Environmental Statistics

with S-PLUS

Steven P. Millard
Nagaraj K. Neerchal

CRC Press
Boca Raton  London  New York  Washington, D.C.
Library of Congress Cataloging-in-Publication Data

I. Steven P.
Environmental statistics with S-Plus / Steven P. Millard, Maganraj K. Neerchal. p. cm.—(CRC applied environmental statistics series)
index bibliographical references and index.
ISBN 0-8493-7168-6 (alk. paper)
Environmental sciences—Statistical methods—Data processing. 2. S-Plus.
chad, Maganraj K. II. Title. III. Series.
573 M55 2000
0727—dc21 00-058565
CIP

Visit the CRC Press Web site at www.crcpress.com

© 2001 by CRC Press LLC

No claim to original U.S. Government works

International Standard Book Number: 0-8493-7168-6
Library of Congress Card Number: 00-058565

Printed in the United States of America: 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

PREFACE

The environmental movement of the 1960s and 1970s resulted in the creation of several laws aimed at protecting the environment, and in the creation of Federal, state, and local government agencies charged with enforcing these laws. Most of these laws mandate monitoring or assessment of the physical environment, which means someone has to collect, analyze, and explain environmental data. Numerous excellent journal articles, guidance documents, and books have been published to explain various aspects of applying statistical methods to environmental data analysis. Only a very few books attempt to provide a comprehensive treatment of environmental statistics in general, and this book is an addition to that category.

This book is a survey of statistical methods you can use to collect and analyze environmental data. It explains what these methods are, how to use them, and where you can find references to them. It provides insight into what to think about before you collect environmental data, how to collect environmental data (via various random sampling schemes), and also how to make sense of it after you have it. Several data sets are used to illustrate concepts and methods, and they are available both with software and on the CRC Press Web so that the reader may reproduce the examples. The appendix includes an extensive list of references.

This book grew out of the authors' experiences as teachers, consultants, and software developers. It is intended as both a reference book for environmental scientists, engineers, and regulators who need to collect or make sense of environmental data, and as a textbook for graduate and advanced undergraduate students in an applied statistics or environmental science course. Readers should have a basic knowledge of probability and statistics, but those with more advanced training will find lots of useful information as well.

A unique and powerful feature of this book is its integration with the commercially available software package S-PLUS, a popular and versatile statistics and graphics package. S-PLUS has several add-on modules useful for environmental data analysis, including ENVIRONMENTALSTATS for S-PLUS, S+SPATIALSTATS, and S-PLUS for ArcView GIS. Throughout this book, when a data set is used to explain a statistical method, the commands for and results from the software are provided. Using the software in conjunction with this text will increase the understanding and immediacy of the methods.

This book follows a more or less sequential progression from elementary ideas about sampling and looking at data to more advanced methods of estimation and testing as applied to environmental data. Chapter 1 provides an introduction and overview, Chapter 2 reviews the Data Quality Objectives (DQO) and Data Quality Assessment (DQA) process necessary in the design
# TABLE OF CONTENTS

1 Introduction ...................................................................................... 1  
  Intended Audience........................................................................ 2  
  Environmental Science, Regulations, and Statistics....................... 2  
  Overview......................................................................................... 7  
  Data Sets and Case Studies............................................................ 10  
  Software......................................................................................... 11  
  Summary ......................................................................................... 12  
  Exercises.......................................................................................... 12  

2 Designing a Sampling Program, Part I ............................................. 13  
  The Basic Scientific Method............................................................. 13  
  What is a Population and What is a Sample?.................................... 15  
  Random vs. Judgment Sampling..................................................... 15  
  The Hypothesis Testing Framework............................................... 16  
  Common Mistakes in Environmental Studies................................... 17  
  The Data Quality Objectives Process............................................. 19  
  Sources of Variability and Independence.................................... 24  
  Methods of Random Sampling..................................................... 26  
  Case Study....................................................................................... 42  
  Summary ......................................................................................... 49  
  Exercises.......................................................................................... 51  

3 Looking at Data .................................................................................. 53  
  Summary Statistics......................................................................... 53  
  Graphs for a Single Variable......................................................... 67  
  Graphs for Two or More Variables................................................. 113  
  Summary ......................................................................................... 133  
  Exercises.......................................................................................... 134  

4 Probability Distributions................................................................. 139  
  What is a Random Variable?......................................................... 139  
  Discrete vs. Continuous Random Variable..................................... 140  
  What is a Probability Distribution?............................................... 141  
  Probability Density Function (PDF).............................................. 145  
  Cumulative Distribution Function (CDF)...................................... 153  
  Quantiles and Percentiles............................................................. 158  
  Generating Random Numbers from Probability Distributions....... 161  
  Characteristics of Probability Distributions................................. 162  
  Important Distributions in Environmental Statistics.................... 167  
  Multivariate Probability Distributions.......................................... 194  
  Summary ......................................................................................... 194  
  Exercises.......................................................................................... 195
INTRODUCTION

The environmental movement of the 1960s and 1970s resulted in the creation of several laws aimed at protecting the environment, and in the creation of Federal, state, and local government agencies charged with enforcing these laws. In the U.S., laws such as the Clean Air Act, the Clean Water Act, the Resource Conservation and Recovery Act, and the Comprehensive Emergency Response and Civil Liability Act mandate some sort of monitoring or comparison to ensure the integrity of the environment. Once you start talking about monitoring a process over time, or comparing observations from two or more sites, you have entered the world of numbers and statistics. In fact, more and more environmental regulations are mandating the use of statistical techniques, and several excellent books, guidance documents, and journal articles have been published to explain how to apply various statistical methods to environmental data analysis (e.g., Berthoux and Brown, 1994; Gibbons, 1994; Gilbert, 1987; Helsel and Hirsch, 1992; McBean and Rovers, 1998; Ott, 1995; Piegorsch and Bailer, 1997; ASTM, 1996; USEPA, 1989a,b,c; 1990; 1991a,b,c; 1992a,b,c,d; 1995a,b,c; 1996a,b, 1997a,b). Only a very few books attempt to provide a comprehensive treatment of environmental statistics in general, and even these omit some important topics.

This explosion of regulations and mandated statistical analysis has resulted in at least four major problems.

- Mandated procedures or those suggested in guidance documents are not always appropriate, or may be misused (e.g., Millard, 1987a; Davis, 1994; Gibbons, 1994).
- Statistical methods developed in other fields of research need to be adapted to environmental data analysis, and there is a need for innovative methods in environmental data analysis.
- The backgrounds of people who need to analyze environmental data vary widely, from someone who took a statistics course decades ago to someone with a Ph.D. doing high-level research.
- There is no single software package with a comprehensive treatment of environmental statistics.

This book is an attempt to solve some of these problems. It is a survey of statistical methods you can use to collect and analyze environmental data. It explains what these methods are, how to use them, and where you can find references to them. It provides insight into what to think about before you collect environmental data, how to collect environmental data (via various
ian of the lognormal distribution (see Equation (4.10)), so the mean is smaller than the mean (see Figure 4.12). The second point is coefficient of variation of $X$, only depends on $\sigma$, the standard deviation.

The formula for the skew of a lognormal distribution can be written as:

$$Skew = 3 \cdot CV + CV^3$$  \hspace{1cm} (4.36)

eq et al., 1993). This equation shows that large values of the CV coefficient lie skewed and starts to resemble a normal distribution. Figure 4.20 shows two different lognormal distributions characterized by the mean parameter Lognormal Distribution

-Parameter lognormal distribution is bounded below at 0. The parameter lognormal distribution includes a threshold parameter $\gamma$ which determines the lower boundary of the random variable. That $\xi (X-\gamma)$ has a normal distribution with mean $\mu$ and standard deviation $\sigma$. Then $X$ is said to have a three-parameter lognormal distribution.

The threshold parameter $\gamma$ affects only the location of the three-parameter lognormal distribution; it has no effect on the variance or the shape of the distribution. Note that when $\gamma = 0$, the three-parameter lognormal distribution reduces to the two-parameter lognormal distribution. The three-parameter lognormal distribution is sometimes used in hydrology to model rainfall, stream flow, pollutant loading, etc. (Stedinger et al., 1993).

Binomial Distribution

After the normal distribution, the binomial distribution is one of the most frequently used distributions in probability and statistics. It is used to model the number of occurrences of a specific event in $n$ independent trials. The outcome for each trial is binary: yes/no, success/failure, 1/0, etc. The binomial random variable $X$ represents the number of "successes" out of the $n$ trials. In environmental monitoring, sometimes the binomial distribution is used to model the proportion of observations of a pollutant that exceed some ambient or cleanup standard, or to compare the proportion of detected values at background and compliance units (USEPA, 1989a, Chapters 7 and 8; USEPA, 1989b, Chapter 8; USEPA, 1992b, p. 5-29; Ott, 1995, Chapter 4).

The probability density (mass) function of a binomial random variable $X$ is given by:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$  \hspace{1cm} (4.37)

where $n$ denotes the number of trials and $p$ denotes the probability of "success" for each trial. It is common notation to say that $X$ has a $B(n, p)$ distribution.

The first quantity on the right-hand side of Equation (4.37) is called the binomial coefficient. It represents the number of different ways you can arrange the $x$ "successes" to occur in the $n$ trials. The formula for the binomial coefficient is:

$$\binom{n}{x} = \frac{n!}{x! (n-x)!}$$  \hspace{1cm} (4.38)

The quantity $n!$ is called "n factorial" and is the product of all of the integers between 1 and $n$. That is,
Probability Distributions

\[ n! = n(n-1)(n-2)\cdots21 \] \hfill (4.39)

Figure 4.5 shows the pdf of a \( B(1, 0.5) \) random variable and Figure 4.10 shows the associated cdf. Figure 4.21 and Figure 4.22 show the pdf's of a \( B(10, 0.5) \) and \( B(10, 0.2) \) random variable, respectively.

The Mean and Variance of the Binomial Distribution

The mean and variance of a binomial random variable are:

\[ E(X) = np \]
\[ Var(X) = np(1-p) \] \hfill (4.40)

The average number of successes in \( n \) trials is simply the probability of a success for one trial multiplied by the number of trials. The variance depends on the probability of success. Figure 4.23 shows the function \( f(p) = p(1-p) \) as a function of \( p \). The variance of a binomial random variable is greatest when the probability of success is \( \frac{1}{2} \), and the variance decreases to 0 as the probability of success decreases to 0 or increases to 1.

Variance of Binomial Distribution

Figure 4.23 The variance of a \( B(1, p) \) random variable as a function of \( p \)
more about what all these Greek letters mean later in this chapter about the lognormal distribution. Also, in the next chapter about how we came up with a mean of 0.6 and a coefficient of 0.5).

Relative frequency (density) histogram, the area of the bar is the probability of falling in that interval. Similarly, for a continuous random variable, the area under the curve between 0 and 1, is simply the area under the pdf between these two limits. Mathematically, this is written as:

\[
\Pr (0.75 \leq X \leq 1) = \int_{0.75}^{1} f(x) \, dx
\]  

(4.3)

normal pdf shown in Figure 4.7, the area under the curve between 0.145, so there is a 14.5% chance that the random variable falls into this interval.

**Probability Density Functions**

Display examples of all of the available probability distributions in S-PLUS and ENVIRONMENTALSTATS for S-PLUS. These probability distributions can be used as models for populations. Almost all of these distributions can be derived from some kind of theoretical mathematical models. The binomial distribution for binary outcomes, the Poisson distribution for counts, the Weibull distribution for extreme values, the normal distribution for sums of several random variables, etc. Later in this chapter we will detail probability distributions that are commonly used in environmental statistics.

To produce the binomial pdf shown in Figure 4.5 using the S-PLUS pull-down menu, follow these steps.

1. On the S-PLUS menu bar, make the following menu choices: Environments > Probability Distributions and Random Numbers > Plot Distribution. This will bring up the Plot Distribution Function dialog box.
2. In the Distribution box, choose Lognormal (Alternative). In the mean box, type 0.6. In the cv box, type 0.5. Click OK or Apply.

**Command**

To produce the binomial pdf shown in Figure 4.5 using the S-PLUS Command or Script Window, type this command.

```
pdfplot("binom", list(size=1, prob=0.5))
```

To produce the lognormal pdf shown in Figure 4.7, type this command.

```
pdfplot("lnorm.alt", list(mean=0.6, cv=0.5))
```

Figure 4.8  Probability distributions in S-PLUS and ENVIRONMENTALSTATS for S-PLUS
Figure 4.8 (continued) Probability distributions in S-PLUS and ENVIRONMENTALSTATS for S-PLUS
Figure 4.8 (continued) Probability distributions in S-PLUS and ENVIROMENTALSTATS for S-PLUS
Computing Values of the Probability Density Function

We can use S-PLUS and ENVIRONMENTALSTATS for S-PLUS to compute the value of the pdf for any of the built-in probability distributions. As we saw in equation (4.1), the value of the pdf for the binomial distribution shown in Figure 4.5 is 0.5 for \( x = 0 \) (a tail) and 0.5 for \( x = 1 \) (a head). From (4.2), you can show that for the lognormal distribution shown in Figure 4.7, the values of the pdf evaluated at 0.5, 0.75, and 1 are about 1.67, 0.35, respectively.

To compute the values of the pdf of the binomial distribution shown in Figure 4.5 using the S-PLUS Command or Script Window, type this command:

\[
\text{dbinom}(0:1, \text{size}=1, \text{prob}=0.5)
\]

To compute the values of the pdf of the lognormal distribution shown in Figure 4.7 for the values 0.5, 0.75, and 1, type this command using ENVIRONMENTALSTATS for S-PLUS:

\[
\text{dlnorm.alt}(c(0.5, 0.75, 1), \text{mean}=0.6, \text{cv}=0.5)
\]

CUMULATIVE DISTRIBUTION FUNCTION (CDF)

The cumulative distribution function (cdf) of a random variable \( X \), sometimes called simply the distribution function, is the function \( F \) such that

\[
F(x) = \Pr(X \leq x)
\]

for all values of \( x \). That is, \( F(x) \) is the probability that the random variable \( X \) is less than or equal to some number \( x \). The cdf can also be defined or computed in terms of the probability density function (pdf) \( f \) as

\[
F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(t) \, dt
\]

for a continuous distribution, and for a discrete distribution it is

\[
F(x) = \Pr(X \leq x) = \sum_{x_i \leq x} f(x_i)
\]
4.9 illustrates the relationship between the probability density function and the cumulative distribution function for the lognormal distribution shown in Figure 4.7.

You can use the cdf to compute the probability that a random variable will fall into some specified interval. For example, the probability that a random variable $X$ falls into the interval $[0.75, 1]$ is given by:

\[
\Pr (0.75 \leq X \leq 1) = \int_{0.75}^{1} f (x) \, dx
\]

\[
= \Pr (X \leq 1) - \Pr (X \leq 0.75) + \Pr (X = 0.75)
\]

For a continuous random variable, the probability that $X$ is exactly equal to 0.75 is 0 (because the area under the pdf between 0.75 and 0.75 is 0), but for a discrete random variable there may be a positive probability of $X$ taking on the value 0.75.

Plotting Cumulative Distribution Functions

Figure 4.10 displays the cumulative distribution function for the binomial random variable whose pdf was shown in Figure 4.5. Figure 4.11 displays the cdf for the lognormal random variable whose pdf was shown in Figure 4.7.

We can see from Figure 4.10 that the cdf of a binomial random variable is a step function (which is also true of any discrete random variable). The cdf is 0 until it hits $x = 0$, at which point it jumps to 0.5 and stays there until it hits $x = 1$, at which point it stays at 1 for all values of $x$ at 1 and greater.

On the other hand, the cdf for the lognormal distribution shown in Figure 4.11 is a smooth curve that is 0 below $x = 0$, and rises towards 1 as $x$ increases.

Menu

To produce the binomial cdf shown in Figure 4.10 using the ENVIRONMENTALSTATS for S-PLUS pull-down menu, follow these steps.
7  HYPOTHESIS TESTS

Comparing Groups to Standards and One Another

In Chapters 2 and 6, we introduced the idea of the hypothesis testing framework. In Chapter 6 we discussed three tools you can use to make an objective decision about whether contamination is present or not: prediction intervals, tolerance intervals, and control charts. In this chapter, we provide a full discussion of the statistical hypothesis testing framework, discuss the relationship between confidence intervals and formal hypothesis tests, and discuss hypothesis tests to make inferences about a single population and compare two or more populations.

THE HYPOTHESIS TESTING FRAMEWORK

We introduced the hypothesis testing framework back in Chapter 2. Our first example involved deciding whether to wear a jacket or not, and our decision depended on our belief about whether it would rain that day (Table 2.1). Our second example involved deciding whether a site or well is contaminated or not (Table 2.2). Table 7.1 below reproduces Table 2.2. In this case, the null hypothesis is that no contamination is present.

<table>
<thead>
<tr>
<th>Reality</th>
<th>No Contamination</th>
<th>Contamination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your Decision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contamination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mistake: Type I Error (Probability = α)</td>
<td>Correct Decision (Probability = 1−β)</td>
<td></td>
</tr>
<tr>
<td>No Contamination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct Decision</td>
<td>Mistake: Type II Error (Probability = β)</td>
<td></td>
</tr>
</tbody>
</table>

Table 7.1  Hypothesis testing framework for deciding on the presence of contamination in the environment when the null hypothesis is "no contamination"

In Step 5 of the DQO process (see Chapter 2), you usually link the principal study question you defined in Step 2 with some population parameter such as the mean, median, 95th percentile, etc. For example, if the study question is "Is the concentration of 1,2,3,4-tetrachlorobenzenes (TcCB) in the
Cleanup site significantly above background levels?" then you may
reformulate this question as "Is the average concentration of TcCB
at the Cleanup site greater than the average concentration of TcCB
at a Reference site?"

A hypothesis test or significance test is a formal mathematical mecha-
nism for objectively making a decision in the face of uncertainty, and is usu-
ally used to answer a question about the value of a population parameter. A
d hypothesis test about a population parameter \( \theta \) (theta) is used to
test the null hypothesis

\[
H_0 : \theta = \theta_0
\] (7.1)

e two-sided alternative hypothesis

\[
H_a : \theta \neq \theta_0
\] (7.2)

(horned "H-naught") denotes the null hypothesis that the true
mean is equal to some specified value \( \theta_0 \). A lower one-

The distribution of the t-statistic under the null hypothesis is shown in Figure 5.9 in Chapter 5 for sample sizes of
5, 10, and \( \infty \). On the other hand, if the true mean \( \mu \) is larger than \( \mu_0 \), then
the sample mean is bouncing around \( \mu \), the numerator of the t-statistic is
bouncing around some positive number. So if the t-statistic is "large" we will probably reject the null hy-
thesis in favor of the alternative hypothesis.

Parametric vs. Nonparametric Tests

For a parametric test, the test statistic \( T \) is usually some estimator of \( \theta \)
(possibly shifted by subtracting a number and scaled by dividing by a num-
ber), and the distribution of \( T \) under the null hypothesis depends on the
distribution of the population (e.g., normal, lognormal, Poisson, etc.). For a
nonparametric or distribution-free test, \( T \) is usually based on the ranks
of the data in the random sample, and the distribution of \( T \) under the null hy-
thesis does not depend on the distribution of the population.

For example, for a two-sample t-test (see below), the test statistic is a
scaled version of the difference between the two sample means, and both
populations are assumed to be normally distributed. For the Wilcoxon rank
sum test (see below), the test statistic is the sum of the ranks in the first sam-
ple, and the distribution of this statistic under the null hypothesis does not
depend on the distribution of the two populations.
Type I and Type II Errors (Significance Level and Power)

As stated above, a hypothesis test involves using a test statistic computed from data collected from an experiment to make a decision. A test statistic is a random quantity (e.g., some expression involving the sample mean); if you at the experiment or get new observations, you will often get a different value for the test statistic. Because you are making your decision based on value of a random quantity, you will sometimes make the "wrong" decision. Table 7.2 illustrates the general hypothesis testing framework; simply a generalization of Table 7.1 above.

<table>
<thead>
<tr>
<th>Your Decision</th>
<th>Reality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>$H_0$ True: Correct Decision (Probability = $1 - \beta$)</td>
</tr>
<tr>
<td></td>
<td>$H_0$ False: Mistake: Type I Error (Probability = $\alpha$)</td>
</tr>
<tr>
<td>Do Not Reject $H_0$</td>
<td>$H_0$ True: Correct Decision</td>
</tr>
<tr>
<td></td>
<td>$H_0$ False: Mistake: Type II Error (Probability = $\beta$)</td>
</tr>
</tbody>
</table>

Table 7.2 The framework of a hypothesis test

P-Values

When you perform a hypothesis test, you usually compute a quantity called the p-value. The p-value is the probability of seeing a test statistic as extreme or more extreme than the one you observed, assuming the null hypothesis is true. Thus, if the p-value is less than or equal to the specified value of $\alpha$ (the Type I error level), you reject the null hypothesis, and if the p-value is greater than $\alpha$, you do not reject the null hypothesis. For hypothesis tests where the test statistic has a continuous distribution, under the null hypothesis (i.e., if $H_0$ is true), the p-value is uniformly distributed between 0 and 1. When the test statistic has a discrete distribution, the p-value can take on only a discrete number of values.

To get an idea of the relationship between p-values and Type I errors, consider the following example. Suppose your friend is a magician and she has a fair coin (i.e., the probability of a "head" and the probability of a "tail" are both 50%) and a coin with beads on both sides (so the probability of a head is 100% and the probability of a tail is 0%). She takes one of these coins out of her pocket, begins to flip it several times, and tells you the outcome after each flip. You have to decide which coin she is flipping. Of course, if a flip comes up tails, then you automatically know she is flipping the two-beaded coin. If you observe five heads in a row, then the p-value associated with this outcome is 0.0312; that is, there is a probability of 3.12% of getting five heads in a row when you flip a fair coin five times. Therefore, the Type I error rate associated with your decision rule is 0.0312. If you want to create a decision rule for which the Type I error rate is lower, you may have to increase the sample size of the experiment (usually denoted $n$), the variability inherent in the data (usually denoted $\sigma$), and the magnitude of the difference between the null and alternative hypothesis (usually denoted $\delta$ or $\Delta$). Conventional choices for $\alpha$ are 1%, 5%, and 10%, but these choices should be made in the context of balancing the cost of a Type I and Type II error. For most hypothesis tests, there is a well-defined relationship between $\alpha$, $\beta$, $n$, and the scaled difference $\delta/\sigma$ (see Chapter 8). A very important fact is that for a specified sample size, if you reduce the Type I error, then you increase the Type II error, and vice-versa.
rate is no greater than 1%, then you will have to wait until you see the
ue of seven flips. If your decision rule is to reject the null hypothesis
seeing \( T = 7 \) heads in a row, then the actual Type I error rate is 0.78%.
when you see seven heads in a row, the p-value is 0.0078.

<table>
<thead>
<tr>
<th># Heads in a Row (7)</th>
<th>Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>12.5</td>
</tr>
<tr>
<td>4</td>
<td>6.25</td>
</tr>
<tr>
<td>5</td>
<td>3.12</td>
</tr>
<tr>
<td>6</td>
<td>1.56</td>
</tr>
<tr>
<td>7</td>
<td>0.78</td>
</tr>
<tr>
<td>8</td>
<td>0.39</td>
</tr>
<tr>
<td>9</td>
<td>0.20</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
</tr>
</tbody>
</table>

7.3 The probability of seeing \( T \) heads in a row in \( n \) flips of a fair coin

In this example, the power associated with your decision rule is the
bility of correctly deciding your friend is flipping the two-headed coin
fact that is the one she is flipping. This is a special example in
power is equal to 100%, because if your friend really is flipping
headed coin then you will always see a head on each flip and no mat-
value of \( T \) you choose for the cut-off, you will always see \( T \) heads
Usually, however, there is an inverse relationship between the Type
Type II error, so that the smaller you set the Type I error, the smaller
wer of the test (see Chapter 8).

nsider the null hypothesis shown in Equation (7.1), where \( \theta \) is some
ation parameter of interest (e.g., mean, proportion, 95th percentile, etc.).
is a one-to-one relationship between hypothesis tests concerning \( \theta \) and
ence interval for this parameter. A \((1-\alpha)\)100% confidence interval
ists of all possible values of \( \theta \) that are associated with not rejecting
1 hypothesis at significance level \( \alpha \). Thus, if you know how to create
ence interval for a parameter, you can perform a hypothesis test for
rayer, and vice-versa. Table 7.4 shows the explicit relationship be-
hypothesis tests and confidence intervals.

Never you report the results of a hypothesis test, you should almost
report the corresponding confidence interval as well. This is because
have a small sample size, you may not have much power to uncover
that the null hypothesis is not true, even if there is a huge difference
the postulated value of \( \theta \) (e.g., \( \theta = 5 \) ppb) and the true value of \( \theta \)
(e.g., \( \theta = 20 \) ppb). On the other hand, if you have a large sample size, you
may be very likely to detect a small difference between the postulated value
of \( \theta \) (e.g., \( \theta = 5 \) ppb) and the true value of \( \theta \) (e.g., \( \theta = 6 \) ppb), but this dif-
ference may not really be important to detect. Confidence intervals help
you sort out the important distinction between a statistically significant
difference and a scientifically meaningful difference.

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Alternative Hypothesis</th>
<th>Corresponding Confidence Interval</th>
<th>Rejection Rule Based on CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-sided</td>
<td>( \theta \neq \theta_0 )</td>
<td>Two-sided ([LCL, UCL])</td>
<td>( LCL &gt; \theta_0 ), or ( UCL &lt; \theta_0 )</td>
</tr>
<tr>
<td>Lower</td>
<td>( \theta &lt; \theta_0 )</td>
<td>Upper ([\infty, UCL])</td>
<td>( UCL &lt; \theta_0 )</td>
</tr>
<tr>
<td>Upper</td>
<td>( \theta &gt; \theta_0 )</td>
<td>Lower ([LCL, \infty])</td>
<td>( LCL &gt; \theta_0 )</td>
</tr>
</tbody>
</table>

Table 7.4 Relationship between hypothesis tests and confidence intervals

OVERVIEW OF UNIVARIATE HYPOTHESIS TESTS

Table 7.5 summarizes the kinds of univariate hypothesis tests that we will
talk about in this chapter. In Chapter 9 we will talk about hypothesis tests
for regression models. We will not discuss hypothesis tests for multivariate
observations. A good introduction to multivariate statistical analysis is John-

<table>
<thead>
<tr>
<th>One-Sample</th>
<th>Two-Samples</th>
<th>Multiple Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goodness-of-Fit</td>
<td>Proportions</td>
<td>Proportions</td>
</tr>
<tr>
<td>Proportion</td>
<td>Locations</td>
<td>Locations</td>
</tr>
<tr>
<td>Location</td>
<td>Variability</td>
<td>Variability</td>
</tr>
</tbody>
</table>

Table 7.5 Summary of the kinds of hypothesis tests discussed in this chapter

GOODNESS-OF-FIT TESTS

Most commonly used parametric statistical tests assume the observations
in the random sample(s) come from a normal population. So how do you
know whether this assumption is valid? We saw in Chapter 3 how to make a
visual assessment of this assumption using Q-Q plots. Another way to verify
this assumption is with a goodness-of-fit test, which lets you specify what
kind of distribution you think the data come from and then compute a test
statistic and a p-value.
goodness-of-fit test may be used to test the null hypothesis that the data come from a specific distribution, such as "the data come from a normal distribution with mean 10 and standard deviation 2," or to test the more general hypothesis that the data come from a particular family of distributions such as "the data come from a lognormal distribution." Goodness-of-fit tests are mostly used to test the latter kind of hypothesis, since in practice we only know or want to specify the parameters of the distribution.

In practice, goodness-of-fit tests may be of limited use for very large or very small sample sizes. Almost any goodness-of-fit test will reject the null hypothesis of the specified distribution if the number of observations is very large since "real" data are never distributed according to any theoretical distribution (Conover, 1980, p. 367). On the other hand, with only a very small number of observations, no test will be able to determine whether the observations appear to come from the hypothesized distribution or some other different looking distribution.

For Normality

Two commonly used tests to test the null hypothesis that the observations come from a normal distribution are the Shapiro-Wilk test (Shapiro and Wilk, 1965), and the Shapiro-Francia test (Shapiro and Francia, 1972). The Shapiro-Wilk test is more powerful at detecting short-tailed (platykurtic) and long-tailed distributions, and less powerful against symmetric, moderately long-tailed (leptokurtic) distributions. Conversely, the Shapiro-Francia test is more powerful against symmetric long-tailed distributions and less powerful against short-tailed distributions (Royston, 1992a; 1993). These tests are considered to be two of the very best tests of normality available (Staiano, 1986b, p. 406).

**Shapiro-Wilk Test**

Shapiro-Wilk test statistic can be written as:

\[
W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)} \right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

(7.8)

\(x_{(i)}\) denotes the \(i^{th}\) ordered observation, \(a_i\) is the \(i^{th}\) element of the vector \(a\), and the vector \(a\) is defined by:

\[
a^T = \eta^T \eta^{-1} / \sqrt{\eta^T \eta^{-1} \eta}
\]

(7.9)

where \(T\) denotes the transpose operator, and \(\eta\) is the vector of expected values and \(V\) is the variance-covariance matrix of the order statistics of a random sample of size \(n\) from a standard normal distribution. That is, the values of \(a\) are the expected values of the standard normal order statistics weighted by their variance-covariance matrix, and normalized so that \(a^T a = 1\). It can be shown that the \(W\)-statistic in Equation (7.8) is the same as the square of the sample correlation coefficient between the vectors \(a\) and \(x_{(i)}\):

\[
W = \left( \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) (x_{(i)} - \bar{x}) \right)^2
\]

(7.10)

where

\[
x = \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \bar{x}) y \right) \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)
\]

(7.11)

(see Chapter 9 for an explanation of the sample correlation coefficient).

Small values of \(W\) yield small p-values and indicate the null hypothesis of normality is probably not true. Royston (1992a) presents an approximation for the coefficients \(a\) necessary to compute the Shapiro-Wilk \(W\)-statistic, and also a transformation of the \(W\)-statistic that has approximately a standard normal distribution under the null hypothesis. Both of these approximations are used in ENVIRONMENTALSTATS for S-PLUS.

**The Shapiro-Francia Test**

Shapiro and Francia (1972) introduced a modification of the \(W\)-test that depends only on the expected values of the order statistics (\(\eta\)) and not on the variance-covariance matrix (\(V\)):

\[
W' = \left( \frac{1}{n} \sum_{i=1}^{n} b_i x_{(i)} \right)^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

(7.12)
Environmental Statistics with S-PLUS

$b_i$ is the $i^{th}$ element of the vector $b$ defined as:

$$b = \sqrt{\frac{n}{m}} m$$  \hspace{1cm} (7.13)

Authors, including Ryan and Joiner (1973), Filliben (1975), and Weisberg and Bingham (1975), note that the $W'$-statistic is intuitively appealing because it is the squared sample correlation coefficient associated with a normal probability plot. That is, if it is the squared correlation between ordered sample values $x_i$ and the expected normal order statistics $\bar{x}_i$:

$$W' = \left\{ x \left[ b, x_i \right] \right\}^2 = \left\{ x \left[ m, \bar{x}_i \right] \right\}^2$$  \hspace{1cm} (7.14)

Weisberg and Bingham (1975) introduced an approximation of the Shapiro-Wilk statistic $W$ that is easier to compute. They suggested using the Blom (1958, pp. 68-75; see Chapter 3) to approximate the elements of $m$:

$$\tilde{W} = \left\{ \frac{\sum_{i=1}^{n} c_i x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right\}^2 = \left\{ x \left[ \bar{x}; \bar{x}_i \right] \right\}^2$$  \hspace{1cm} (7.15)

$c_i$ is the $i^{th}$ element of the vector $c$ defined by:

$$c = \sqrt{\frac{n}{m}} m$$  \hspace{1cm} (7.16)

$$\bar{m} = \Phi^{-1} \left( \frac{i - 3/8}{n + 1/4} \right)$$  \hspace{1cm} (7.17)

and $\Phi$ denotes the standard normal cdf. That is, the values of the elements of $m$ in Equation (7.13) are replaced with their estimates based on the usual Blom plotting positions for a normal distribution (see Chapter 3).

Filliben (1975) proposed the probability plot correlation coefficient (PPCC) test that is essentially the same test as the test of Weisberg and Bingham (1975), but Filliben used different plotting positions. Looney and Gulledge (1985) investigated the characteristics of Filliben's PPCC test using various plotting position formulas and concluded that the PPCC test based on Blom plotting positions performs slightly better than tests based on other plotting positions. The Weisberg and Bingham (1975) approximation to the Shapiro-Francia $W'$-statistic is the square of Filliben's PPCC test statistic based on Blom plotting positions. Royston (1992c) provides a method for computing p-values associated with the Weisberg-Bingham approximation to the Shapiro-Francia $W'$-statistic, and this method is implemented in ENVIRONMENTALSTATS for S-PLUS.

The Shapiro-Wilk and Shapiro-Francia tests can be used to test whether observations appear to come from a normal distribution, or the transformed observations (e.g., Box-Cox transformed) come from a normal distribution. Hence, these tests can test whether the data appear to come from a normal, lognormal, or three-parameter lognormal distribution for example, as well as a zero-modified normal or zero-modified lognormal distribution.

**Example 7.1: Testing the Normality of the Reference Area TcCB Data**

In Chapter 3 we saw that the Reference area TcCB data appear to come from a lognormal distribution based on histograms (Figures 3.1, 3.2, 3.10, and 3.11), an empirical cdf plot (Figure 3.16), normal Q-Q plots (Figures 3.18 and 3.19), Tukey mean-difference Q-Q plots (Figures 3.21 and 3.22), and a plot of the PPCC vs. $\lambda$ for a variety of Box-Cox transformations (Figure 3.24). Here we will formally test whether the Reference area TcCB data appear to come from a normal or lognormal distribution.

<table>
<thead>
<tr>
<th>Assumed Distribution</th>
<th>Shapiro-Wilk ($W$)</th>
<th>Shapiro-Francia ($W'$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.918 (p=0.003)</td>
<td>0.923 (p=0.006)</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.979 (p=0.006)</td>
<td>0.987 (p=0.006)</td>
</tr>
</tbody>
</table>

Table 7.6 Results of tests for normality and lognormality for the Reference area TcCB data

Table 7.6 lists the results of these two tests. The second and third columns show the test statistics with the p-values in parentheses. The p-values clearly indicate that we should not assume the Reference area TcCB data come from a normal distribution, but the assumption of a lognormal distri-