PLIED ENVIRONMENTAL STATISTICS

Steven P. Millard, *Series Editor*

·onmental Statistics with S-PLUS
ı P. Millard and Nagaraj K. Neerchal

ʼHCOMING TITLES

ndwater Monitoring and Regulations
›s Davis

stical Tools for Environmental Quality
el E. Ginevan and Douglas E. Splitstone

Environmental Statistics

with S-PLUS

Steven P. Millard
Nagaraj K. Neerchal

CRC

CRC Press
Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

l, Steven P.
/ironmental statistics with S-Plus / Steven P. Millard, Nagaraj K. Neerchal.
p. cm.—(CRC applied environmental statistics series)
ludes bibliographical references and index.
iN 0-0893-7168-6 (alk. paper)
Environmental sciences—Statistical methods—Data processing. 2. S-Plus.
chal, Nagaraj K. II. Title. III. Series.

373 M55 2000
107'27—dc21) 00-058565
CIP

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

International Standard Book Number 0-0893-7168-6
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

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; 1994a,b,c; 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.

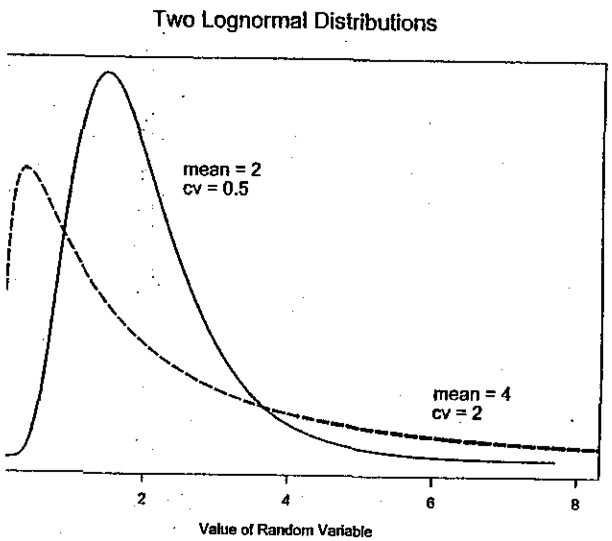
- Mandated procedures or those suggested in guidance documents are not always appropriate, or may be misused (e.g., Millard, 1987a; Davis, 1994; Gibbons, 1994).
- Statistical methods developed in other fields of research need to be adapted to environmental data analysis, and there is a need for innovative 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 me-
'ays smaller than the mean (see Figure 4.12). The second point is
coefficient of variation of X, only depends on σ, the standard de-
Y.
:mula for the skew of a lognormal distribution can be written as:

$$Skew = 3\,CV + CV^3 \tag{4.36}$$

et al., 1993). This equation shows that large values of the CV
: to very skewed distributions. As τ gets small, the distribution
:ss skewed and starts to resemble a normal distribution. Figure
; two different lognormal distributions characterized by the mean
:.



4.20 Probability density functions for two lognormal distributions

ameter Lognormal Distribution

-parameter lognormal distribution is bounded below at 0. The
ıeter lognormal distribution includes a threshold parameter γ
ıt determines the lower boundary of the random variable. That
$\cdot g(X-\gamma)$ has a normal distribution with mean μ and standard de-
ıen X is said to have a three-parameter lognormal distribution.

The threshold parameter γ 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:

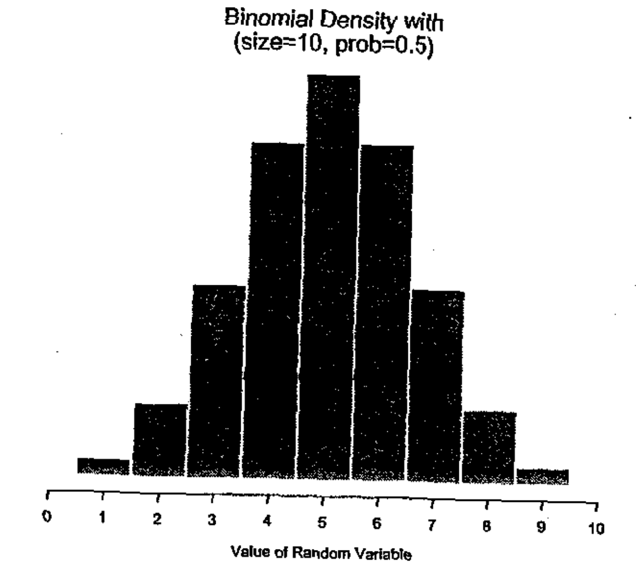
$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n \tag{4.37}$$

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.

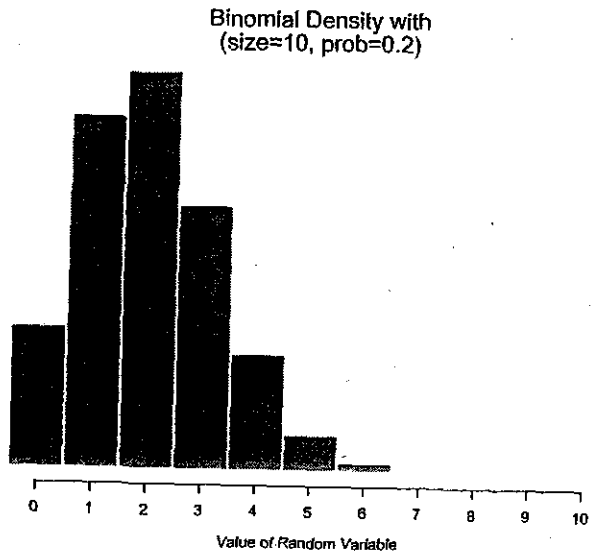
The first quantity on the right-hand side of Equation (4.37) is called the binomial coefficient. It represents the number of different ways you can arrange the x "successes" to occur in the n trials. The formula for the binomial coefficient is:

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!} \tag{4.38}$$

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



re 4.21 Probability density function of a B(10, 0.5) random variable



re 4.22 Probability density function of a B(10, 0.2) random variable

$$n! = n(n-1)(n-2)\cdots 2\ 1 \tag{4.39}$$

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:

$$\begin{aligned} E(X) &= np \\ Var(X) &= np(1-p) \end{aligned} \tag{4.40}$$

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 ½, and the variance decreases to 0 as the probability of success decreases to 0 or increases to 1.

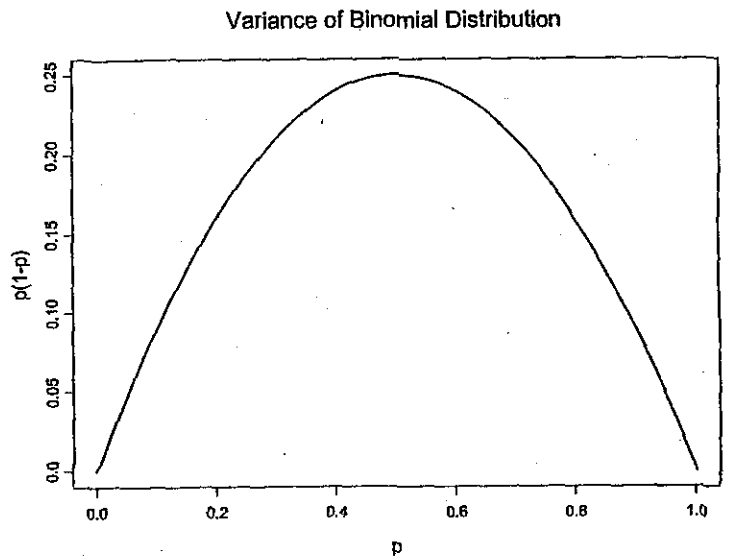


Figure 4.23 The variance of a B(1, p) random variable as a function of p

‹ more about what all these Greek letters mean later in this chap- talk 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).

lative frequency (density) histogram, the area of the bar is the of falling in that interval. Similarly, for a continuous random probability that the random variable falls into some interval, say 5 and 1, is simply the area under the pdf between these two in- ints. Mathematically, this is written as:

$$\Pr\left(0.75 \le X \le 1\right) = \int_{0.75}^{1} f\left(x\right) dx \qquad (4.3)$$

ormal pdf shown in Figure 4.7, the area under the curve between s about 0.145, so there is a 14.5% chance that the random vari- into this interval.

obability Density Functions

8 display examples of all of the available probability distribu- LUS and ENVIRONMENTALSTATS for S-PLUS. These probability can be used as models for populations. Almost all of these dis- n be derived from some kind of theoretical mathematical model omial distribution for binary outcomes, the Poisson distribution ents, the Weibull distribution for extreme values, the normal dis- sums of several random variables, etc.). Later in this chapter we in detail probability distributions that are commonly used in en- statistics.

uce the binomial pdf shown in Figure 4.5 using the NTALSTATS for S-PLUS pull-down menu, follow these steps.

the S-PLUS menu bar, make the following menu choices: **ronmentalStats>Probability Distributions and Random ıbers>Plot Distribution.** This will bring up the Plot Distribu- Function dialog box.
e Distribution box, choose **Binomial**. In the size box, type **1**. In rob box, type **0.5**.
: **OK** or **Apply**.

ne lognormal pdf shown in Figure 4.7, follow these steps.

1. On the S-PLUS menu bar, make the following menu choices: **EnvironmentalStats>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 ENVIRONMENTALSTATS for 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))
```

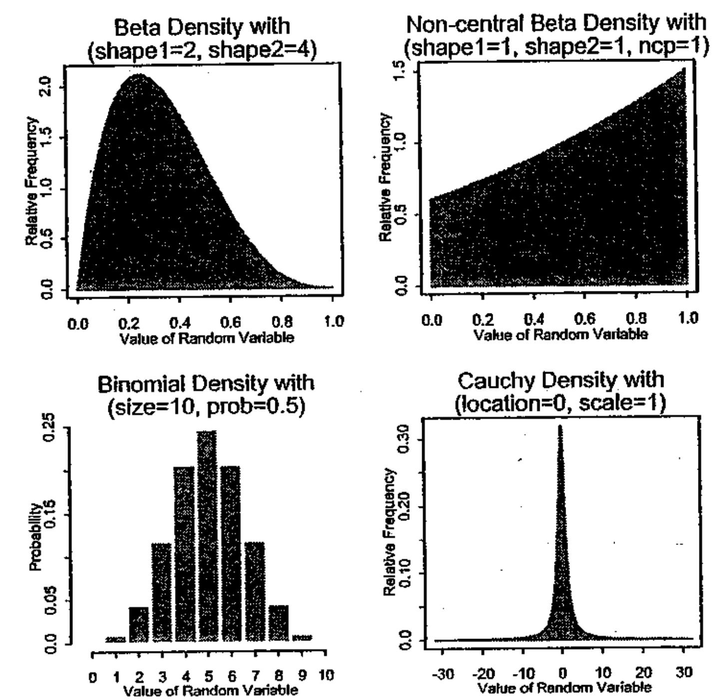


Figure 4.8 Probability distributions in S-PLUS and ENVIRONMENTALSTATS for S-PLUS

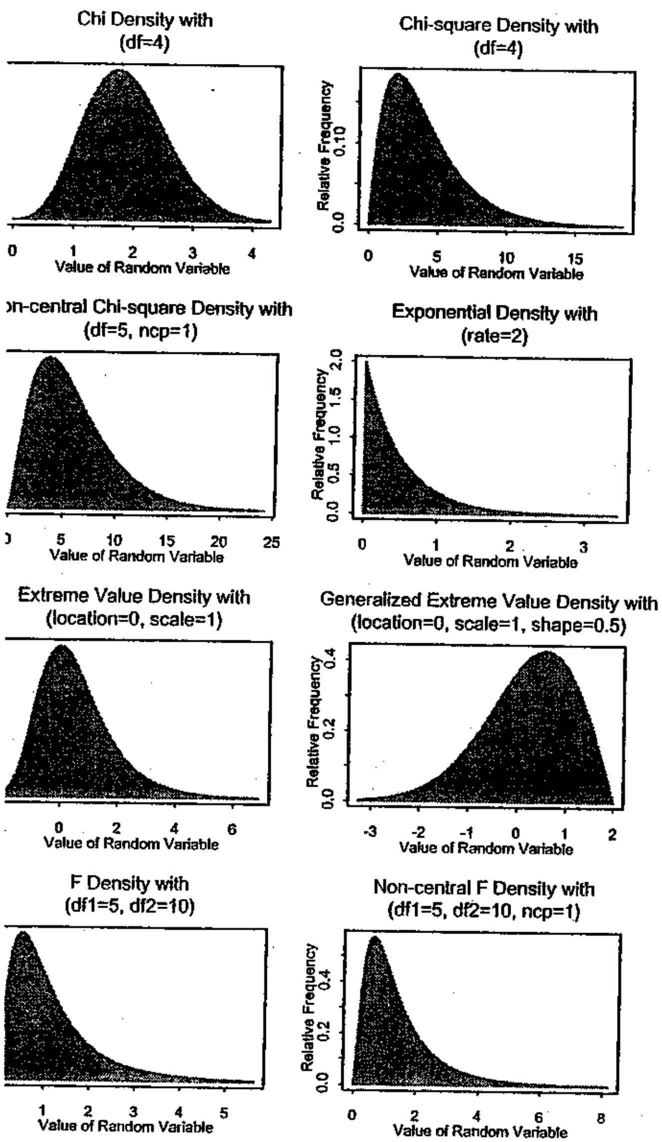


Figure 4.8 (continued) Probability distributions in S-PLUS and ENVIRONMENTALSTATS for S-PLUS

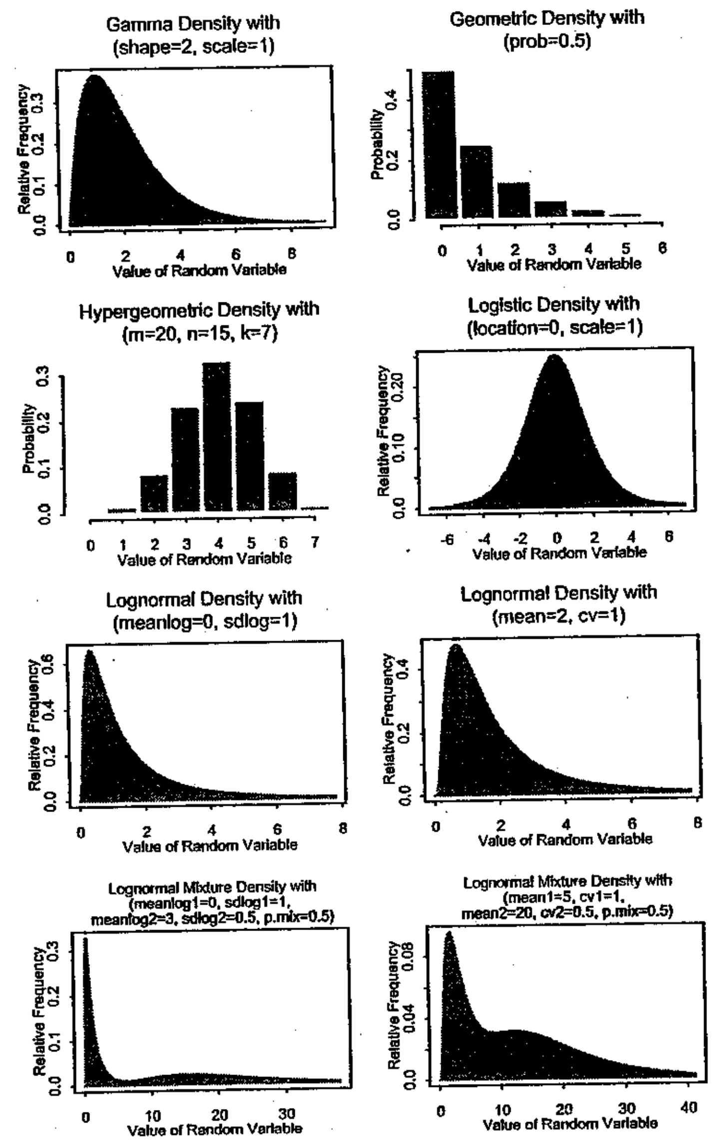
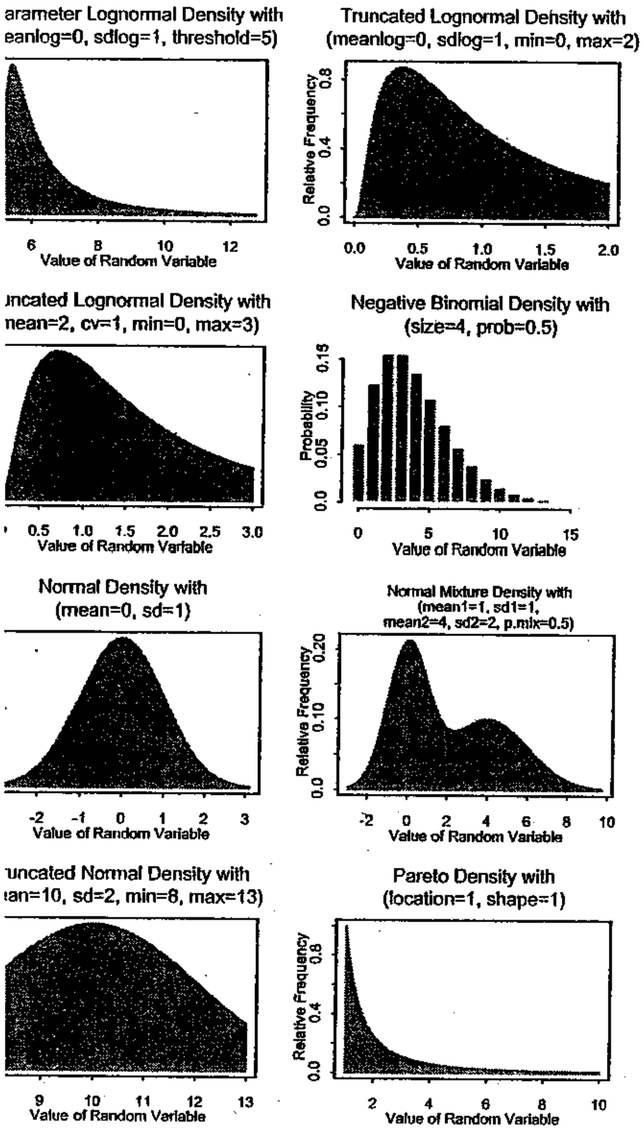


Figure 4.8 (continued) Probability distributions in S-PLUS and ENVIRONMENTALSTATS for S-PLUS



igure 4.8 (continued) Probability distributions in S-PLUS and ENVIRONMENTALSTATS for S-PLUS

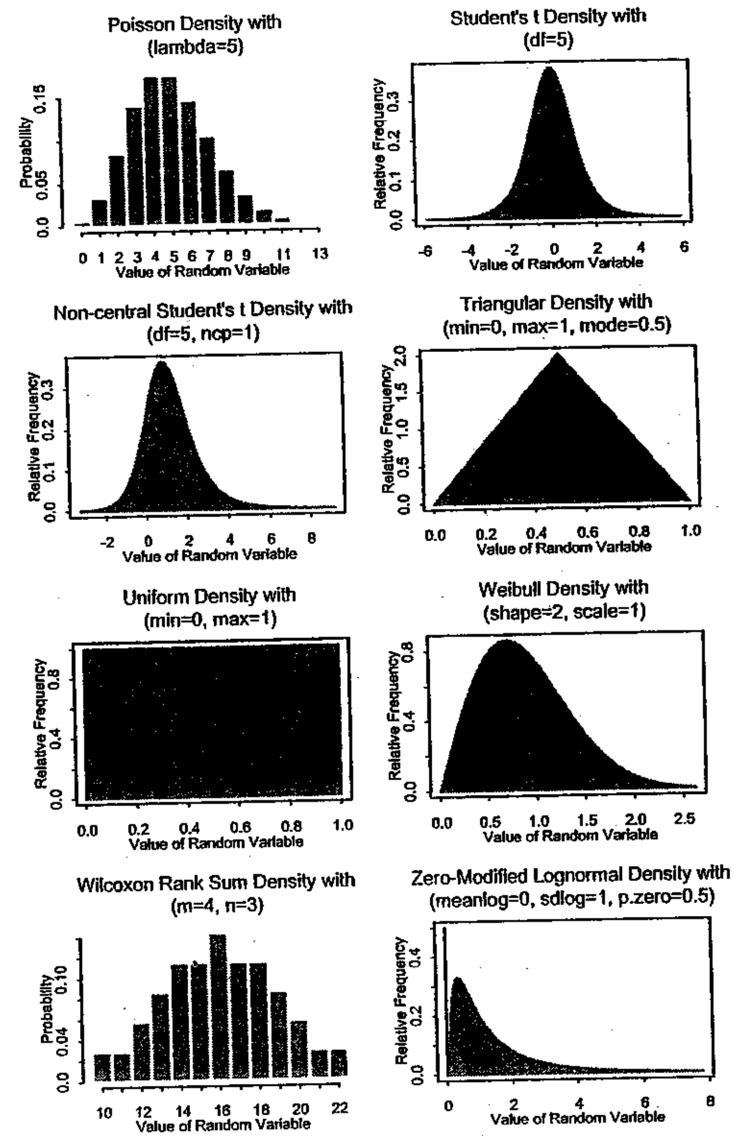


Figure 4.8 (continued) Probability distributions in S-PLUS and ENVIRONMENTALSTATS for S-PLUS

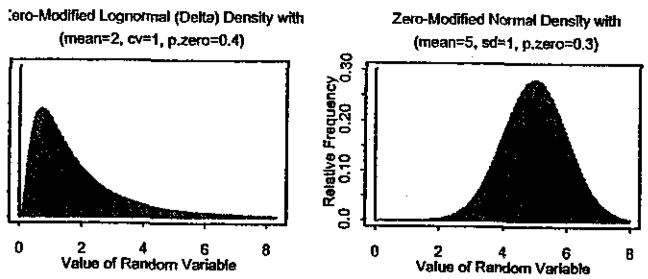


Figure 4.8 (continued) Probability distributions in S-PLUS and ENVIRONMENTALSTATS for S-PLUS

ting Values of the Probability Density Function

:an use S-PLUS and ENVIRONMENTALSTATS for S-PLUS to compute : of the pdf for any of the built-in probability distributions. As we Equation (4.1), the value of the pdf for the binomial distribution . 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 7, the values of the pdf evaluated at 0.5, 0.75, and 1 are about 1.67, : 0.35, respectively.

ımpute the values of the pdf of the binomial distribution shown in .5 using the ENVIRONMENTALSTATS for S-PLUS pull-down menu, ese steps.

)n the S-PLUS menu bar, make the following menu choices: **:nvironmentalStats>Probability Distributions and Random ıumbers>Density, CDF, Quantiles.** This will bring up the Densi- ies, Cumulative Probabilities, or Quantiles dialog box.
'or the Data to Use buttons, choose **Expression.** In the Expression ıox, type **0:1**. In the Distribution box, choose **Binomial.** In the size ıox type **1**. In the prob box type **0.5**. Under the Probability or)uantile group, make sure the **Density** box is checked.
:lick **OK** or **Apply**.

ute the values of the pdf of the lognormal distribution shown in 7 for the values 0.5, 1, and 1.5, follow these steps.

)n the S-PLUS menu bar, make the following menu choices: **:nvironmentalStats>Probability Distributions and Random**

Numbers>Density, CDF, Quantiles. This will bring up the Densities, Cumulative Probabilities, or Quantiles dialog box.

2. For the Data to Use buttons, choose **Expression.** In the Expression box, type **c(0.5, 0.75, 1)**. In the Distribution box, choose **Lognormal (Alternative).** In the mean box type **0.6**. In the cv box type **0.5**. Under the Probability or Quantile group, make sure the Density box is checked.
3. Click **OK** or **Apply**.

Command

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.

```
dbinom(0:1, size=1, 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.

```
dlnorm.alt(c(0.5, 0.75, 1), mean=0.6, 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

$$F(x) = \Pr(X \le x) \tag{4.4}$$

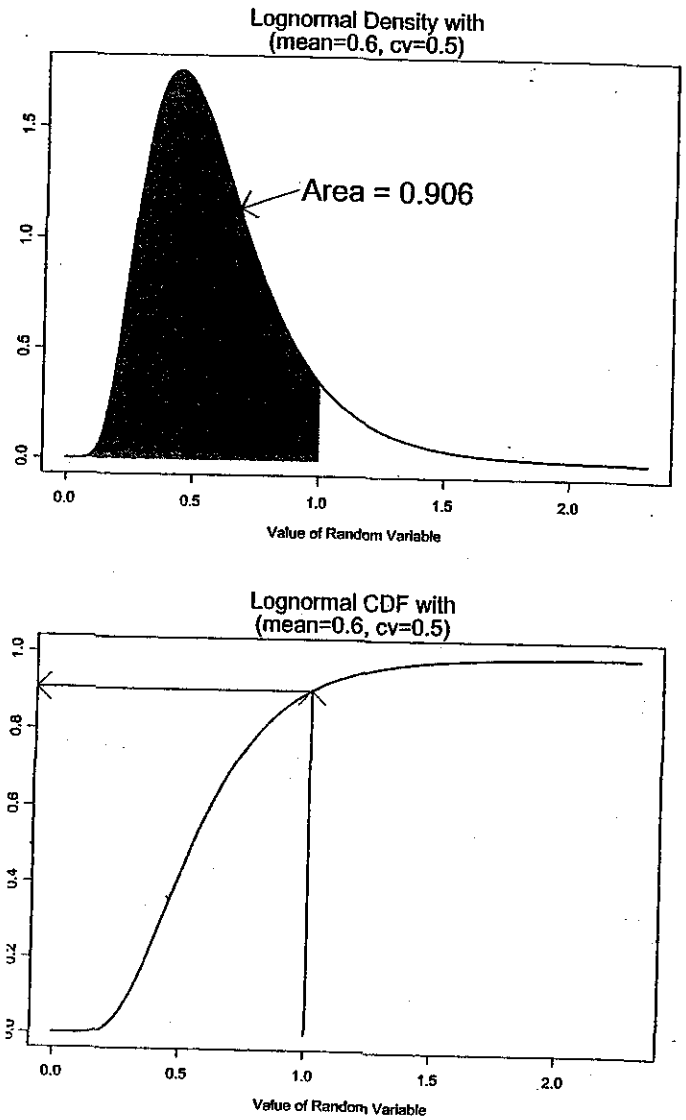
for all values of x. That is, $F(x)$ is the probability that the random variable X is less than or equal to some number x. The cdf can also be defined or computed in terms of the probability density function (pdf) f as

$$F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(t)\,dt \tag{4.5}$$

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

$$F(x) = \Pr(X \le x) = \sum_{x_i \le x} f(x_i) \tag{4.6}$$

ıre 4.9 illustrates the relationship between the probability density func- and the cumulative distribution function for the lognormal distribution vn in Figure 4.7.



4.9 Relationship between the pdf and the cdf for a lognormal distribution

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:

$$\Pr(0.75 \le X \le 1) = \int_{0.75}^{1} f(x)\,dx$$

$$= \Pr(X \le 1) - \Pr(X \le 0.75) + \Pr(X = 0.75) \qquad (4.7)$$

$$= F(1) - F(0.75) + \Pr(X = 0.75)$$

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.

	Reality	
Your Decision	*No Contamination*	*Contamination*
Contamination	Mistake: Type I Error (Probability = α)	Correct Decision (Probability = $1-\beta$)
No Contamination	Correct Decision	Mistake: Type II Error (Probability = β)

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-tetrachlorobenzen (TcCB) in the

Cleanup site significantly above background levels?" then you may) reformulate this question as "Is the average concentration of TcCB il at the Cleanup site greater than the average concentration of TcCB il at a Reference site?"

pothesis test or *significance test* is a formal mathematical mecha- objectively making a decision in the face of uncertainty, and is usu- l to answer a question about the value of a population parameter. A *d hypothesis test* about a population parameter θ (theta) is used to ıull hypothesis

$$H_0 : \theta = \theta_0 \qquad (7.1)$$

.e two-sided alternative hypothesis

$$H_a : \theta \neq \theta_0 \qquad (7.2)$$

) (pronounced "H-naught") denotes the null hypothesis that the true θ is equal to some specified value θ_0 (theta-naught). A *lower one- othesis test* is used to test the null hypothesis

$$H_0 : \theta \geq \theta_0 \qquad (7.3)$$

e lower one-sided alternative hypothesis

$$H_a : \theta < \theta_0 \qquad (7.4)$$

per one-sided hypothesis test is used to test the null hypothesis

$$H_0 : \theta \leq \theta_0 \qquad (7.5)$$

e upper one-sided alternative hypothesis

$$H_a : \theta > \theta_0 \qquad (7.6)$$

nmental monitoring, we are almost always concerned only with l hypotheses, such as "The average concentration of TcCB in the

soil at the Cleanup site is less than or equal to 2 ppb," or "The average value of pH at the compliance well is greater than or equal to 7." We rarely consider two-sided hypothesis tests.

Hypotheses tests are usually based on a *test statistic*, say T, which is computed from a random sample from the population. If T is "too extreme," then we decide to reject the null hypothesis in favor of the alternative hypothesis. For example, suppose we are interested in determining whether the true mean of a distribution is less than or equal to some hypothesized value μ_0 vs. the alternative that the true mean is bigger than μ_0. This is simply the one-sided upper hypothesis given in Equations (7.5) and (7.6) with θ replaced by μ. We can use Student's t-statistic, which we will discuss in more detail later in this chapter, to test this hypothesis. Student's t-statistic is a scaled version of the sample mean minus the hypothesized mean:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad (7.7)$$

Since the sample mean is an unbiased estimator of the true mean μ, if the true mean is equal to μ_0, then the sample mean is "bouncing around" μ_0 and the t-statistic is "bouncing around" 0. The distribution of the t-statistic under the null hypothesis is shown in Figure 5.9 in Chapter 5 for sample sizes of $n = 2$, 5, and ∞. On the other hand, if the true mean μ is larger than μ_0, then the sample mean is bouncing around μ, the numerator of the t-statistic is bouncing around $\mu - \mu_0$, and the t-statistic is bouncing around some positive number. So if the t-statistic is "large" we will probably reject the null hypothesis in favor of the alternative hypothesis.

Parametric vs. Nonparametric Tests

For a *parametric test*, the test statistic T is usually some estimator of θ (possibly shifted by subtracting a number and scaled by dividing by a number), 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 hypothesis 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 sample, and the distribution of this statistic under the null hypothesis does not depend on the distribution of the two populations.

›e I and Type II Errors (Significance Level and Power)

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

	Reality	
Your Decision	H_0 True	H_0 False
Reject H_0	Mistake: Type I Error (Probability = α)	Correct Decision (Probability = $1-\beta$)
›o Not Reject H_0	Correct Decision	Mistake: Type II Error (Probability = β)

Table 7.2 The framework of a hypothesis test

s we explained in Chapter 2, statisticians call the two kinds of mistakes ;an make a ***Type I error*** and a ***Type II error***. Of course, in the real l, once you make a decision, you take an action (e.g., clean up the site nothing), and you hope to find out eventually whether the decision you was the correct decision. For a specific hypothesis test, the probability ıking a Type I error is usually denoted with the Greek letter α (alpha) ; called the ***significance-level*** or ***α-level*** of the test. This probability is alled the ***false positive rate***. The probability of making a Type II error ıally denoted with the Greek letter β (beta). This probability is also the ***false negative rate***. The probability $1-\beta$ denotes the probability of ›tly deciding to reject the null hypothesis when in fact it is false. This ›ility is called the ***power*** of the hypothesis test.

›te that in Table 7.2 above the phrase "Do Not Reject H_0" is used in- ›f "Decide H_0 is True." This is because in the framework of hypothe- ting, you assume the null hypothesis is true unless you have enough ıce to reject it. If you end up not rejecting H_0 because of the value of est statistic, you may have unknowingly committed a Type II error; .e alternative hypothesis is really true, but you did not have enough .ce to reject H_0.

:ritical aspect of any hypothesis test is deciding on the acceptable val- : the probabilities of making a Type I and Type II error. This is part of of the DQO process. **The choice of values for α and β is a subjec- ›licy decision.** The possible choices for α and β are limited by the

sample size of the experiment (usually denoted n), the variability inherent in the data (usually denoted σ), and the magnitude of the difference between the null and alternative hypothesis (usually denoted δ or Δ). Conventional choices for α 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 α, β, n and the scaled difference δ/σ (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.

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 α (the Type I error level), you reject the null hypothesis, and if the p-value is greater than α, 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 heads 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 fair coin. If the coin keeps coming up heads, however, how many flips in a row coming up heads will you let go by before you decide to say the coin is the two-headed coin?

Table 7.3 displays the probability of seeing various numbers of heads in a row under the null hypothesis that your friend is flipping the fair coin. (Of course, under the alternative hypothesis, all flips will result in a head and the probability of seeing any number of heads in a row is 100%.) Suppose you decide you will make your decision after seeing the results of five flips, and if you see $T = 5$ heads in a row then you will reject the null hypothesis and say your friend is flipping the two-headed 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 rate is no greater than 1%, then you will have to wait until you see the ɔme 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%. u do see seven heads in a row, the p-value is 0.0078.

# Heads in a Row (T)	Probability (%)
1	50
2	25
3	12.5
4	6.25
5	3.12
6	1.56
7	0.78
8	0.39
9	0.20
10	0.10

7.3 *The probability of seeing T heads in a row in T flips of a fair coin*

ɪr this example, the power associated with your decision rule is the bility of correctly deciding your friend is flipping the two-headed coin in fact that is the one she is flipping. This is a special example in . the power is equal to 100%, because if your friend really is flipping o headed coin then you will always see a head on each flip and no mat- ıat value of T you choose for the cut-off, you will always see T heads ɪw. 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).

ionship between Hypothesis Tests and Confidence Intervals

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

ıenever 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 t that the *null* hypothesis is not true, even if there is a huge difference n the postulated value of θ (e.g., $\theta_0 \le 5$ ppb) and the true value of θ

(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 θ (e.g., $\theta_0 \le 5$ ppb) and the true value of θ (e.g., $\theta \approx 6$ ppb), but this difference may not really be important to detect. Confidence intervals help you sort out the important distinction between a ***statistically significantly difference*** and a ***scientifically meaningful difference***.

Test Type	Alternative Hypothesis	Corresponding Confidence Interval	Rejection Rule Based on CI
Two-sided	$\theta \ne \theta_0$	Two-sided [LCL, UCL]	$LCL > \theta_0$ or $UCL < \theta_0$
Lower	$\theta < \theta_0$	Upper [$-\infty$, UCL]	$UCL < \theta_0$
Upper	$\theta > \theta_0$	Lower [LCL, ∞]	$LCL > \theta_0$

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 Johnson and Wichern (1998).

One-Sample	Two-Samples	Multiple Samples
Goodness-of-Fit		
Proportion	Proportions	Proportions
Location	Locations	Locations
Variability	Variability	Variability

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 me from a specific distribution, such as "the data come from a normal ition with mean 10 and standard deviation 2," or to test the more gen- ll hypothesis that the data come from a particular family of distribu- uch as "the data come from a lognormal distribution." Goodness-of- are mostly used to test the latter kind of hypothesis, since in practice ly know or want to specify the parameters of the distribution.

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

for Normality

o commonly used tests to test the null hypothesis that the observations from a normal distribution are the Shapiro-Wilk test (Shapiro and .965), and the Shapiro-Francia test (Shapiro and Francia, 1972). The)-Wilk test is more powerful at detecting short-tailed (platykurtic) and l distributions, and less powerful against symmetric, moderately long- (leptokurtic) distributions. Conversely, the Shapiro-Francia test is owerful against symmetric long-tailed distributions and less powerful short-tailed distributions (Royston, 1992b; 1993). These tests are ered to be two of the very best tests of normality available ostino, 1986b, p. 406).

apiro-Wilk Test

Shapiro-Wilk test statistic can be written as:

$$W = \left[\sum_{i=1}^{n} a_i x_{(i)} \right]^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (7.8)$$

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

$$\underline{a}^T = \underline{m}^T V^{-1} \Big/ \sqrt{\underline{m}^T V^{-1} V^{-1} \underline{m}} \qquad (7.9)$$

where T denotes the transpose operator, and $\underline{m}$ 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 $\underline{a}$ are the expected values of the standard normal order statistics weighted by their variance-covariance matrix, and normalized so that $\underline{a}^T \underline{a} = 1$. 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 $\underline{a}$ and $\underline{x}_{()}$:

$$W = \left\{ r\left[\underline{a}, \underline{x}_{(\,)} \right] \right\}^2 \qquad (7.10)$$

where

$$r(\underline{x}, \underline{y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \qquad (7.11)$$

(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 $\underline{a}$ 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 ($\underline{m}$) and not on the variance-covariance matrix (V):

$$W' = \left[\sum_{i=1}^{n} b_i x_{(i)} \right]^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (7.12)$$

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

$$\underline{b} = \underline{m} \Big/ \sqrt{\underline{m}^T \underline{m}} \tag{7.13}$$

l authors, including Ryan and Joiner (1973), Filliben (1975), and ɜrg and Bingham (1975), note that the W'-statistic is intuitively ing because it is the squared sample correlation coefficient associated normal probability plot. That is, it is the squared correlation between lered sample values $\underline{x}_{()}$ and the expected normal order statistics $\underline{m}$:

$$W' = \left\{ r\left[\underline{b}, \underline{x}_{()}\right] \right\}^2 = \left\{ r\left[\underline{m}, \underline{x}_{()}\right] \right\}^2 \tag{7.14}$$

isberg and Bingham (1975) introduced an approximation of the Shaancia W'-statistic that is easier to compute. They suggested using cores (Blom, 1958, pp. 68–75; see Chapter 3) to approximate the eleɔf $\underline{m}$:

$$\tilde{W}' = \frac{\left[\sum_{i=1}^{n} c_i x_{(i)}\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \left\{ r\left[\underline{c}, \underline{x}_{()}\right] \right\}^2 \tag{7.15}$$

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

$$\underline{c} = \underline{\tilde{m}} \Big/ \sqrt{\underline{\tilde{m}}^T \underline{\tilde{m}}} \tag{7.16}$$

$$\tilde{m}_i = \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right) \tag{7.17}$$

and Φ denotes the standard normal cdf. That is, the values of the elements of $\underline{m}$ in Equation (7.13) are replaced with their estimates based on the usual 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. λ 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.

Assumed Distribution	**Shapiro-Wilk (W)**	**Shapiro-Francia (W')**
Normal	0.918 (p=0.003)	0.923 (p=0.006)
Lognormal	0.979 (p=0.55)	0.987 (p=0.78)

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-