercise 1.9.)

If a covariate is available, it may be possible to use it to improve the quantile estimate. This is often the case in simulation studies, where the covariate is called a control variable.

1.3.3 q-q Plots

The quantile-quantile plot or q-q plot is a useful graphical display for comparing two distributions, two samples, or a sample and a given theoretical distribution. The most common use is for comparing an observed sample with a given theoretical or reference distribution. Either order statistics of the sample or sample quantiles are plotted along one axis and either expected values of order statistics or quantiles of the reference distribution corresponding to appropriate quantile-levels are plotted along the other axis. The choices have to do with the appropriate definitions of empirical quantiles, as in equation (1.19), or of population quantiles corresponding to quantile-levels, as in equation (1.20). The sample points may be on the vertical axis and the population points on the horizontal axis, or they can be plotted in the other way.

The extent to which the q-q scatterplot fails to lie along a straight line provides a visual assessment of whether the points plotted along the two axes are from distributions or samples whose quantiles are linearly related. The points will fall along a straight line if the two distributions are in the sample location-scale family; that is, if the two underlying random variables have a relationship of the form \( Y = aX + b \), where \( a \neq 0 \).

One of the most common reference distributions, of course, is the normal. The normal family of distributions is a location-scale family, so a q-q plot for a normal reference distribution does not depend on the mean or variance. The R function `qqnorm` plots order statistics from a given sample against normal quantiles.

The R function `qqplot` creates more general q-q plots.

Figure 1.3 shows a q-q plot that compares a sample with two gamma reference distributions. The sample was generated from a gamma(10, 10) distribution, and the order statistics from that sample are plotted along the vertical axes. The reference distribution on the left-hand side in Figure 1.3 is a gamma(10, 1) distribution, which is in the same scale family as a gamma(10, 10); hence, the points fall close to a straight line. This plot was produced by the R statements

```
plot(qgamma(ppoints(length(x)),10),sort(x))
abline(lsfit(qgamma(ppoints(x),10),sort(x)))
```
Note that the “worst fit” is in the tails; that is, near the end points of the plot. This is characteristic of q-q plots and is a result of the greater skewness of the extreme order statistics.

The reference distribution in the plot on the right-hand side in Figure 1.3 is a gamma(1, 1) distribution. The points do not seem to lie on a straight line. The extremes of the sample do not match the quantiles well at all. The pattern that we observe for the smaller observations (that is, that they are below a straight line that fits most of the data) is characteristic of data with a heavier left tail than the reference distribution to which it is being compared. Conversely, the larger observations, being below the straight line, indicate that the data have a lighter right tail than the reference distribution.

An important property of the q-q plot is that its shape is independent of the location and the scale of the data. In Figure 1.3, the sample is from a gamma distribution with a scale parameter of 10, but the distribution quantiles are from a population with a scale parameter of 1.

For a random sample from the distribution against whose quantiles it is plotted, the points generally deviate most from a straight line in the tails. This is because of the larger variability of the extreme order statistics. Also, because the distributions of the extreme statistics are skewed, the deviation from a straight line is in a specific direction (toward lighter tails) more than half of the time (see Exercise 1.10, page 48).

A q-q plot is an informal visual goodness-of-fit test. (There is no “significance level”.) The sup absolute difference between the ECDF and the reference
CDF is the *Kolmogorov distance*, which is the basis for the Kolmogorov test (and the Kolmogorov-Smirnov test) for distributions. The Kolmogorov distance does poorly in measuring differences in the tails of the distribution. A q-q plot, on the other hand, generally is very good in revealing differences in the tails.

A q-q plot is a useful application of the ECDF. As I have mentioned, the ECDF is not so meaningful for multivariate distributions. Plots based on the ECDF for of a multivariate dataset are generally difficult to interpret.

### 1.4 The Role of Optimization in Inference

Important classes of estimators are defined as the point at which some function that involves the parameter and the random variable achieves an optimum. There are, of course, many functions that involve the parameter and the random variable; an example is the probability density.

In the use of function optimization in inference, once the objective function is chosen, observations on the random variable are taken and are then considered to be fixed; the parameter in the function is considered to be a variable (the “decision variable”, in the parlance often used in the literature on optimization). The function is then optimized with respect to the parameter variable. The nature of the function determines the meaning of “optimized”; if the function is the probability density, for example, “optimized” would logically mean “maximized”. (This leads to maximum likelihood estimation, which we discuss below.)

In discussing the use of optimization in statistical estimation, we must be careful to distinguish between a symbol that represents a fixed parameter and a symbol that represents a “variable” parameter. When we denote a probability density function as \( p(y \mid \theta) \), we generally expect “\( \theta \)” to represent a fixed, but possibly unknown, parameter. A family of probability models specifies a parameter space, say \( \Theta \), that determines the possible values of \( \theta \) in that model.

In an estimation method that involves optimizing some function, “\( \theta \)” is often used as a variable placeholder. I prefer to use some other symbol for a variable quantity that serves in the place of the parameter; therefore, in the following discussion, I will generally use “\( t \)” in place of “\( \theta \)” when I want to treat is as a variable. For a family of probability models with parameter space \( \Theta \), we define a set \( T \supseteq \Theta \) and require \( t \in T \). For other common symbols of parameters, such as \( \beta \), when I want to treat them as variables in an optimization algorithm, I use corresponding Latin letters, such as \( b \).

In an iterative algorithm, I use \( t^{(k)} \) to represent a fixed value in the \( k \)th iteration.
1.4 The Role of Optimization in Inference

Some Comments on Optimization

The solution to an optimization problem is in some sense “best” for that particular problem and its objective functions; this may mean, however, that it is considerably less good for some other optimization problem. It is often the case, therefore, that an optimal solution is not robust to assumptions about the phenomenon being studied.

Optimization, because it is performed in the context of a set of assumptions, is likely to magnify the effects of the assumptions.

1.4.1 Fitting Statistical Models by Optimization

A statistical model such as \( Y = \beta_0 + \beta_1 x + E \) in equation (1.2) specifies a family of probability distributions of the data. The specific member of that family of probability distributions depends on the values of the parameters in model, \( \beta_0, \beta_1, \) and \( \sigma^2 \) (see equation (1.3)). Simpler models, such as the statement “\( X \) is distributed as \( N(\mu, \sigma^2) \),” also specify families of probability distributions. An important step in statistical inference is to use observed data to estimate the parameters in the model. This is called “fitting the model.” The question is, how do we estimate the parameters? What properties should the estimators have?

In the following pages we discuss two general ways in which optimization is used in to estimate parameters or to fit models. One is to minimize deviations of observed values from what a model would predict (think “least squares”, as an example). This is an intuitive procedure which may be chosen without regard to the nature of the data-generating process. The justification for a particular form of the objective function, however, may arise from assumptions about a probability distribution underlying the data-generating process. Another way in which optimization is used in statistical inference is in maximizing the “likelihood”, which we will discuss more precisely in Section 1.4.3, beginning on page 30. The correct likelihood function depends on the probability distribution underlying the data-generating process, which, of course, is not known and can only be assumed. How poor the maximum likelihood estimator is depends on both the true distribution and the assumed distribution.

In the discussion below, we briefly describe particular optimization techniques that assume that the objective function is a continuous function of the decision variables, or the parameters. We also assume that there are no a priori constraints on the values of the parameters. Techniques appropriate for other situations, such as for discrete optimization and constrained optimization, are available in the general literature on optimization.

We must also realize that mathematical expressions below do not necessarily imply computational methods. There are many additional considerations for the numerical computations. A standard example of this point is in the solution of the linear full-rank system of \( n \) equations in \( n \) unknowns: \( Ax = b \).
While we may write the solution as \( x = A^{-1} b \), we would almost never compute the solution by forming the inverse and then multiplying \( b \) by it (see Gentle (2007), Chapter 6).

### 1.4.2 Estimation by Minimizing Residuals

In many applications, we can express the expected value of a random variable as a function of a parameter. For example, in the model \( Y = \beta_0 + \beta_1 x + E \) in equation (1.2), we have

\[
E(Y) = \beta_0 + \beta_1 x,
\]

which involves an observable covariate. In general, we may write the expected value as

\[
E(Y) = f(x, \theta),
\]

where \( f \) is some function, \( x \) may be some observable covariates (possibly a vector), and \( \theta \) is some unobservable parameter (possibly a vector). (The more difficult and interesting problems, of course, involve the determination of the form of the function \( f \), but for the time being, we concentrate on the simpler problem of determining an appropriate value of \( \theta \), assuming that the form of the function \( f \) is known.)

Assuming that we have observations \( y_1, \ldots, y_n \) on \( Y \) (and observations on the covariates if there are any), we ask what value of \( \theta \) would make the model fit the data best. There are various ways of approaching this question, but a reasonable first step would be to look at the differences in the observations and their expected (or “predicted”) values:

\[
y_i - f(x_i, \theta).
\]

For any particular value of \( \theta \), say \( t \), we have the residuals,

\[
r_i(t) = y_i - f(x_i, t). \tag{1.23}
\]

A reasonable estimator of \( \theta \) would be the value of \( t \) that minimizes some norm of \( r(t) \), the \( n \)-vector of residuals \( r_i(t) \).

Notice that I am using \( t \) in place of \( \theta \). The logical reason for doing this is that \( \theta \) itself is some unknown constant. I want to work with a variable whose value I can choose according to my own criterion. (While you should understand this reasoning, I will often use the same symbol for the variable that I use for the constant unknown parameter. This sloppiness is common in the statistical literature.)

We have now formulated our estimation problem as an optimization problem:

\[
\min_{t} \| r(t) \|, \tag{1.24}
\]

where \( \| \cdot \| \) represents a vector norm. We often choose the norm as the \( L_p \) norm, which we sometimes represent as \( \| \cdot \|_p \). For the \( n \)-vector \( v \), the \( L_p \) vector norm is
For the $L_p$ norm, we minimize a function of an $L_p$ norm of the residuals,

$$s_p(t) = \sum_{i=1}^{n} |y_i - f(x_i, t)|^p,$$

(1.25)

for some $p \geq 1$, to obtain an $L_p$ estimator. Simple choices are the sum of the absolute values and the sum of the squares. The latter choice yields the least squares estimator.

More generally, we could minimize

$$s_\rho(t) = \sum_{i=1}^{n} \rho(y_i - f(x_i, t))$$

(1.26)

for some nonnegative function $\rho(\cdot)$ to obtain an “M estimator”. (The name comes from the similarity of this objective function to the objective function for some maximum likelihood estimators.)

**Other Types of Residuals**

Recall our original estimation problem. We have a model, say

$$Y = f(x_i, \theta) + E,$$

where $E$ is a random variable with unknown variance, say $\sigma^2$. We have formulated an optimization problem involving residuals formed from the decision variable $t$, for the estimation of $\theta$. What about estimation of $\sigma^2$? To use the same heuristic, that is, to form a residual involving $\sigma^2$, we would use the residuals that involve $t$:

$$g \left( \sum_{i=1}^{n} (r(t))^2 \right) - v^2,$$

(1.27)

where $g$ is some function and where the variable $v^2$ is used in place of $\sigma^2$. This residual may look somewhat unfamiliar, and there is no obvious way of choosing $g$. In certain simple models, however, we can formulate expression $g \left( \sum_{i=1}^{n} (r(t))^2 \right)$ in terms of the expected values of $Y - f(x, \theta)$. In the case of the model (1.2) $Y = \beta_0 + \beta_1 x + E$, this leads to the estimator $v^2 = \sum_{i=1}^{n} (r(t))^2 / (n - 2)$.

Formation of residuals in other estimation problems may not always be so straightforward. For example, consider the problem of estimation of the parameters $\alpha$ and $\beta$ in a beta($\alpha$, $\beta$) distribution. If the random variable $Y$ has
this distribution, then \( E(Y) = \alpha / (\alpha + \beta) \). Given observations \( y_1, \ldots, y_n \), we could form residuals similar to what we did above:

\[
r_{1i}(a, b) = y_i - a / (a + b).
\]

Minimizing a norm of this, however, does not yield unique individual values for \( a \) and \( b \).

We need other, mathematically independent, residuals. ("Mathematically independent" refers to unique solutions of equations; not statistical independence of random variables.) If the random variable \( Y \) has the beta distribution as above, then we know

\[
E(Y^2) = V(Y^2) + (E(Y))^2 = \frac{\alpha^2(\alpha + \beta) + \alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.
\]

We could now form a second set of residuals:

\[
r_{2i}(a, b) = y_i^2 - \left( \frac{a^2(a + b) + ab}{(a + b)^2(a + b + 1)} \right).
\]

We now form a minimization problem involving these two sets of residuals.

There are, of course, various minimization problems that could be formulated. The most direct problem, analogous to (1.24), is

\[
\min_{a,b} \left( \|r_{1}(a, b)\| + \|r_{2}(a, b)\| \right).
\]

Another optimization problem is the sequential one:

\[
\min_{a,b} \|r_{1}(a, b)\|, \tag{1.28}
\]

and then, subject to a conditional minimum,

\[
\min_{\hat{a}, \hat{b}} \|r_{2}(\hat{a}, \hat{b})\|. \tag{1.29}
\]

In Exercise 1.13a, you are asked to estimate the parameters in a gamma distribution using these ideas.

**Computational Methods for Estimation by Minimizing Residuals**

Standard techniques for optimization can be used to determine estimates that minimize various functions of the residuals, that is, for some appropriate function of the residuals \( s(\cdot) \), to solve

\[
\min_{t} s(t). \tag{1.30}
\]

Here, I am thinking of the model parameter \( \theta \), and so I am using \( t \) as a variable in place of \( \theta \). I will denote the optimal value of \( t \), according to the criteria embodied in the the function \( s(\cdot) \), as \( \hat{\theta} \).

*Elements of Computational Statistics, Second Edition ©2013 James E. Gentle*
1.4 The Role of Optimization in Inference

Except for special forms of the objective function, the algorithms to solve expression (1.30) are iterative. If $s$ is twice differentiable, one algorithm is Newton’s method, in which the minimizing value of $t$, $\hat{\theta}$, is obtained as a limit of the iterates

$$
t^{(k)} = t^{(k-1)} - \left( H_s(t^{(k-1)}) \right)^{-1} \nabla s(t^{(k-1)}),
$$

(1.31)

where $H_s(t)$ denotes the Hessian of $s$ and $\nabla s(t)$ denotes the gradient of $s$, both evaluated at $t$. (Newton’s method is sometimes called the Newton-Raphson method.)

The function $s(\cdot)$ is usually chosen to be differentiable, at least piecewise.

For various computational considerations, instead of the exact Hessian, a matrix $\tilde{H}_s$ approximating the Hessian is often used. In this case, the technique is called a quasi-Newton method.

Newton’s method or a quasi-Newton method often overshoots the best step. The direction $t^{(k)} - t^{(k-1)}$ may be the best direction, but the distance $\|t^{(k)} - t^{(k-1)}\|$ may be too great. A variety of methods using Newton-like iterations involve a system of equations of the form

$$
\tilde{H}_s(t) d = -\nabla s(t).
$$

(1.32)

These equations are solved for the direction $d$, and the new point is taken as the old $t$ plus $\alpha d$, for some damping factor $\alpha$.

There are various ways of deciding when an iterative optimization algorithm has converged. In general, convergence criteria are based on the size of the change in $t^{(k)}$ from $t^{(k-1)}$, or the size of the change in $s(t^{(k)})$ from $s(t^{(k-1)})$.

**Statistical Properties of Minimum-Residual Estimators**

It is generally difficult to determine the variance or other high-order statistical properties of an estimator defined as above (that is, defined as the minimizer of some function of the residuals). In many cases, all that is possible is to approximate the variance of the estimator in terms of some relationship that holds for a normal distribution. (In robust statistical methods, for example, it is common to see a “scale estimate” expressed in terms of some mysterious constant times a function of some transformation of the residuals.)

There are two issues that affect both the computational method and the statistical properties of the estimator defined as the solution to the optimization problem. One consideration has to do with the acceptable values of the...
parameter $\theta$. In order for the model to make sense, it may be necessary that the parameter be in some restricted range. In some models, a parameter must be positive, for example. In these cases, the optimization problem has constraints. Such a problem is more difficult to solve than an unconstrained problem. Statistical properties of the solution are also more difficult to determine. More extreme cases of restrictions on the parameter may require the parameter to take values in a countable set. Obviously, in such cases, Newton’s method cannot be used because the derivatives cannot be defined. In those cases, a combinatorial optimization algorithm must be used instead. Other situations in which the function is not differentiable also present problems for the optimization algorithm. In such cases, if the domain is continuous, a descending sequence of simplexes can be used.

Secondly, it may turn out that the optimization problem (1.30) has local minima. This depends on the nature of the function $f(\cdot)$ in equation (1.22). Local minima present problems for the computation of the solution because the algorithm may get stuck in a local optimum. Local minima also present conceptual problems concerning the appropriateness of the estimation criterion itself. As long as there is a unique global optimum, it seems reasonable to seek it and to ignore local optima. It is not so clear what to do if there are multiple points at which the global optimum is attained.

**Least Squares Estimation**

Least squares estimators are generally more tractable than estimators based on other functions of the residuals. They are more tractable both in terms of solving the optimization problem to obtain the estimate and in approximating statistical properties of the estimators, such as their variances.

Consider the model in equation (1.22), $E(Y) = f(x, \theta)$, and assume that $\theta$ is an $m$-vector and that $f(\cdot)$ is a smooth function in $\theta$. Letting $y$ be the $n$-vector of observations, we can write the least squares objective function corresponding to equation (1.25) as

$$s(t) = (r(t))^T r(t), \quad (1.33)$$

where the superscript $T$ indicates the transpose of a vector or matrix.

The gradient and the Hessian for a least squares problem have special structures that involve the Jacobian of the residuals, $J_r(t)$. The gradient of $s$ is

$$\nabla s(t) = 2(J_r(t))^T r(t). \quad (1.34)$$

Taking derivatives of $\nabla s(t)$, we see that the Hessian of $s$ can be written in terms of the Jacobian of $r$ and the individual residuals:

$$H_s(t) = 2(J_r(t))^T J_r(t) + 2 \sum_{i=1}^n r_i(t)H_{r_i}(t). \quad (1.35)$$
In the vicinity of the solution \( \hat{\theta} \), the residuals \( r_i(t) \) should be small, and \( \mathbf{H}_r(t) \) may be approximated by neglecting the second term:

\[
\mathbf{H}_r(t) \approx 2(\mathbf{J}_r(t))^T \mathbf{J}_r(t).
\]

Using equation (1.34) and this approximation for equation (1.35) in the gradient descent equation (1.32), we have the system of equations

\[
(\mathbf{J}_r(t(k-1)))^T \mathbf{J}_r(t(k-1)) \mathbf{d}^{(k)} = - (\mathbf{J}_r(t(k-1)))^T r(t(k-1)) \tag{1.36}
\]

to be solved for \( \mathbf{d}^{(k)} \), where

\[
\mathbf{d}^{(k)} \propto \mathbf{t}^{(k)} - \mathbf{t}^{(k-1)}.
\]

It is clear that the solution \( \mathbf{d}^{(k)} \) is a descent direction; that is, if \( \nabla s(t(k-1)) \neq 0 \),

\[
(\mathbf{d}^{(k)})^T \nabla s(t(k-1)) = - \left( (\mathbf{J}_r(t(k-1)))^T \mathbf{d}^{(k)} \right)^T \left( (\mathbf{J}_r(t(k-1)))^T \mathbf{d}^{(k)} \right) < 0.
\]

The update step is determined by a line search in the appropriate direction:

\[
\mathbf{t}^{(k)} - \mathbf{t}^{(k-1)} = \alpha^{(k)} \mathbf{d}^{(k)}.
\]

This method is called the \textit{Gauss-Newton algorithm}. (The method is also sometimes called the “modified Gauss-Newton algorithm” because many years ago no damping was used in the Gauss-Newton algorithm, and \( \alpha^{(k)} \) was taken as the constant 1. Without an adjustment to the step, the Gauss-Newton method tends to overshoot the minimum in the direction \( \mathbf{d}^{(k)} \).) In practice, rather than a full search to determine the best value of \( \alpha^{(k)} \), we just consider the sequence of values \( 1, \frac{1}{2}, \frac{1}{3}, \ldots \) and take the largest value so that \( s(t(k)) < s(t(k-1)) \). The algorithm terminates when the change is small.

If the residuals are not small or if \( \mathbf{J}_r(t(k)) \) is poorly conditioned, the Gauss-Newton method can perform very poorly. One possibility is to add a conditioning matrix to the coefficient matrix in equation (1.36). A simple choice is \( \tau^{(k)} \mathbf{I}_m \), and the equation for the update becomes

\[
\left((\mathbf{J}_r(t(k-1)))^T \mathbf{J}_r(t(k-1)) + \tau^{(k)} \mathbf{I}_m\right) \mathbf{d}^{(k)} = - (\mathbf{J}_r(t(k-1)))^T r(t(k-1)),
\]

where \( \mathbf{I}_m \) is the \( m \times m \) identity matrix. A better choice may be an \( m \times m \) scaling matrix, \( \mathbf{S}^{(k)} \), that takes into account the variability in the columns of \( \mathbf{J}_r(t(k-1)) \); hence, we have for the update

\[
\left((\mathbf{J}_r(t(k-1)))^T \mathbf{J}_r(t(k-1)) + \lambda^{(k)} (\mathbf{S}^{(k)})^T \mathbf{S}^{(k)}\right) \mathbf{d}^{(k)} = - (\mathbf{J}_r(t(k-1)))^T r(t(k-1)) \tag{1.37}
\]

The basic requirement for the matrix \( (\mathbf{S}^{(k)})^T \mathbf{S}^{(k)} \) is that it improve the condition of the coefficient matrix. There are various ways of choosing this matrix. One is to transform the matrix \( (\mathbf{J}_r(t(k-1)))^T \mathbf{J}_r(t(k-1)) \) so that it has 1's
along the diagonal (this is equivalent to forming a correlation matrix from a variance-covariance matrix), and to use the scaling vector to form \( S^{(k)} \). The nonnegative factor \( \lambda^{(k)} \) can be chosen to control the extent of the adjustment. The sequence \( \lambda^{(k)} \) must go to 0 for the algorithm to converge.

Equation (1.37) can be thought of as a Lagrange multiplier formulation of the constrained problem,

\[
\min_x \frac{1}{2} \left\| J_r(t^{(k-1)})x + r(t^{(k-1)}) \right\|
\]

s.t. \( \left\| S^{(k)}x \right\| \leq \delta_k \). 

The Lagrange multiplier \( \lambda^{(k)} \) is zero if \( d^{(k)} \) from equation (1.36) satisfies \( \|d^{(k)}\| \leq \delta_k \); otherwise, it is chosen so that \( \|S^{(k)}d^{(k)}\| = \delta_k \).

Use of an adjustment such as in equation (1.37) is called the Levenberg-Marquardt algorithm. This is probably the most widely used method for nonlinear least squares.

**Variance of Least Squares Estimators**

In a data-generating process following the simple model \( E(Y) = f(\theta) \), if \( Y \) has finite second moment, then sample mean \( \overline{Y} \) is a consistent estimator (in mean square) of \( f(\theta) \). Furthermore, if \( Y \) has finite fourth moment, then the minimum residual norm \( (r(\hat{\theta}))^T r(\hat{\theta}) \) divided by \( (n - m) \) is a consistent estimator of the variance of \( Y \), say \( \sigma^2 \), that is

\[
\sigma^2 = E(Y - f(\theta))^2.
\]

We denote this estimator as \( \hat{\sigma}^2 \):

\[
\hat{\sigma}^2 = (r(\hat{\theta}))^T r(\hat{\theta}) / (n - m).
\]

This estimator, strictly speaking, is not a least squares estimator of \( \sigma^2 \), but it is based on least squares estimators of the other parameters, and so we call it the least squares estimator.

The variance of the least squares estimator \( \hat{\theta} \), however, is not easy to work out, except in special cases. In the simplest case, \( f \) is linear and \( Y \) has a normal distribution, and we have the familiar linear regression estimates of \( \theta \) and of the variance of the estimator of \( \theta \).

Without the linearity property, however, even with the assumption of normality, it may not be possible to write a simple expression for the variance-covariance matrix of an estimator that is defined as the solution to the least squares optimization problem. Using a linear approximation, however, we may estimate an approximate variance-covariance matrix for \( \theta \) as

\[
\left( (J_r(\hat{\theta}))^T J_r(\hat{\theta}) \right)^{-1} \hat{\sigma}^2.
\]

(1.39)
Compare this linear approximation to the expression for the estimated variance-covariance matrix of the least squares estimator \( \hat{\beta} \) in the linear regression model \( E(Y) = X\beta \), in which \( J_r(\hat{\beta}) \) is just \( X \). An estimate of \( \sigma^2 \) based on the least squares estimates of the parameters is the sum of the squared residuals divided by \( n - m \), where \( m \) is the number of estimated elements in \( \theta \).

If the residuals are small, the Hessian is approximately equal to the cross-product of the Jacobian, as we see from equation (1.35), so an alternate expression for the estimated variance-covariance matrix is

\[
(\mathbf{H}_s(\hat{\theta}))^{-1}\hat{\sigma}^2.
\]

This latter expression is more useful if Newton’s method or a quasi-Newton method is used instead of the Gauss-Newton method for the solution of the least squares problem because in these methods the Hessian or an approximate Hessian is used in the computations.

**Iteratively Reweighted Least Squares**

Often in applications, the residuals in equation (1.23) are not given equal weight for estimating \( \theta \). This may be because the reliability or precision of the observations may be different. For *weighted least squares*, instead of equation (1.33) we have the objective function

\[
s_w(t) = \sum_{i=1}^{n} w_i (r_i(t))^2.
\]

The weights add no complexity to the problem, and the Gauss-Newton methods of the previous section apply immediately, with

\[
\hat{r}(t) = W r(t),
\]

where \( W \) is a diagonal matrix containing the weights.

The simplicity of the computations for weighted least squares suggests a more general usage of the method. Suppose that we are to minimize some other \( L_p \) norm of the residuals \( r_i \), as in equation (1.25). The objective function can be written as

\[
s_p(t) = \sum_{i=1}^{n} \frac{1}{|y_i - f(t)|^{2-p}} |y_i - f(t)|^2.
\]

This leads to an iteration on the least squares solutions. Beginning with \( y_i - f(t^{(0)}) = 1 \), we form the recursion that results from the approximation

\[
s_p(t^{(k)}) \approx \sum_{i=1}^{n} \frac{1}{|y_i - f(t^{(k-1)})|^{2-p}} \left| y_i - f(t^{(k)}) \right|^2.
\]
Hence, we solve a weighted least squares problem, and then form a new weighted least squares problem using the residuals from the previous problem. This method is called iteratively reweighted least squares, or IRLS. The iterations over the residuals are outside the loops of iterations to solve the least squares problems, so in nonlinear least squares, IRLS results in nested iterations.

There are some problems with the use of reciprocals of powers of residuals as weights. The most obvious problem arises from very small residuals. This is usually handled by use of a fixed large number as the weight.

Iteratively reweighted least squares can also be applied to other norms,

$$s_\rho(t) = \sum_{i=1}^{n} \rho(y_i - f(t)),$$

but the approximations for the updates may not be as good.

### 1.4.3 Estimation by Maximum Likelihood

One of the most commonly used approaches to statistical estimation is maximum likelihood. The concept has an intuitive appeal, and the estimators based on this approach have a number of desirable mathematical properties, at least for broad classes of distributions.

Given a sample $y_1, \ldots, y_n$ from a distribution with probability density or probability mass function $p(y \mid \theta)$, a reasonable estimate of $\theta$ is the value that maximizes the joint density or joint probability with variable $t$ at the observed sample value: $\prod_i p(y_i \mid t)$, so long as that maximizing value is within the parameter space of the model, or at least within the closure of that space.

We define the likelihood function as a function of a variable in place of the parameter:

$$L_n(t; y) = \prod_{i=1}^{n} p(y_i \mid t), \quad \text{for } t \in \Theta, \quad (1.44)$$

where $\Theta$ is the parameter space of the probability model.

Note the reversal in roles of variables and parameters. The likelihood function appears to represent a “posterior probability”, but that is not an appropriate interpretation. $L_n$ is not even a PDF (in $t$), even though $\prod_{i=1}^{n} p(y_i \mid t)$ is a PDF (in the variables $y_i$).

Just as in the case of estimation by minimizing residuals, the more difficult and interesting problems involve the determination of the form of the function $p(y_i \mid \theta)$. In these sections, as above, however, we concentrate on the simpler problem of determining an appropriate value of $t$ as an estimator of $\theta$, assuming that the form of $p$ is known.

The value of $t$ for which $L$ attains its maximum value is the maximum likelihood estimate (MLE) of $\theta$ for the given data, $y$. The data—that is, the
realizations of the variables in the density function—are considered as fixed, and the parameters are considered as variables of the optimization problem,

$$\max_{t \in \Theta} L_n(t ; y).$$  \hfill (1.45)

This optimization problem can be much more difficult than the optimization problem (1.24) that results from an estimation approach based on minimization of some norm of a residual vector. As we discussed in that case, there can be both computational and statistical problems associated either with restrictions on the set of possible parameter values or with the existence of local optima of the objective function. These problems also occur in maximum likelihood estimation.

Applying constraints in the optimization problem to force the solution to be within the set of possible parameter values is called restricted maximum likelihood estimation, or REML estimation. In addition to problems due to constraints or due to local optima, other problems may arise if the likelihood function is bounded. The conceptual difficulties resulting from an unbounded likelihood are much deeper. In practice, for computing estimates in the unbounded case, the general likelihood principle may be retained, and the optimization problem redefined to include a penalty that keeps the function bounded. Adding a penalty to form a bounded objective function in the optimization problem, or to dampen the solution is called penalized maximum likelihood estimation.

For a broad class of distributions, the maximum likelihood criterion yields estimators with good statistical properties. The conditions that guarantee certain optimality properties are called the “regular case”.

Although in practice, the functions of residuals that are minimized are almost always differentiable, and the optimum occurs at a stationary point, this is often not the case in maximum likelihood estimation. A standard example in which the MLE does not occur at a stationary point is a distribution in which the range depends on the parameter, and the simplest such distribution is the uniform \( U(0, \theta) \). In this case, the MLE is the max order statistic.

An important family of probability distributions are those whose probability densities are members of the exponential family, that is, densities of the form

$$p(y | \theta) = h(y) \exp \left( \theta^T g(y) - a(\theta) \right), \quad \text{if } y \in \mathcal{Y},$$

$$= 0, \quad \text{otherwise}, \quad (1.46)$$

where \( \mathcal{Y} \) is some set, \( \theta \) is an \( m \)-vector, and \( g(\cdot) \) is an \( m \)-vector-valued function. Maximum likelihood estimation is particularly straightforward for distributions in the exponential family. Whenever \( \mathcal{Y} \) does not depend on \( \theta \), and \( g(\cdot) \) and \( a(\cdot) \) are sufficiently smooth, the MLE has certain optimal statistical properties. This family of probability distributions includes many of the familiar distributions, such as the normal, the binomial, the Poisson, the gamma, the Pareto, and the negative binomial.
The log-likelihood function,
\[ l_{L_n}(t; y) = \log L_n(t; y), \quad (1.47) \]
is a sum rather than a product. The form of the log-likelihood in the exponential family is particularly simple:
\[ l_{L_n}(t; y) = \sum_{i=1}^{n} t_i ^T g(y_i) - n a(t) + c, \]
where \( c \) depends on the \( y_i \) but is constant with respect to the variable of interest.

The logarithm is monotone, so the optimization problem (1.45) can be solved by solving the maximization problem with the log-likelihood function:
\[ \max_{t \in \Theta} l_{L_n}(t; y). \quad (1.48) \]

In the following discussion, we will find it convenient to drop the subscript \( n \) in the notation for the likelihood and the log-likelihood. We will also often work with the likelihood and log-likelihood as if there is only one observation. (A general definition of a likelihood function is any nonnegative function that is proportional to the density or the probability mass function; that is, it is the same as the density or the probability mass function except that the arguments are switched, and its integral or sum over the domain of the random variable need not be 1.)

If the likelihood is twice differentiable and if the range does not depend on the parameter, Newton’s method (see equation (1.32)) could be used to solve the optimization problem (1.48). Newton’s equation
\[ H_{L_L}(t^{(k-1)}; y) d^{(k)} = \nabla l_{L_L}(t^{(k-1)}; y) \quad (1.49) \]
is used to determine the step direction in the \( k^{th} \) iteration. A quasi-Newton method, as we mentioned on page 25, uses a matrix \( \tilde{H}_{L_L}(t^{(k-1)}) \) in place of the Hessian \( H_{L_L}(t^{(k-1)}) \). At this point, we should remind the reader of the comments on page 22. Mathematical expressions do not necessarily imply computational methods. There are many additional considerations for the numerical computations, and the expressions below, such as equations (1.52), (1.54), and (1.55), rarely should be used directly in a computer program.

Statistical Models

In the foregoing I have been careful, at the cost of some notational complexity, to distinguish between a parameter and a variable in a computational method used to estimate that parameter. Most other authors do not do this, and I do not always do it myself. It is important to make the distinction in data analysis when the objective is to fit a statistical model. In analysis of a statistical model
(as opposed to using the model to analyze data), it is often important to think
of the parameter as a variable. In fact, we may take the derivative of a PDF
with respect to the parameter, which obviously requires that the parameter
be a variable.

In this section, we will consider some important useful relationships be-
tween the log-likelihood function and properties of probability models as they
range over the parameter space, that is, when the parameter is considered to
be a variable.

If it exists, the derivative of the log-likelihood is the relative rate of change,
with respect to the parameter placeholder \( \theta \), of the probability density func-
tion at a fixed observation. If \( \theta \) is a scalar, some positive function of the
derivative such as its square or its absolute value is obviously a measure of
the effect of change in the parameter or in the estimate of the parameter. More
generally, an outer product of the derivative with itself is a useful measure
of the changes in the components of the parameter at any given point in the
parameter space:

\[
\nabla l_L(\theta ; y) \ (\nabla l_L(\theta ; y))^T.
\]

The average of this quantity with respect to the probability density of the
random variable \( Y \),

\[
I(\theta | Y) = E_{\theta} \left( \nabla l_L(\theta | Y) \ (\nabla l_L(\theta | Y))^T \right), \tag{1.50}
\]
is called the information matrix, or the Fisher information matrix, that an
observation on \( Y \) contains about the parameter \( \theta \).

The optimization problem (1.45) or (1.48) can be solved by Newton’s
method, equation (1.31) on page 25, or by a quasi-Newton method. (We should
first note that this is a maximization problem, so the signs are reversed from
our previous discussion of a minimization problem.)

If \( \theta \) is a scalar, the expected value of the square of the first deriv ative is
the expected value of the negative of the second derivative,

\[
E \left( \left( \frac{\partial}{\partial \theta} l_L(\theta ; y) \right)^2 \right) = -E \left( \frac{\partial^2}{\partial \theta^2} l_L(\theta ; y) \right),
\]
or, in general,

\[
E \left( \nabla l_L(\theta ; y) \ (\nabla l_L(\theta ; y))^T \right) = -E \left( H_L(\theta ; y) \right). \tag{1.51}
\]
The expected value of the second derivative, or an approximation of it, can
be used in a Newton-like method to solve the maximization problem.

### Computations

A common quasi-Newton method for optimizing \( l_L(\theta ; y) \) is Fisher scoring, in
which the Hessian in Newton’s method is replaced by its expected value. The
expected value can be replaced by an estimate, such as the sample mean. The iterates then are

\[ \theta^{(k)} = \theta^{(k-1)} - \left( \bar{E}(\theta^{(k-1)}) \right)^{-1} \nabla l_L(\theta^{(k-1)} ; y), \tag{1.52} \]

where \( \bar{E}(\theta^{(k-1)}) \) is an estimate or an approximation of \( E\left(H_l L(\theta^{(k-1)} | Y)\right) \), which is itself an approximation of \( E\theta^* (H_l L(\theta | Y)) \). By equation (1.51), this is the negative of the Fisher information matrix if the differentiation and expectation operators can be interchanged. (This is one of the “regularity conditions” we alluded to earlier.) The most common practice is to take \( \bar{E}(\theta^{(k-1)}) \) as the Hessian evaluated at the current value of the iterations on \( \theta \); that is, as \( H_l L(\theta^{(k-1)} ; y) \). This is called the observed information matrix.

In some cases a covariate \( x_i \) may be associated with the observed \( y_i \), and the distribution of \( Y \) with given covariate \( x_i \) has a parameter \( \mu \) that is a function of \( x_i \) and \( \theta \). (The linear regression model is an example, with \( \mu_i = x_i^T \theta \).) We may in general write \( \mu = x_i(\theta) \). In these cases, another quasi-Newton method may be useful. The Hessian in equation (1.49) is replaced by

\[ \left( X(\theta^{(k-1)}) \right)^T K(\theta^{(k-1)}) X(\theta^{(k-1)}), \tag{1.54} \]

where \( K(\theta^{(k-1)}) \) is a positive definite matrix that may depend on the current value \( \theta^{(k-1)} \). (Again, think of this in the context of a regression model, but not necessarily linear regression.) This method is called the Delta algorithm because of its similarity to the delta method for approximating a variance-covariance matrix (described on page 40).

In some cases, when \( \theta \) is a vector, the optimization problem (1.45) or (1.48) can be solved by alternating iterations on the elements of \( \theta \). In this approach, iterations based on equations such as (1.49) are

\[ \bar{H}_L \left( \theta_i^{(k-1)} ; \theta_j^{(k-1)} , y \right) \frac{\delta_i^{(k)}}{\delta_j^{(k)}} = \nabla l_L \left( \theta_i^{(k-1)} ; \theta_j^{(k-1)} , y \right), \tag{1.55} \]

where \( \theta = (\theta_i, \theta_j) \) (or \( (\theta_j, \theta_i) \)), \( \delta_i \) is the update direction for \( \theta_i \), and \( \theta_j \) is considered to be constant in this step. In the next step, the indices \( i \) and \( j \) are exchanged. This is called componentwise optimization. For some objective functions, the optimal value of \( \theta_i \) for fixed \( \theta_j \) can be determined in closed form. In such cases, componentwise optimization may be the best method.

Sometimes, we may be interested in the MLE of \( \theta_i \) given a fixed value of \( \theta_j \), so the iterations do not involve an interchange of \( i \) and \( j \) as in componentwise optimization. Separating the arguments of the likelihood or log-likelihood function in this manner leads to what is called profile likelihood, or concentrated likelihood.
1.4 The Role of Optimization in Inference

As a purely computational device, the separation of $\theta$ into smaller vectors makes for a smaller optimization problem for which the number of computations is reduced by more than a linear amount. The iterations tend to zigzag toward the solution, so convergence may be quite slow. If, however, the Hessian is block diagonal, or almost block diagonal (with sparse off-diagonal submatrices), two successive steps of the alternating method are essentially equivalent to one step with the full $\theta$. The rate of convergence would be the same as that with the full $\theta$. Because the total number of computations in the two steps is less than the number of computations in a single step with a full $\theta$, the method may be more efficient in this case.

Statistical Properties of MLE

Under suitable regularity conditions we referred to earlier, maximum likelihood estimators have a number of desirable properties. For most distributions used as models in practical applications, the MLEs are consistent. Furthermore, in those cases, the MLE $\hat{\theta}$ is asymptotically normal (with mean $\theta^*$) with variance-covariance matrix

$$
\left( \mathbb{E}_{\theta^*} \left( -H_{\theta^*} (\theta^* \mid Y) \right) \right)^{-1},
$$

which is the inverse of the Fisher information matrix. A consistent estimator of the variance-covariance matrix is the inverse of the Hessian at $\hat{\theta}$. (Note that there are two kinds of asymptotic properties and convergence issues. Some involve the iterative algorithm, and the others are the usual statistical asymptotics in terms of the sample size.)

EM Methods

As we mentioned above, the computational burden in a single iteration for solving the MLE optimization problem can be reduced by more than a linear amount by separating $\theta$ into two subvectors. The MLE is then computed by alternating between computations involving the two subvectors, and the iterations proceed in a zigzag path to the solution. Each of the individual sequences of iterations is simpler than the sequence of iterations on the full $\theta$.

Another alternating method that arises from an entirely different approach alternates between updating $\hat{\theta}^{(k)}$ using maximum likelihood and conditional expected values. This method is called the EM method because the alternating steps involve an expectation and a maximization. The method was described and analyzed by Dempster et al. (1977). Many additional details and alternatives as well as many examples of applications of the EM algorithm are discussed by Ng et al. (2012).

The EM methods can be explained most easily in terms of a random sample that consists of two components, one observed and one unobserved or
missing. A simple example of missing data occurs in life-testing, when, for example, a number of electrical units are switched on and the time when each fails is recorded. In such an experiment, it is usually necessary to curtail the recordings prior to the failure of all units. The failure times of the units still working are unobserved. The data are said to be right censored. The number of censored observations and the time of the censoring obviously provide information about the distribution of the failure times.

The missing data can be missing observations on the same random variable that yields the observed sample, as in the case of the censoring example; or the missing data can be from a different random variable that is related somehow to the random variable observed.

Many common applications of EM methods do involve missing-data problems, but this is not necessary. Often, an EM method can be constructed based on an artificial “missing” random variable to supplement the observable data.

Let \( Y = (U, V) \), and assume that we have observations on \( U \) but not on \( V \). We wish to estimate the parameter \( \theta \), which figures in the distribution of both components of \( Y \). An EM method uses the observations on \( U \) to obtain a value of \( \theta^{(k)} \) that increases the likelihood and then uses an expectation based on \( V \) that increases the likelihood further.

Let \( L_{\text{c}}(\theta \mid u, v) \) and \( l_{\text{c}}(\theta \mid u, v) \) denote, respectively, the likelihood and the log-likelihood for the complete sample. The likelihood for the observed \( U \) is
\[
L(\theta \mid u) = \int L_{\text{c}}(\theta \mid u, v) \, dv,
\]
and \( l_{L}(\theta \mid u) = \log L(\theta \mid u) \). The EM approach to maximizing \( L(\theta \mid u) \) has two alternating steps. The first one begins with a value \( \theta^{(0)} \). The steps are iterated until convergence.

- **E step**: compute \( q^{(k)}(\theta) = \mathbb{E}_{V \mid u, \theta^{(k-1)}} \left( l_{\text{c}}(\theta \mid u, V) \right) \).
- **M step**: determine \( \theta^{(k)} \) to maximize \( q^{(k)}(\theta) \), subject to any constraints on acceptable values of \( \theta \).

The sequence \( \theta^{(1)}, \theta^{(2)}, \ldots \) converges to a local maximum of the observed-data likelihood \( L(\theta \mid u) \) under fairly general conditions (including, of course, the nonexistence of a local maximum near enough to \( \theta^{(0)} \)). The EM method can be very slow to converge, however.

For a simple example of the EM method, see Exercise 1.16, in which the problem in Dempster et al. (1977) is described. In this case, the unobserved data is a result of the model being overparametrized. Another simple example is the case of a sample from a mixture of distributions. The missing data are the indicators of which distribution gave rise to the individual observations (see Exercise 1.17). As mentioned above, life-testing experiments or other experiments in which there is an “outcome” often yield missing data because some of the experimental units are removed from the experiment before the outcome actually occurs. The times-to-outcome of the censored experimental
1.4 The Role of Optimization in Inference

units are not observed; only bounds on the times are known. For instance, consider the simple exponential model for failure times

\[ p(t) = \frac{1}{\theta} e^{-t/\theta}. \]

A total of \( n \) items are put on test, and their ending times are to be recorded. However, as is often the case, only the failure times \( t_1, \ldots, t_{n_1} \) of the first \( n_1 \) items to fail are recorded, and for the remaining \( n - n_1 \) items all we know are the times at which they were taken from the test (or “censored”), \( c_{n_1 + 1}, \ldots, c_n \). The MLE of \( \theta \) can easily be obtained, however, because the log-likelihood is

\[ l(\theta) = -n_1 \log(\theta) - \left( \sum_{i=1}^{n_1} x_i + \sum_{i=m_1+1}^{n} c_i \right)/\theta, \]

and its maximum clearly occurs at

\[ \hat{\theta} = \left( \sum_{i=1}^{n_1} x_i + \sum_{i=m_1+1}^{n} c_i \right)/n_1. \]

This example, however, provides useful insight into the EM method, and in Exercise 1.18, you are asked to formulate this estimation problem to solve as an EM method.

As a further example of the EM method, consider an experiment described by Flury and Zoppec (2000). It is assumed that the lifetime of light bulbs follows an exponential distribution with mean \( \theta \). To estimate \( \theta \), \( n \) light bulbs were tested until they all failed. Their failure times were recorded as \( u_1, \ldots, u_n \). In a separate experiment, \( m \) bulbs were tested, but the individual failure times were not recorded. Only the number of bulbs, \( r \), that had failed at time \( t \) was recorded. The missing data are the failure times of the bulbs in the second experiment, \( v_1, \ldots, v_m \). We have

\[ l_{L}(\theta ; u, v) = -n (\log \theta + \bar{u}/\theta) - \sum_{i=1}^{m} (\log \theta + v_i/\theta). \]

The expected value, \( E_{V|u,\theta(k-1)} \), of this is

\[ q^{(k)}(\theta) = -(n+m) \log \theta - \frac{1}{\theta} \left( n\bar{u} + (m-r)(t + \theta^{(k-1)}) + r(\theta^{(k-1)} - t\theta^{(k-1)}) \right), \]

where \( h^{(k-1)} \) is given by

\[ h^{(k-1)} = \frac{e^{-t/\theta^{(k-1)}}}{1 - e^{-t/\theta^{(k-1)}}}. \]

The \( k \)th M step determines the maximum, which, given \( \theta^{(k-1)} \), occurs at

\[ Elements of Computational Statistics, Second Edition ©2013 James E. Gentle \]
\[ \theta^{(k)} = \frac{1}{n + m} \left( n\bar{u} + (m - r)(t + \theta^{(k-1)}) + r(\theta^{(k-1)} - th^{(k-1)}) \right). \]

Starting with a positive number \( \theta^{(0)} \), this equation is iterated until convergence.

This example is interesting because if we assume that the distribution of the light bulbs is uniform, \( U(0, \theta) \) (such bulbs are called “heavybulbs”!), the EM algorithm cannot be applied. As we have pointed out above, maximum likelihood methods must be used with some care whenever the range of the distribution depends on the parameter. In this case, however, there is another problem. It is in computing \( q^{(k)}(\theta) \), which does not exist for \( \theta < \theta^{(k-1)} \).

Although in the paper that first provided a solid description of the EM method (Dempster et al. (1977)), specific techniques were used for the computations in the two steps, it is not necessary for the EM method to use those same inner-loop algorithms. There are various other ways to perform each of these computations. A number of papers since 1977 have suggested specific methods for the computations and have given new names to methods based on those inner-loop computations.

For the expectation step, there are not as many choices. In the happy case of an exponential family or some other nice distributions, the expectation can be computed in closed form. Otherwise, computing the expectation is a numerical quadrature problem. There are various procedures for quadrature, including Monte Carlo (see page 69). The additional Monte Carlo computations add a lot to the overall time required for convergence of the EM method. Even the variance-reducing methods discussed in Section 2.6 can do little to speed up the method. Monte Carlo samples from previous expectation steps may be used in order to reduce the costs of Monte Carlo sampling.

An additional problem in using Monte Carlo in the expectation step may be that the distribution of \( Y \) is difficult to simulate. The versatile Gibbs method (page 66) is often useful in this context.

The convergence criterion for optimization methods that involve Monte Carlo generally should be tighter than for deterministic methods.

For the maximization step, there are more choices, as we have seen in the discussion of maximum likelihood estimation above.

For the maximization step, Dempster et al. (1977) suggested requiring only an increase in the expected value; that is, take \( \theta^{(k)} \) so that \( q_k(\theta^{(k)}) \geq q_{k-1}(\theta^{(k-1)}) \). This is called a generalized EM algorithm, or GEM. There are several simple possibilities for achieving an increase. One way is just to take \( \theta^{(k)} \) as the point resulting from a single Newton step. Another way of simplifying the the M-step is just to use a componentwise maximization, as in the update step of equation (1.55) on page 34; that is, if \( \theta = (\theta_1, \theta_2) \), first \( \theta_1^{(k)} \) is determined to maximize \( q \) subject to the constraint \( \theta_2 = \theta_2^{(k-1)} \); then \( \theta_2^{(k)} \) is determined to maximize \( q \) subject to the constraint \( \theta_1 = \theta_1^{(k)} \). This sometimes simplifies the maximization problem so that it can be done in closed form.
As is usual for estimators defined as solutions to optimization problems, we may have some difficulty in determining the statistical properties of the estimators. One method of estimating the variance-covariance matrix of the estimator is by use of the gradient and Hessian of the complete-data log-likelihood, \( l_{\text{LC}}(\theta; u, v) \), as in equation (1.40). Bootstrap methods (see Chapter 4) can also be used for estimation of the variance-covariance matrix.

It is interesting to note that under certain assumptions on the distribution, the iteratively reweighted least squares method discussed on page 30 can be formulated as an EM method (see Dempster et al. (1980)).

### 1.5 Inference about Functions

#### Functions of Parameters and Functions of Estimators

Suppose that instead of estimating the parameter \( \theta \), we wish to estimate \( g(\theta) \), where \( g(\cdot) \) is some function. If the function \( g(\cdot) \) is monotonic or has certain other regularity properties, it may be the case that the estimator that results from the minimum residuals principle or from the maximum likelihood principle is invariant; that is, the estimator of \( g(\theta) \) is merely the function \( g(\cdot) \) evaluated at the solution to the optimization problem for estimating \( \theta \). The statistical properties of a \( T \) for estimating \( \theta \), however, do not necessarily carry over to \( g(T) \) for estimating \( g(\theta) \).

As an example of why a function of an unbiased estimator may not be unbiased, consider a simple case in which \( T \) and \( g(T) \) are scalars. Let \( R = g(T) \) and consider \( \text{E}(R) \) and \( g(\text{E}(T)) \) in the case in which \( g \) is a convex function. (A function \( g \) is a convex function if for any two points \( x \) and \( y \) in the domain of \( g \), \( g(\frac{1}{2}(x + y)) \leq \frac{1}{2}(g(x) + g(y)) \).) In this case, obviously

\[
\text{E}(R) \leq g(\text{E}(T)), \tag{1.57}
\]

so \( R \) is biased for \( g(\theta) \). (This relation is Jensen’s inequality.) An opposite inequality obviously also applies to a concave function, in which case the bias is positive.

It is often possible to adjust \( R \) to be unbiased for \( g(\theta) \); and properties of \( T \), such as sufficiency for \( \theta \), may carry over to the adjusted \( R \). Some of the applications of the jackknife and the bootstrap that we discuss later are in making adjustments to estimators of \( g(\theta) \) that are based on estimators of \( \theta \).

The variance of \( R = g(T) \) can often be approximated in terms of the variance of \( T \). Let \( T \) and \( \theta \) be \( m \)-vectors and \( R \) be a \( k \)-vector. In a simple but common case, we may know that \( T \) in a sample of size \( n \) has an approximate normal distribution with mean \( \theta \) and some variance-covariance matrix, say \( V(T) \), and \( g \) is a smooth function (that is, it can be approximated by a truncated Taylor series about \( \theta \)).
Basic Computations in Statistical Inference

\[
R_i = g_i(T) \\
\approx g_i(\theta) + J_{g_i}(\theta)(T - \theta) + \frac{1}{2}(T - \theta)^T H_{g_i}(\theta)(T - \theta).
\]

Because the variance of \( T \) is \( O(n^{-1}) \), the remaining terms in the expansion go to zero in probability at the rate of at least \( n^{-1} \).

This yields the approximations

\[
E(R) \approx g(\theta)
\]

and

\[
V(R) \approx J_{g}(\theta) V(T) (J_{g}(\theta))^T.
\]

This method of approximation of the variance is called the delta method.

A common form of a simple estimator that may be difficult to analyze and may have unexpected properties is a ratio of two statistics,

\[
R = \frac{T}{S},
\]

where \( S \) is a scalar. An example is a studentized statistic, in which \( T \) is a sample mean and \( S \) is a function of squared deviations. If the underlying distribution is normal, a statistic of this form may have a well-known and tractable distribution. In particular, if \( T \) is a mean and \( S \) is a function of an independent chi-squared random variable, the distribution is that of a Student’s \( t \). If the underlying distribution has heavy tails, however, the distribution of \( R \) may have unexpectedly light tails. An asymmetric underlying distribution may also cause the distribution of \( R \) to be very different from a Student’s \( t \) distribution. If the underlying distribution is positively skewed, the distribution of \( R \) may be negatively skewed (see Exercise 1.19).

**Linear Estimators**

A functional \( \Theta \) is linear if, for any two functions \( f \) and \( g \) in the domain of \( \Theta \) and any real number \( a \),

\[
\Theta(af + g) = a\Theta(f) + \Theta(g).
\]

A statistic is linear if it is a linear functional of the ECDF. A linear statistic can be computed from a sample using an online algorithm, and linear statistics from two samples can be combined by addition. Strictly speaking, this definition excludes statistics such as means, but such statistics are essentially linear in the sense that they can be combined by a linear combination if the sample sizes are known.
1.6 Probability Statements in Statistical Inference

There are two important instances in statistical inference in which statements about probability are associated with the decisions of the inferential methods. In hypothesis testing, under assumptions about the distributions, we base our inferential methods on probabilities of two types of errors. In confidence intervals the decisions are associated with probability statements about coverage of the parameters. For both cases the probability statements are based on the distribution of a random sample, $Y_1, \ldots, Y_n$.

In computational inference, probabilities associated with hypothesis tests or confidence intervals are estimated by simulation of a hypothesized data-generating process or by resampling of an observed sample.

1.6.1 Tests of Hypotheses

Often statistical inference involves testing a “null” hypothesis, $H_0$, about the parameter. In a simple case, for example, we may test the hypothesis

$$H_0 : \theta = \theta_0$$

versus an alternative hypothesis that $\theta$ takes on some other value or is in some set that does not include $\theta_0$. The straightforward way of performing the test involves use of a test statistic, $T$, computed from a random sample of data, $Y_1, \ldots, Y_n$. Associated with $T$ is a rejection region $C$ such that if the null hypothesis is true, $Pr (T \in C)$ is some preassigned (small) value, $\alpha$, and $Pr (T \in C)$ is greater than $\alpha$ if the null hypothesis is not true. Thus, $C$ is a region of more “extreme” values of the test statistic if the null hypothesis is true. If $T \in C$, the null hypothesis is rejected. It is desirable that the test have a high probability of rejecting the null hypothesis if indeed the null hypothesis is not true. The probability of rejection of the null hypothesis is called the power of the test.

A procedure for testing that is mechanically equivalent to this is to compute the test statistic $t$ and then to determine the probability that $T$ is more extreme than $t$. In this approach, the realized value of the test statistic determines a region $C_t$ of more extreme values. The probability that the test statistic is in $C_t$ if the null hypothesis is true, $Pr (T \in C_t)$, is called the “p-value” or “significance level” of the realized test statistic.

If the distribution of $T$ under the null hypothesis is known, the critical region or the p-value can be determined. If the distribution of $T$ is not known, some other approach must be used. A common method is to use some approximation to the distribution. The objective is to approximate a quantile of $T$ under the null hypothesis. The approximation is often based on an asymptotic distribution of the test statistic. In Monte Carlo tests, discussed in Section 2.3, the quantile of $T$ is estimated by simulation of the distribution of the underlying data.
1.6.2 Confidence Intervals

Our usual notion of a confidence interval relies on a frequency approach to probability, and it leads to the definition of a $1 - \alpha$ confidence interval for the (scalar) parameter $\theta$ as the random interval $(T_L, T_U)$ that has the property

$$\Pr(T_L \leq \theta \leq T_U) = 1 - \alpha.$$  \hfill (1.60)

This is also called a $(1 - \alpha)100\%$ confidence interval. The endpoints of the interval, $T_L$ and $T_U$, are functions of a sample, $Y_1, \ldots, Y_n$. The interval $(T_L, T_U)$ is not uniquely determined.

The concept extends easily to vector-valued parameters. Rather than taking vectors $T_L$ and $T_U$, however, we generally define an ellipsoidal region, whose shape is determined by the covariances of the estimators.

A realization of the random interval, say $(t_L, t_U)$, is also called a confidence interval. Although it may seem natural to state that the "probability that $\theta$ is in $(t_L, t_U)$ is $1 - \alpha$", this statement can be misleading unless a certain underlying probability structure is assumed.

In practice, the interval is usually specified with respect to an estimator of $\theta$, $T$. If we know the sampling distribution of $T - \theta$, we may determine $c_1$ and $c_2$ such that

$$\Pr(c_1 \leq T - \theta \leq c_2) = 1 - \alpha;$$ \hfill (1.61)

and hence

$$\Pr(T - c_1 \leq \theta \leq c_2 - T) = 1 - \alpha.$$  

If either $T_L$ or $T_U$ in equation (1.60) is infinite or corresponds to a bound on acceptable values of $\theta$, the confidence interval is one-sided. For two-sided confidence intervals, we may seek to make the probability on either side of $T$ to be equal, to make $c_1 = -c_2$, and/or to minimize $|c_1|$ or $|c_2|$. This is similar in spirit to seeking an estimator with small variance.

For forming confidence intervals, we generally use a function of the sample that also involves the parameter of interest, $f(T, \theta)$. The confidence interval is then formed by separating the parameter from the sample values.

Whenever the distribution depends on parameters other than the one of interest, we may be able to form only conditional confidence intervals that depend on the value of the other parameters. A class of functions that are particularly useful for forming confidence intervals in the presence of such nuisance parameters are called pivotal values, or pivotal functions. A function $f(T, \theta)$ is said to be a pivotal function if its distribution does not depend on any unknown parameters. This allows exact confidence intervals to be formed for the parameter $\theta$. We first form

$$\Pr\left(f_{(\alpha/2)}(T, \theta) \leq f(T, \theta) \leq f_{(1-\alpha/2)}\right) = 1 - \alpha,$$ \hfill (1.62)

where $f_{(\alpha/2)}$ and $f_{(1-\alpha/2)}$ are quantiles of the distribution of $f(T, \theta)$; that is,
\[ \Pr(f(T, \theta) \leq f(\pi)) = \pi. \]

If, as in the case considered above, \( f(T, \theta) = T - \theta \), the resulting confidence interval has the form

\[ \Pr \left( T - f(1 - \alpha/2) \leq \theta \leq T - f(\alpha/2) \right) = 1 - \alpha. \]

For example, suppose that \( Y_1, \ldots, Y_n \) is a random sample from a \( N(\mu, \sigma^2) \) distribution, and \( \bar{Y} \) is the sample mean. The quantity

\[ f(\bar{Y}, \mu) = \frac{\sqrt{n(n-1)} (\bar{Y} - \mu)}{\sqrt{\sum (Y_i - \bar{Y})^2}} \]

has a Student’s \( t \) distribution with \( n - 1 \) degrees of freedom, no matter what is the value of \( \sigma^2 \). This is one of the most commonly used pivotal values.

The pivotal value in equation (1.63) can be used to form a confidence value for \( \theta \) by first writing

\[ \Pr \left( t_{(\alpha/2)} \leq f(\bar{Y}, \mu) \leq t_{(1 - \alpha/2)} \right) = 1 - \alpha, \]

where \( t_{(\pi)} \) is a percentile from the Student’s \( t \) distribution. Then, after making substitutions for \( f(\bar{Y}, \mu) \), we form the familiar confidence interval for \( \mu \):

\[ \left( \bar{Y} - t_{(1 - \alpha/2)} s/\sqrt{n}, \quad \bar{Y} - t_{(\alpha/2)} s/\sqrt{n} \right), \]

where \( s^2 \) is the usual sample variance, \( \sum (Y_i - \bar{Y})^2 / (n - 1) \).

Other similar pivotal values have \( F \) distributions. For example, consider the usual linear regression model in which the \( n \)-vector random variable \( Y \) has a \( N_n(X\beta, \sigma^2 I) \) distribution, where \( X \) is an \( n \times m \) known matrix, and the \( m \)-vector \( \beta \) and the scalar \( \sigma^2 \) are unknown. A pivotal value useful in making inferences about \( \beta \) is

\[ g(\hat{\beta}, \beta) = \frac{(X(\hat{\beta} - \beta))^T X(\hat{\beta} - \beta)/m}{(Y - X\hat{\beta})^T (Y - X\hat{\beta})/(n - m)}, \]

where

\[ \hat{\beta} = (X^T X)^+ X^T Y. \]

The random variable \( g(\hat{\beta}, \beta) \) for any finite value of \( \sigma^2 \) has an \( F \) distribution with \( m \) and \( n - m \) degrees of freedom.

For a given parameter and family of distributions, there may be multiple pivotal values. For purposes of statistical inference, such considerations as unbiasedness and minimum variance may guide the choice of a pivotal value to use. Alternatively, it may not be possible to identify a pivotal quantity for a particular parameter. In that case, we may seek an approximate pivot.
function is asymptotically pivotal if a sequence of linear transformations of
the function is pivotal in the limit as $n \to \infty$.

If the distribution of $T$ is known, $c_1$ and $c_2$ in equation (1.61) can be
determined. If the distribution of $T$ is not known, some other approach must
be used. A method for computational inference, discussed in Section 4.3, is to
use bootstrap samples from the ECDF.

1.7 Computational Methods

Mathematics provides us models for data and for computations. There are
two significant ways in which mathematical models differ from the actual
quantities we work with and how we work with them.

- The actual numbers we work with are not the same as the mathematical
  abstraction $\mathbb{R}$, the real numbers.
- Useful mathematical expressions often do not indicate good computational
  methods.

Computer Arithmetic

Numerical quantities can be represented in various ways in a computer. The
standard representation that corresponds most closely to the mathematical
reals $\mathbb{R}$ is the floating-point system, $\mathbb{F}$. Numerical quantities are represented
in $\mathbb{F}$ using a fixed number of bits separated into three distinct components of
a number: its sign, its order of magnitude (expressed as a power of 2), and its
value within a fixed range determined by its magnitude. This representation
immediately implies that the values represented cannot be dense, in fact the
number of different values must be finite.

The computer arithmetic built onto $\mathbb{F}$ corresponds closely to the arith-
metic on $\mathbb{R}$, but the fact that many numbers in $\mathbb{R}$ cannot be represented
exactly in $\mathbb{F}$ means that the arithmetic in the two systems cannot be exactly
the same. Three basic properties of arithmetic on $\mathbb{R}$ that cannot be preserved
in a finite system are closure (of addition or of multiplication), associativ-
ity (of addition or of multiplication), and distribution of multiplication over
addition.

The IEEE Standard 754 (which is also the ISO/IEC/IEEE 60559 standard)
prescribes how the real numbers are to be represented in either 32 or 64 bits,
and some minimal requirements for arithmetic on those numbers.

The fact that the three components of the floating-point representation
each utilize a fixed number of bits means that the number of distinct values
within one range of values is different from the number of distinct values
within another range of the same length. For example, there are twice as
many distinct values between 1 and 2 that can be represented in $\mathbb{F}$ than there
are between 2 and 3 that can be represented.
Algorithms

There are many useful mathematical expressions that should not be followed in order to compute the quantity represented by the expression. For example, the least-squares estimate of $\beta$ in the full-rank model $y = X\beta + \epsilon$ is given by $(X^T X)^{-1} X^T y$. To evaluate this quantity directly from the expression, we would first form $X^T X$, then take its inverse, then form $X^T y$ and finally multiply $X^T y$ by the inverse of $X^T X$. Such computations are subject to bad accumulated rounding error. Better computational methods to compute the least squares estimate of $\beta$ never form $X^T X$, whose condition number is the square of the condition number of $X$.

For some quantities that we want to evaluate, there is a closed form, such as the expression above for the least-squares estimate of $\beta$ in the linear model. For many quantities of interest, there is no closed form expression, rather we have only statements of its properties, such as in the case of least-squares estimates in nonlinear models. In that case, we must use a sequence of iterations, perhaps the Newton iterates in equation (1.31), on page 25.

Iterative methods need a rule for deciding when to stop, and take the latest iteration as the final result. Convergence criteria are usually based on changes in the magnitude of some quantity from one iteration to the next. In general, we should never base the convergence decision on no change; that is, we should not test that the value of one floating-point variable is equal to the value of another floating-point variable. The test for convergence should be based on two values’ being within a small range of each other.

An important feature of some algorithms is that an individual datum need enter the computations only once; that is, the computations can be made and updated by passing through the data only once. Such algorithms are called online. For example, algorithms to compute a sum or to compute the least-squares solutions in a linear model can be online, which an algorithm to compute the median of a set of data cannot be online. Iterative algorithms, in general, cannot be online.

For any given problem there are usually more than one algorithm. Some algorithms may be expected to be more accurate than others; that is, they are less affected by roundoff, or else they are less affected by the characteristics of the dataset to which they are applied. Also, some algorithms generally take less time to perform the computations than other algorithms. We often measure the efficiency of algorithms in terms of the number of floating-point operations. Such counts are often expressed as an order in terms of some measure of the size of the problem. For example, to sort $n$ elements, where $n$ is a large number, one algorithm may require approximately $n^2$ comparisons, whereas another algorithm may require only approximately $n \log(n)$ comparisons. We denote these levels of efficiency as $O(n^2)$ and $O(n \log(n))$. The values of the data, not just the number of data elements, may affect the efficiency of algorithms, so measures of efficiency must be based on some specific characteristics of datasets, We sometimes speak of a worst-case order or an average order.
Chapters 2 and 3 in Gentle (2009) discuss computer arithmetic and algorithms in more detail.

Exercises

Some of these exercises require you to simulate output from a given data-generating process by using random number generators (see Appendix B). For these exercises you do not need to know much about random number generation, which we will briefly discuss in Chapter 2. Some exercises require you to conduct small Monte Carlo studies to assess the performance of statistical techniques (see Appendix A).

1.1. a) How would you describe, in nontechnical terms, the structure of the dataset displayed in Figure 1.1, page 7?
   b) How would you describe the structure of the dataset in more precise mathematical terms? (Obviously, without having the actual data, your equations must contain unknown quantities. The question is meant to make you think about how you would do this—that is, what would be the components of your model.)

1.2. Prove that if $X$ is a random variable with an absolutely continuous distribution function $P_X$, the random variable $P_X(X)$ has a $U(0, 1)$ distribution.

1.3. Show that the variance of the ECDF at a point $y$ is the expression in equation (1.8) on page 12. Use the definition of the variance in terms of expected values, and represent $E\left((P_n(y))^2\right)$ in a manner similar to how $E(P_n(y))$ was represented in equations (1.7). Do not just say that it is a binomial random variable.

1.4. The variance functional.
   a) Express the variance of a random variable as a functional of its CDF as was done in equation (1.10) for the mean.
   b) What is the same functional of the ECDF?
   c) What is the plug-in estimate of the variance?
   d) What are the statistical properties of the plug-in estimate of the variance? (Is it unbiased? Is it consistent? Is it an MLE?, etc.)

1.5. For a distribution with CDF $F$ and given $\alpha \in (0, 1)$, the functional that defines the $\alpha$ quantile is the quantile function, equation (1.18). What is the plug-in estimator of the $\alpha$ quantile? Is it unbiased for some $\alpha$ and for some nondegenerate distribution?

1.6. Give examples of
   a) a parameter that is defined by a linear functional of the distribution function, and
   b) a parameter that is not a linear functional of the distribution function.
   c) Is the variance a linear functional of the distribution function?
1.7. a) Let $U_1, \ldots, U_n$ with $n \geq 2$ be a random sample from a $U(0,1)$ distribution with order statistics $U_{1:n}, \ldots, U_{n:n}$. Show that if $U_{n:n} = v$ the conditional distribution of $vU_{n-1:n}$, given $U_{n:n} = v$ is the same as the unconditional distribution of $U_{n-1:n}$. (This is a simple special case of expression (1.17).

b) Generate a random sample of 100 $U(0,1)$ variates. Call the order statistics from the sample $u_{1:100}, \ldots, u_{100:100}$. Let $v = u_{51:100}$. Now generate an independent random sample of 50 $U(0,v)$ variates. Call the order statistics from the sample $w_{1:50}, \ldots, w_{50:50}$. Plot $w_{1:50}, \ldots, w_{50:50}$ versus $u_{1:50}, \ldots, u_{50:50}$. What do you observe?

1.8. Assume a random sample of size 10 from a normal distribution. With $\nu = 1 - 2\iota$ in equation (1.20), determine the value of $\iota$ that makes the empirical quantile of the 9th order statistic unbiased for the normal quantile corresponding to 0.90.

1.9. The Harrell-Davis quantile estimator (see Harrell and Davis (1982)) is based on the expected value of the $k$th order statistic in a sample of size $n$, which we can work out using equation 1.15:

$$E(X(k)) = \binom{n}{k} \int_{-\infty}^{\infty} x \left( F(x) \right)^{k-1} \left( 1 - F(x) \right)^{n-k} dF(x).$$

If $F$ in equation 1.15 is invertible, then we can make a change of variables $y = F(x)$ in the integral:

$$E(X(k)) = \binom{n}{k} \int_{0}^{1} F^{-1}(y) y^{k-1} (1 - y)^{n-k} dy.$$

Now, for the $\alpha$ quantile, we consider the quantity $\alpha(n+1)$. As $n$ increases without bound, $E(X(\alpha(n+1))) \to F^{-1}(\alpha)$; hence, we consider $E(X(\alpha(n+1)))$ in the form above. We substitute the continuous analogue of the combinatorial, $\binom{n}{k} = 1/B(k, n-k+1)$, where $B(k, n-k+1)$ is the complete beta function (see Appendix C). Also, in place of the CDF, we use the ECDF. For the expected value of the “$\alpha(n+1)$th order statistic” (note that $\alpha(n+1)$ is not necessarily an integer), we have, in the form of the integral above,

$$x_{\alpha} = \frac{1}{B(\alpha(n+1), (n+1)(1-\alpha))} \int_{0}^{1} F_{n}^{-1}(y) y^{\alpha(n+1)-1} (1 - y)^{(n+1)(1-\alpha)-1} dy.$$

Because $F_{n}^{-1}$ is a step function of the sample values that is constant on the intervals $((i-1)/n, i/n)$ and the steps are the order statistics, this expression can be written as a linear combination of the order statistics:

$$x_{\alpha} = \sum_{i=1}^{n} W_{n,i} x_{(i)},$$

where
\[ W_{n,i} = \frac{1}{B(\alpha(n + 1), (n + 1)(1 - \alpha))} \int_{(i-1)/n}^{i/n} y^{\alpha(n+1)-1} (1 - y)^{(n+1)(1-\alpha)-1} \, dy = P(i/n) - P((i - 1)/n), \]

where \( P(\cdot) \) is the CDF of the beta distribution with parameters \( \alpha(n + 1) \) and \( (n + 1)(1 - \alpha) \).

a) Write a function that accepts a univariate dataset and for given \( \alpha \), computes the Harrell-Davis \( \alpha \) quantile estimate. (The R function `hdquantile` in the `Hmisc` package computes this, but you should write one yourself. All you need are functions to sort and to compute the beta CDF.)

b) Conduct a small Monte Carlo study to compare the Harrell-Davis estimate of median with the sample median as an estimate of the median of a Weibull distribution (see Table D.6 in Appendix D for a definition of the Weibull distribution). The sample median for an odd-sized sample is the central order statistic, and the sample median for an even-sized sample is the average of the two central order statistics. Use sample sizes of 10, 50, and 100. Compare the estimators based on their bias and their variance. The median of the Weibull(\( \alpha, \beta \)) distribution is \( (\beta \log 2)^{1/\alpha} \).

1.10. Generate a sample of pseudorandom numbers from a normal (0,1) distribution and produce a quantile plot of the sample against a normal (0,1) distribution, similar to Figure 1.3 on page 19. Do the tails of the sample seem light? (How can you tell?) If they do not, generate another sample and plot it. Does the erratic tail behavior indicate problems with the random number generator? Why might you expect often (more than 50% of the time) to see samples with light tails?

1.11. Consider the least squares estimator of \( \beta \) in the usual linear regression model, \( E(Y) = X\beta \).

a) Use expression (1.39) on page 28 to derive the variance-covariance matrix for the estimator.

b) Use expression (1.40) to derive the variance-covariance matrix for the estimator.

c) Use both equations 1.39 and 1.40 to estimate the variance-covariance matrix of \( (\beta_0, \beta_1, \beta_2) \).

1.12. a) For \( t = 1 : 1000 \) generate 1,000 observations from the model

\[ y_i = \beta_1 e^{-(\beta_2 t_i + \beta_3 t_i^2)} + \epsilon_i, \]

where \( \beta_1 = 10, \beta_2 = 0.001, \beta_3 = 0.000001 \), and \( \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \) with \( \sigma^2 = 1 \).

b) Use least squares to estimate \( \beta_0, \beta_1, \beta_2 \), and \( \sigma^2 \).

c) Use both equations 1.39 and 1.40 to estimate the variance-covariance matrix of \( (\beta_0, \beta_1, \beta_2) \).

1.13. Assume a random sample \( y_1, \ldots, y_n \) from a gamma distribution with parameters \( \alpha \) and \( \beta \).
a) What are the least squares estimates of \( \alpha \) and \( \beta \)? (Recall \( \text{E}(Y) = \alpha \beta \) and \( \text{V}(Y) = \alpha \beta^2 \).)

b) Write a function in a language such as R, Matlab, or Fortran that accepts a sample of size \( n \) and computes the least squares estimator of \( \alpha \) and \( \beta \) and an approximation of the variance-covariance matrix using both expression (1.39) and expression (1.40).

c) Try out your program in Exercise 1.13b by generating a sample of size 500 from a gamma(2,3) distribution and then computing the estimates. (See Appendix B for information on software for generating random deviates.)

d) Formulate the optimization problem for determining the MLE of \( \alpha \) and \( \beta \). Does this problem have a closed-form solution?

e) Write a function in a language such as R, Matlab, or Fortran that accepts a sample of size \( n \) and computes the maximum likelihood estimator of \( \alpha \) and \( \beta \) and computes an approximation of the variance-covariance matrix using expression (1.56), page 35.

f) Try out your program in Exercise 1.13e by computing the estimates from an artificial sample of size 500 from a gamma(2,3) distribution.


a) Write a program to use IRLS to determine the \( L_p \) estimates of \( \beta \) in the linear regression model \( y = X\beta + E \). The program should accept \( p \), the \( n \)-vector \( y \), and the \( n \times m \) matrix \( X \), and should return the \( L_p \) estimates of \( \beta \) as well as the sample mean squared error and the \( L_p \) norm of the residuals. Make sure that your program handles small residuals well. Residuals that are numerically zero can either be assigned to a very large value or, counter-intuitively, omitted.

b) Let \( n = 100 \), \( m = 3 \), and \( \beta = (1, 1, 1) \). Assume that the model contains an intercept; that is, the first column of \( X \) is constant (all 1’s). Now generate random data for the matrix \( X \). Generate the values independently and identically from a \( U(1, 10) \) distribution. Now generate \( y \) from the model and \( X \) plus an error that is \( \text{N}(0, 1) \). Finally use your program to compute the \( L_1 \), \( L_{1.5} \), \( L_2 \), \( L_3 \) estimates of \( \beta \).

c) Now, take the same data as in Exercise 1.14b, and add 10 to \( y_1 \) and to \( y_2 \). Use your program to compute the \( L_1 \), \( L_{1.5} \), \( L_2 \), \( L_3 \) estimates of \( \beta \), and compare these estimates with the estimates from data without the outliers.

d) Combine the two previous parts into a small Monte Carlo study to compare the performance of the \( L_1 \) and \( L_2 \) estimators in the presence of location outliers.

1.15. For the random variable \( Y \) with a distribution in the exponential family and whose density is expressed in the form of equation (1.46) on page 31, and assuming that the first two moments of \( g(Y) \) exist and \( a(\cdot) \) is twice differentiable, show that
Basic Computations in Statistical Inference

\[ E(g(Y)) = \nabla a(\theta) \]

and

\[ V(g(Y)) = H_{a}(\theta). \]

**Hint:** First show that

\[ E(\nabla \log(p(Y | \theta))) = 0, \]

where the differentiation is with respect to \( \theta \).

1.16. Dempster et al. (1977) consider the multinomial distribution with four outcomes, that is, the multinomial with probability function,

\[
p(x_1, x_2, x_3, x_4) = \frac{n!}{x_1!x_2!x_3!x_4!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3} \pi_4^{x_4},
\]

with \( n = x_1 + x_2 + x_3 + x_4 \) and \( 1 = \pi_1 + \pi_2 + \pi_3 + \pi_4 \). They assumed that the probabilities are related by a single parameter, \( \theta \):

\[
\begin{align*}
\pi_1 &= \frac{1}{2} + \frac{1}{4} \theta \\
\pi_2 &= \frac{1}{4} - \frac{1}{4} \theta \\
\pi_3 &= \frac{1}{4} - \frac{1}{4} \theta \\
\pi_4 &= \frac{1}{4} \theta,
\end{align*}
\]

where \( 0 \leq \theta \leq 1 \). (This model goes back to an example discussed by Fisher, 1925, in Statistical Methods for Research Workers.) Given an observation \((x_1, x_2, x_3, x_4)\), the log-likelihood function is

\[
l(\theta) = x_1 \log(2 + \theta) + (x_2 + x_3) \log(1 - \theta) + x_4 \log(\theta) + c
\]

and

\[
\frac{dl(\theta)}{d\theta} = \frac{x_1}{2 + \theta} - \frac{x_2 + x_3}{1 - \theta} + \frac{x_4}{\theta}.
\]

The objective is to estimate \( \theta \).

a) Determine the MLE of \( \theta \). (Just solve a simple polynomial equation.)

Evaluate the estimate using the data that Dempster, et al. used: \( n = 197 \) and \( x = (125, 18, 20, 34) \).

b) Although the optimum is easily found as in the previous part of this exercise, it is instructive to use Newton’s method (as in equation (1.31) on page 25). Write a program to determine the solution by Newton’s method, starting with \( \hat{\theta}(0) = 0.5 \).

c) Write a program to determine the solution by scoring (which is the quasi-Newton method given in equation (1.52) on page 34), again starting with \( \hat{\theta}(0) = 0.5 \).
d) Write a program to determine the solution by the EM algorithm, again starting with \( \hat{\theta}^{(0)} = 0.5 \).

e) How do these methods compare? (Remember, of course, that this is a particularly simple problem.)

1.17. A common application of EM methods is for estimation of the parameters in a mixture distribution. In a simple instance of such a problem, we have observations \( y_1, \ldots, y_n \) some of which come from population \( P_1 \) and the rest from population \( P_2 \). We can think of the data as pairs, \( (y_i, u_i) \), where the unobserved \( u_i \) is either 1 or 2, depending on whether the observation is from the first or the second population.

Consider the model in which the proportion \( \omega \) of the observations are from a \( N(\mu_1, \sigma_1^2) \) distribution and \( 1-\omega \) of the observations are from a \( N(\mu_2, \sigma_2^2) \) distribution.

a) Let \( \omega = 0.7, \mu_1 = 0, \mu_2 = 2, \) and \( \sigma_1^2 = \sigma_2^2 = 1 \). Generate 1,000 observations from this mixture distribution. (That means, of course, that there may be more or less than 70% from the first distribution, and the observations themselves are not grouped into 2 classes.)

b) Assume you know that \( \sigma_1^2 = \sigma_2^2 = 1 \). Use the EM method to estimate \( \omega, \mu_1, \) and \( \mu_2 \).

c) Now assume you do not know \( \sigma_1^2 \) and \( \sigma_2^2 \). Use the EM method to estimate \( \omega, \mu_1, \mu_2, \sigma_1^2, \) and \( \sigma_2^2 \).

1.18. Consider the simple exponential model for failure times

\[
p(t) = \frac{1}{\theta} e^{-t/\theta},
\]

and data from \( n \) items that are put on test, and for which only the failure times of the first \( n_1 \) items to fail are recorded, \( t_1, \ldots, t_{n_1} \). For the remaining \( n - n_1 \) items all we know are the times at which they were taken from the test (or “censored”), \( c_{n_1+1}, \ldots, c_n \). We have seen that the log-likelihood for the observed data is

\[
l(\theta) = -n_1 \log(\theta) - \left( \sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^{n} c_i \right) / \theta,
\]

and its maximum clearly occurs at

\[
\hat{\theta} = \left( \sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^{n} c_i \right) / n_1.
\]

Formulate the problem in the EM framework, where the sum

\[
\sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^{n} c_i
\]

is estimated in the E step.

*Elements of Computational Statistics, Second Edition ©2013 James E. Gentle*
1.19. Assume that \( \{X_1, X_2\} \) is a random sample of size 2 from an exponential distribution with parameter \( \theta \). Consider the random variable formed as a Student’s t,
\[
T = \frac{\bar{X} - \theta}{\sqrt{S^2/2}},
\]  
(1.66)
where \( \bar{X} \) is the sample mean and \( S^2 \) is the sample variance,
\[
\frac{1}{n-1} \sum (X_i - \bar{X})^2.
\]

a) Show that the distribution of \( T \) is negatively skewed (although the distribution of \( X \) is positively skewed). The distribution is quite complicated. Note that \( n = 2 \) in this exercise.

b) Use Monte Carlo samples to explore the skewness of \( T \) for larger sample sizes, \( n = 10, 20, 100 \). (Note that the denominator in equation (1.66) in general is \( \sqrt{S^2/n} \).)

c) Give a heuristic explanation of the negative skewness of \( T \).

The properties illustrated in the exercise relate to the robustness of statistical procedures that use Student’s \( t \). While those procedures may be robust to some departures from normality, they are often not robust to skewness. These properties also have relevance to the use of statistics like a Student’s \( t \) in the bootstrap.

1.20. A function \( T \) of a random variable \( X \) with distribution parametrized by \( \theta \) is said to be \textit{sufficient} for \( \theta \) if the conditional distribution of \( X \) given \( T(X) \) does not depend on \( \theta \). Discuss (compare and contrast) pivotal and sufficient functions. (Start with the basics: Are they statistics? In what way do they both depend on some universe of discourse, that is, on some family of distributions?)

1.21. Use the pivotal value \( g(\hat{\beta}, \beta) \) in equation (1.65) on page 43 to form a \((1 - \alpha)100\%\) confidence region for \( \beta \) in the usual linear regression model.

1.22. Assume a random sample \( y_1, \ldots, y_n \) from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \). Determine an unbiased estimator of \( \sigma \) based on the sample variance, \( s^2 \). (Note that \( s^2 \) is sufficient and unbiased for \( \sigma^2 \).)

1.23. Why are there twice as many distinct values between 1 and 2 that can be represented in IF as there are between 2 and 3 that can be represented in that system? (You do not need to know the details of the representation to answer this. All you need are the given facts that the representation uses a binary base, and that the mantissa uses a fixed number of bits.)