Chapter 5

Frequentist Statistical Inference

5.1. BACKGROUND

Statistical inference refers broadly to the process of drawing conclusions from a limited data sample about the characteristics of a (possibly hypothetical) "population," or generating process, from which the data were drawn. Put another way, inferential methods are meant to extract information from data samples about the process or processes that generated them.

The most familiar instance of statistical inference is in the formal testing of statistical hypotheses, also known as *significance testing*. In their simplest form, these tests yield a binary decision that a particular hypothesis about the phenomenon generating the data may be true or not, so that this process is also known as *hypothesis testing*. However, limiting statistical inferences to such binary conclusions is unnecessarily restrictive and potentially misleading (e.g., Nicholls, 2001). It is usually better to consider and communicate elements of the inferential procedures beyond just a binary result in order to address degrees of confidence in the inferences. The most familiar of these procedures are based in the frequentist, or relative frequency, view of probability, and are usually covered extensively in introductory courses in statistics. Accordingly, this chapter will review only the basic concepts behind these most familiar formal hypothesis tests, and subsequently emphasize aspects of inference that are particularly relevant to applications in the atmospheric sciences. A different approach to characterizing confidence about statistical inferences is provided by Bayesian statistics, based on the subjective view of probability, an introduction to which is provided in Chapter 6.

5.1.1. Parametric Versus Nonparametric Inference

There are two contexts in which frequentist statistical inferences are addressed; broadly, there are two types of tests and inferences. *Parametric tests* and inferences are those conducted in situations where we know or assume that a particular parametric distribution is an appropriate representation for the data and/or the test statistic. *Nonparametric tests* are conducted without assumptions that particular parametric forms are appropriate in a given situation.

Very often, parametric tests consist essentially of making inferences about particular distribution parameters. Chapter 4 presented a selection of parametric distributions that have been found to be useful for describing atmospheric data. Fitting such a distribution amounts to distilling the information contained in a sample of data, so that the distribution parameters can be regarded as representing (at least some aspects of) the nature of the underlying data-generating process of interest. Thus a parametric statistical test concerning a physical process of interest can reduce to a test pertaining to a distribution parameter, such as a Gaussian mean μ.

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00005-5

Nonparametric, or *distribution-free tests* and inferences proceed without the necessity of assumptions about what, if any, parametric distribution can well describe the data at hand. Nonparametric inferential methods proceed along one of two basic lines. One approach is to construct the procedure similarly to parametric procedures, but in such a way that the distribution of the data is unimportant, so that data from any distribution can be treated in the same way. In the following, procedures following this approach are referred to as *classical* nonparametric methods, since they were devised before the advent of cheap and abundant computing power. In the second approach, crucial aspects of the relevant distribution are inferred directly from the data, by repeated computer manipulations of the data themselves. These nonparametric methods are known broadly as *resampling* procedures.

5.1.2. The Sampling Distribution

The concept of the *sampling distribution* is fundamental to both parametric and nonparametric inferential methods. Recall that a statistic is some numerical quantity computed from a batch of data. The sampling distribution for a statistic is the probability distribution describing batch-to-batch variations of that statistic. Since the batch of data from which any sample statistic (including the test statistic for a hypothesis test) has been computed is subject to sampling variations, sample statistics are subject to sampling variations as well. The value of a statistic computed from a particular batch of data will in general be different from that for the same statistic computed using a different batch of the same kind of data. For example, average January temperature is obtained by averaging daily temperatures during that month at a particular location for a given year. This statistic is different from year to year.

The random batch-to-batch variations of sample statistics can be described using probability distributions just as the random variations of the underlying data can be described using probability distributions. Thus sample statistics can be viewed as having been drawn from probability distributions, and these distributions are called sampling distributions. The sampling distribution provides a probability model describing the relative frequencies of possible values of the statistic.

5.1.3. The Elements of Any Hypothesis Test

Any hypothesis test proceeds according to the following five steps:

1. Identify a *test statistic* that is appropriate to the data and question at hand. The test statistic is the quantity computed from the data values that will be the subject of the test. In parametric settings the test statistic will often be the sample estimate of a parameter of a relevant distribution. In nonparametric resampling tests there is nearly unlimited freedom in the definition of the test statistic.
2. Define a *null hypothesis*, usually denoted H_0. The null hypothesis defines a specific logical frame of reference against which to judge the observed test statistic. Often the null hypothesis will be a "straw man" that we hope to reject.
3. Define an *alternative hypothesis*, H_A. Many times the alternative hypothesis will be as simple as "H_0 is not true," although more complex and specific alternative hypotheses are also possible.
4. Obtain the *null distribution*, which is simply the sampling distribution for the test statistic, if the null hypothesis is true. Depending on the situation, the null distribution may be an exactly known parametric distribution, a distribution that is well approximated by a known parametric distribution, or an empirical distribution obtained by resampling the data. Identifying the null distribution is the crucial step in the construction of a hypothesis test.

5. Compare the observed test statistic to the null distribution. If the test statistic falls in a sufficiently improbable region of the null distribution, H_0 is rejected as too implausible to have been true given the observed evidence. If the test statistic falls within the range of ordinary values described by the null distribution, the test statistic is seen as consistent with H_0, which is then not rejected. Note that not rejecting H_0 does not necessarily mean that the null hypothesis is true, only that there is insufficient evidence to reject this hypothesis. Not rejecting H_0 does not mean that it is "accepted." Rather, when H_0 is not rejected, it is more precise to say that it is "not inconsistent" with the observed data.

5.1.4. Test Levels and *p* Values

The sufficiently improbable region of the null distribution just referred to is defined by the *rejection level*, or simply the *level*, of the test. The null hypothesis is rejected if the probability (according to the null distribution) of the observed test statistic, *and all other results at least as unfavorable to the null hypothesis*, is less than or equal to the test level. The test level is chosen in advance of the computations, but it depends on the particular investigator's judgment and taste, so that there is usually a degree of arbitrariness about its specific value. Table 5.1 lists qualitative characterizations of the strength of evidence provided by various test levels (i.e., ceilings on p values), as formulated by Fisher (Efron et al., 2001). Commonly the 5% level is chosen, although tests conducted at the 10% level or the 1% level are not unusual. In situations where penalties can be associated quantitatively with particular test errors (e.g., erroneously rejecting H_0), however, the test level can be optimized (see Winkler, 1972b).

The p value is the specific probability that the observed value of the test statistic, together with all other possible values of the test statistic that are at least as unfavorable to the null hypothesis, will occur (according to the null distribution). Thus the null hypothesis is rejected if the p value is less than or equal to the test level, and is not rejected otherwise.

Unfortunately, p values are commonly misused and misinterpreted, which has led the American Statistical Association (ASA) to formulate an official statement on their use and purposes (Wasserstein and Lazar, 2016). Among the six principles listed in this statement is the important caution that a p value is *not* the probability that H_0 is true (see also Ambaum, 2010). The ASA statement also notes that it is more informative to report the p value for a hypothesis test rather than simply a reject/not-reject decision at a particular test level, because the p value also communicates the confidence with which a null hypothesis has or has not been rejected; and that a p value does not measure the size of an effect or the importance of a result.

Another important principle from the ASA statement is that conducting multiple analyses and reporting only nominally significant results (colloquially known as "p-hacking") renders the values so derived to be essentially uninterpretable. Relatedly, it is necessary for the hypotheses being examined

TABLE 5.1 Fisher's Scale of Evidence Against a Null Hypothesis, Mapped to Test Levels (i.e., Ceilings on p Values)

Test level	0.20	0.10	0.05	0.025	0.01	0.005	0.001
Strength of evidence	Null	Borderline	Moderate	Substantial	Strong	Very strong	Overwhelming

From Efron et al. (2001).

to have been formulated without having seen the specific data that will be used to evaluate them. This separation may be problematic especially in settings like climate research, in which new data accumulate slowly. A somewhat fanciful counterexample illustrating this point has been provided by Von Storch (1995) and Von Storch and Zwiers (1999), Chapter 6).

5.1.5. Error Types and the Power of a Test

Another way of looking at the level of a test is as the probability of falsely rejecting the null hypothesis, given that it is true. This false rejection is called a *Type I error*, and its probability (the level of the test) is often denoted α. Type I errors are defined in contrast to *Type II errors*, which occur if H_0 is not rejected when it is in fact valid. The probability of a Type II error usually is denoted β.

Figure 5.1 illustrates the relationship of Type I and Type II errors for a test conducted at the 5% level. A test statistic falling to the right of a critical value, corresponding to the quantile in the null distribution yielding the test level as a tail probability, results in rejection of the null hypothesis. Since the area under the probability density function of the null distribution to the right of the critical value in Figure 5.1 (horizontal hatching) is 0.05, this is the probability of a Type I error. The portion of the horizontal axis corresponding to H_0 rejection is sometimes called the *rejection region* or the *critical region*. Outcomes in this range are not impossible under H_0, but rather have some small probability α of occurring. It is clear from this illustration that, although we would like to minimize the probabilities of both Type I and Type II errors, this is not possible. Their probabilities, α and β, can be adjusted by adjusting the level of the test, which corresponds to moving the critical value to the left or right; but decreasing α in this way necessarily increases β, and vice versa.

The level of the test, α, can be prescribed, but the probability of a Type II error, β, usually cannot. This is because alternative hypotheses are defined more generally than the null hypothesis, and typically consists of the union of many specific alternative hypotheses. The probability α depends on the null distribution, which must be known in order to conduct a test, but β depends on which specific alternative

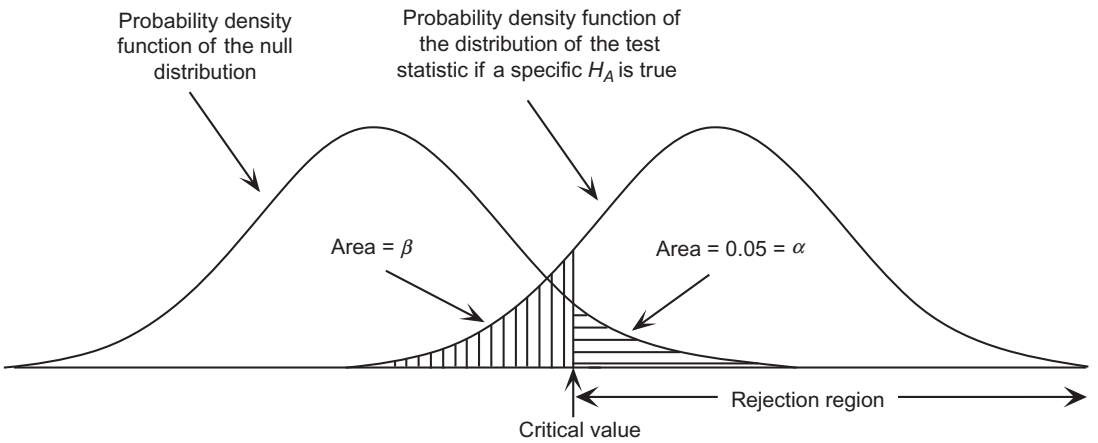

FIGURE 5.1 Illustration of the relationship of the rejection level, α, corresponding to the probability of a Type I error (horizontal hatching); and the probability of a Type II error, β (vertical hatching); for a test conducted at the 5% level. The horizontal axis represents possible values of the test statistic. Decreasing the probability of a Type I error necessarily increases the probability of a Type II error, and vice versa.

hypothesis would be applicable, and this is generally not known. Figure 5.1 illustrates the relationship between α and β for only one of a potentially infinite number of possible alternative hypotheses.

It is sometimes useful, however, to examine the behavior of β over a range of the possibilities for H_A. This investigation usually is done in terms of the quantity $1 - \beta$, which is known as the *power* of a test against a specific alternative. Geometrically, the power of the test illustrated in Figure 5.1 is the area under the sampling distribution on the right (i.e., for a particular H_A) that does not have vertical hatching. The relationship between the power of a test and a continuum of specific alternative hypotheses is called the *power function*. The power function expresses the probability of rejecting the null hypothesis, as a function of how far wrong it is. One reason why we might like to choose a less stringent test level (say, $\alpha = 0.10$) would be to better balance error probabilities for a test known to have low power.

5.1.6. One-Sided Versus Two-Sided Tests

A statistical test can be either *one-sided* or *two-sided*. This dichotomy is sometimes expressed in terms of tests being either *one-tailed* or *two-tailed*, since it is the probability in the extremes (tails) of the null distribution that governs whether a test result is interpreted as being significant. Whether a test is one-sided or two-sided depends on the nature of the hypothesis being tested.

A one-sided test is appropriate if there is a prior (e.g., a physically based) reason to expect that violations of the null hypothesis will lead to values of the test statistic on a particular side of the null distribution. This situation is illustrated in Figure 5.1, which has been drawn to imply that alternative hypotheses producing smaller values of the test statistic have been ruled out on the basis of prior information. In such cases the alternative hypothesis would be stated in terms of the true value being larger than the null hypothesis value (e.g., H_A: $\mu > \mu_0$), rather than the more vague alternative hypothesis that the true value is not equal to the null value (H_A: $\mu \neq \mu_0$). In the one-sided test situation illustrated in Figure 5.1, any test statistic larger than the $100 \cdot (1 - \alpha)$ percentile of the null distribution results in the rejection of H_0 at the α level, whereas very small values of the test statistic would not lead to a rejection of H_0.

A one-sided test is also appropriate when only values on one tail or the other of the null distribution are unfavorable to H_0 because of the way the test statistic has been constructed. For example, a test statistic involving a squared difference will be near zero if the difference is small, but will take on large positive values if the difference is large. In this case, results on the left tail of the null distribution could be quite supportive of H_0, so that only right-tail probabilities would result in H_0 rejection.

Two-sided tests are appropriate when either very large or very small values of the test statistic are unfavorable to the null hypothesis. Usually such tests pertain to the very general alternative hypothesis "H_0 is not true." The rejection region for two-sided tests consists of both the extreme left and extreme right tails of the null distribution. These two portions of the rejection region are delineated in such a way that the sum of their two probabilities under the null distribution yields the level of the test, α. That is, the null hypothesis is rejected at the α level if the test statistic is larger than $100 \cdot (1 - \alpha/2)$ percentile of the null distribution on the right tail, or is smaller than the $100 \cdot (\alpha/2)$ percentile of this distribution on the left tail. Thus a test statistic must be further out on the tail (i.e., more unusual with respect to H_0) to be declared significant in a two-tailed test as compared to a one-tailed test, at a specified test level. That the test statistic must be more extreme to reject the null hypothesis in a two-tailed test is appropriate, because generally one-tailed tests are used when additional (i.e., external to the test data) information exists, which then allows stronger inferences to be made.

5.1.7. Confidence Intervals: Inverting Hypothesis Tests

Hypothesis testing ideas can be used to construct *confidence intervals* around sample statistics. A typical use of confidence intervals is to construct *error bars* around plotted sample statistics in a graphical display, but more generally they allow a fuller appreciation of possible magnitudes of effects being investigated.

In essence, a confidence interval is derived from a hypothesis test in which the value of an observed sample statistic plays the role of the population parameter value under a hypothetical null hypothesis. The confidence interval around this sample statistic then consists of other possible values of the statistic for which that hypothetical H_0 would not be rejected. Whereas hypothesis tests evaluate probabilities associated with an observed test statistic in the context of a null distribution, conversely confidence intervals are constructed by finding the values of the test statistic that would not fall into the rejection region. In this sense, confidence interval construction is the inverse operation to hypothesis testing. That is, there is a duality between a one-sample hypothesis test and the computed confidence interval around the observed statistic, such that the $100\cdot(1-\alpha)\%$ confidence interval around an observed statistic will not contain the null-hypothesis value of the test if the test is significant at the α level, and will contain the null value if the test is not significant at the α level. Expressing the results of a hypothesis test in terms of the corresponding confidence interval will typically be more informative than simply reporting a reject/not-reject decision, because the width of the confidence interval and the distance of its endpoint from the null-hypothesis value will also communicate information about the degree of uncertainty in the sample estimate and about the strength of the inference.

It is tempting to think of a $100\cdot(1-\alpha)\%$ confidence interval as being wide enough to contain the true value with probability $1-\alpha$, but this interpretation is not correct. The reason is that, in the frequentist view, a population parameter is a fixed if unknown constant. Therefore once a confidence interval has been constructed, the true value is either inside the interval or not. The correct interpretation is that $100\cdot(1-\alpha)\%$ of a large number of hypothetical similar confidence intervals, each computed on the basis of a different batch of data of the same kind (and therefore each being somewhat different from each other), will contain the true value.

Example 5.1. A Hypothesis Test Involving the Binomial Distribution

The hypothesis testing framework can be illustrated with a simple, although artificial, example. Suppose that advertisements for a tourist resort in the sunny desert southwest claim that, on average, 6 days out of 7 are cloudless during winter. To examine this claim, we would need to observe the sky conditions in the area on a number of winter days, and then compare the fraction observed to be cloudless with the claimed proportion of $6/7 = 0.857$. Assume that we could arrange to take observations on 25 independent occasions. (These would not be consecutive days, because of the serial correlation of daily weather values.) If cloudless skies are observed on 15 of those 25 days, is this observation consistent with, or does it justify questioning, the claim?

This problem fits neatly into the parametric setting of the binomial distribution. A given day is either cloudless or it is not, and observations have been taken sufficiently far apart in time that they can be considered to be independent. By confining observations to only a relatively small portion of the year, we can expect that the probability, p, of a cloudless day is approximately constant from observation to observation.

The first of the five hypothesis testing steps has already been completed, since the test statistic of $X = 15$ out of $N = 25$ days has been dictated by the form of the problem. The null hypothesis is that

the resort advertisement was correct in claiming $p = 0.857$. Understanding the nature of advertising, it is reasonable to anticipate that, should the claim be false, the true probability will be lower. Thus the alternative hypothesis is that $p < 0.857$. That is, the test will be one-tailed, since results indicating $p > 0.857$ are not of interest with respect to possibly rejecting the truth of the claim. Our prior information regarding the nature of advertising claims will allow stronger inference than would have been the case if we were to have regarded alternatives with $p > 0.857$ as plausible.

Now the crux of the problem is to find the null distribution. That is, what is the sampling distribution of the test statistic X if the true probability of cloudless conditions is 0.857? This X can be thought of as the sum of 25 independent 0's and 1's, with the 1's having some constant probability of occurring on each of the 25 occasions. These are the conditions for the binomial distribution. Thus for this test the null distribution is binomial, with parameters $p = 0.857$ and $N = 25$.

It remains to compute the probability that 15 or fewer cloudless days would have been observed on 25 independent occasions if the true probability p is in fact 0.857. This probability is the p value for the test, which is a different usage for this symbol than the binomial distribution parameter, p. The direct, but tedious, approach to this computation is summation of the terms given by

$$\Pr\{X \le 15\} = \sum_{x=0}^{15} \binom{25}{x} 0.857^x (1 - 0.857)^{25-x}. \tag{5.1}$$

Here the terms for the outcomes for $X < 15$ must be included in addition to $\Pr\{X = 15\}$, since observing, say, only 10 cloudless days out of 25 would be even more unfavorable to H_0 than is $X = 15$. The p value for this test as computed from Equation 5.1 is only 0.0015. Thus $X \le 15$ is a highly improbable result if the true probability of a cloudless day is 6/7, and this null hypothesis would be resoundingly rejected. According to this test, the observed data provide very convincing evidence that the true probability is smaller than 6/7.

A much easier approach to the p-value computation in this example is to use the *Gaussian approximation to the binomial distribution*. This approximation follows from the Central Limit Theorem since, as the sum of some number of 0's and 1's, the random variable X will follow approximately the Gaussian distribution if N is sufficiently large. Here sufficiently large means roughly that $0 < p \pm 3[p(1-p)/N]^{1/2} < 1$, in which case the binomial X can be characterized to good approximation using a Gaussian distribution with

$$\mu \approx Np \tag{5.2a}$$

and

$$\sigma \approx \sqrt{Np(1-p)}. \tag{5.2b}$$

In the current example these parameters are $\mu \approx (25)(0.857) = 21.4$ and $\sigma \approx [(25)(0.857)(1 - 0.857)]^{1/2} = 1.75$. However, $p + 3[p(1-p)/N]^{1/2} = 1.07$, which suggests that use of the Gaussian approximation is questionable in this example. Figure 5.2 compares the exact binomial null distribution with its Gaussian approximation. The correspondence is close, although the Gaussian approximation ascribes nonnegligible probability to the impossible outcomes $\{X > 25\}$, and correspondingly too little probability is assigned to the left tail. Nevertheless, the Gaussian approximation will be carried forward here to illustrate its use.

One small technical issue that must be faced here relates to the representation of discrete probabilities using a continuous probability density function. The p value for the exact binomial test is given by the discrete sum in Equation 5.1 yielding $\Pr\{X \le 15\}$, but its Gaussian approximation is given by the integral

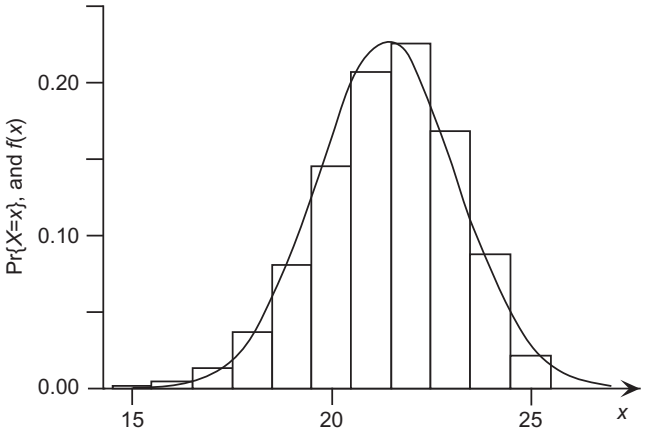

FIGURE 5.2 Relationship of the binomial null distribution (histogram bars) for Example 5.1, and its Gaussian approximation (*smooth curve*). The observed $X = 15$ falls on the far left tail of the null distribution. The exact p value from Equation 5.1 is $\Pr\{X \le 15\} = 0.0015$. Its approximation using the Gaussian distribution, including the continuity correction, is $\Pr\{X \le 15.5\} = \Pr\{Z \le -3.37\} = 0.00038$.

of the Gaussian PDF over the corresponding portion of the real line. This integral should include values >15 but closer to 15 than 16, since these also approximate the discrete $X = 15$. Thus the relevant Gaussian probability will be $\Pr\{X \le 15.5\} = \Pr\{Z \le (15.5 - 21.4)/1.75\} = \Pr\{Z \le -3.37\} = 0.00038$, again leading to rejection but with too much confidence (too small a p value) because the Gaussian approximation puts insufficient probability on the left tail. The additional increment of 0.5 between the discrete $X = 15$ and the continuous $X = 15.5$ is called a *continuity correction*.

The Gaussian approximation to the binomial, Equations 5.2a and 5.2b, can also be used to construct a confidence interval (error bars) around the observed estimate of the binomial $\hat{p} = 15/25 = 0.6$. To do this, imagine a test whose null hypothesis is that the true binomial probability for this situation is 0.6. This test is then solved in an inverse sense to find the values of the test statistic defining the boundaries of the rejection region. That is, how large or small a value of x/N would be tolerated before this new null hypothesis would be rejected?

If a 95% confidence region is desired, the test to be inverted will be at the 5% level. Since the true binomial p could be either larger or smaller than the observed x/N, a two-tailed test (rejection regions for both very large and very small x/N) is appropriate. Referring to Table B.1, since this null distribution is approximately Gaussian, the standardized Gaussian variable cutting off probability equal to $0.05/2 = 0.025$ at the upper and lower tails is $z = \pm 1.96$. (This is the basis of the useful rule of thumb that a 95% confidence interval consists approximately of the mean value ± 2 standard deviations.) Using Equation 5.2a, the mean number of cloudless days should be $(25)(0.6) = 15$, and from Equation 5.2b the corresponding standard deviation is $[(25)(0.6)(1 - 0.6)]^{1/2} = 2.45$. Using Equation 4.28 with $z = \pm 1.96$ yields $x = 10.2$ and $x = 19.8$, leading to the 95% confidence interval bounded by $p = x/N = 0.408$ and 0.792. Notice that the claimed binomial p of $6/7 = 0.857$ falls outside this interval. The confidence interval computed exactly from the binomial probabilities is [0.40, 0.76], with which the Gaussian approximation agrees very nicely. For the Gaussian approximation used to construct this confidence interval, $p \pm 3[p(1 - p)/N]^{1/2}$ ranges from 0.306 to 0.894, which is comfortably within the range [0, 1].

Finally, what is the power of this test? That is, we might like to calculate the probability of rejecting the null hypothesis as a function of the true but unknown binomial p. As illustrated in Figure 5.1 the answer to this

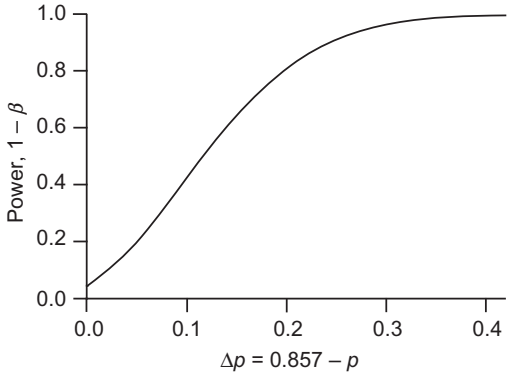

FIGURE 5.3 Power function for the test in Example 5.1. The vertical axis shows the probability of rejecting the null hypothesis, as a function of the difference between the true (and unknown) binomial p and the binomial p for the null distribution (=0.857).

question will depend on the level of the test, since it is more likely (with probability $1-\beta$) to correctly reject a false null hypothesis if α is relatively large. Assuming a test at the 5% level, and again assuming the Gaussian approximation to the binomial distribution for simplicity, the (one-sided) critical value will correspond to $z=-1.645$ relative to the null distribution; or $-1.645=(Np-21.4)/1.75$, yielding $Np=18.5$. The power of the test for a given alternative hypothesis is the probability observing the test statistic $X=\{$number of cloudless days out of $N\}$ less than or equal to 18.5, given the true binomial p corresponding to that alternative hypothesis, and will equal the area to the left of 18.5 in the approximate Gaussian sampling distribution for X defined by that binomial p and $N=25$. Collectively, these probabilities for a range of alternative hypotheses constitute the power function for the test.

Figure 5.3 shows the resulting power function. Here the horizontal axis indicates the difference between the true binomial p and that assumed by the null hypothesis (=0.857). For $\Delta p=0$ the null hypothesis is true, and Figure 5.3 indicates a 5% chance of rejecting it, which is consistent with the test being conducted at the 5% level. We do not know the true value of p, but Figure 5.3 shows that the probability of rejecting the null hypothesis increases as the true p is increasingly different from 0.857, until we are virtually assured of rejecting H_0 with a sample size of $N=25$ if the true probability is smaller than about 0.5. If $N>25$ days had been observed, the resulting power curve would be above that shown in Figure 5.3, so that probabilities of rejecting false null hypotheses would be greater (i.e., their power functions would climb more quickly toward 1), indicating more sensitive tests. Conversely, corresponding tests involving fewer samples would be less sensitive, and their power curves would lie below the one shown in Figure 5.3. ◇

5.2. SOME COMMONLY ENCOUNTERED PARAMETRIC TESTS

5.2.1. One-Sample *t*-Test

By far, the most commonly encountered parametric tests in classical statistics relate to the Gaussian distribution. Tests based on the Gaussian are so pervasive because of the strength of the Central Limit Theorem. As a consequence of this theorem, many non-Gaussian problems can be treated at least approximately in the Gaussian framework. The example test for the binomial parameter p in Example 5.1 is one such case.

Probably the most familiar statistical test is the *one-sample t-test*, which examines the null hypothesis that an observed sample mean has been drawn from a population or generating process centered at some previously specified mean, μ_0. If the number of data values making up the sample mean is large enough for its sampling distribution to be essentially Gaussian (by the Central Limit Theorem), then the test statistic

$$t = \frac{\overline{x} - \mu_0}{[\hat{\mathrm{Var}}(\overline{x})]^{1/2}} \tag{5.3}$$

follows a distribution known as *Student's t*, or simply the *t distribution*. Equation 5.3 resembles the standard Gaussian variable z (Equation 4.26), except that a sample estimate of the variance of the sample mean (denoted by the "hat" accent) has been substituted in the denominator.

The t distribution is a symmetrical distribution that is very similar to the standard Gaussian distribution, although with more probability assigned to the tails. That is, the t distribution has heavier tails than the Gaussian distribution. The t distribution is controlled by a single parameter, ν, called the *degrees of freedom*. The parameter ν can take on any positive integer value, with the largest differences from the Gaussian being produced for small values of ν. For the test statistic in Equation 5.3, $\nu = n - 1$, where n is the number of independent observations being averaged in the sample mean in the numerator.

Tables of t distribution probabilities are available in almost any introductory statistics textbook. However, for even moderately large values of n (and therefore of ν) the variance estimate in the denominator becomes sufficiently precise that the t distribution is closely approximated by the standard Gaussian distribution. The differences in tail quantiles are about 4% and 1% for $\nu = 30$ and 100, respectively, so for sample sizes of this magnitude and larger it is usually quite acceptable to evaluate probabilities associated with the test statistic in Equation 5.3 using standard Gaussian probabilities.

Use of the standard Gaussian PDF (Equation 4.25) as the null distribution for the test statistic in Equation 5.3 can be understood in terms of the Central Limit Theorem, which implies that the sampling distribution of the sample mean in the numerator will be approximately Gaussian if n is sufficiently large. Subtracting the mean μ_0 in the numerator will center that Gaussian distribution on zero if the null hypothesis, to which μ_0 pertains, is true. If n is also large enough that the standard deviation of the sampling distribution of the sample mean (the denominator) can be estimated sufficiently precisely, then the sampling distribution of the quantity in Equation 5.3 will also have unit standard deviation to good approximation. A Gaussian distribution with zero mean and unit standard deviation is the standard Gaussian distribution.

The variance of the sampling distribution of a mean of n independent observations, in the denominator of Equation 5.3, is estimated according to

$$\hat{\mathrm{Var}}[\overline{x}] = s^2/n, \tag{5.4}$$

where s^2 is the sample variance (the square of Equation 3.6) of the individual x's being averaged. Equation 5.4 is clearly true for the simple case of $n = 1$, but also makes intuitive sense for larger values of n. We expect that averaging together, say, pairs ($n = 2$) of x's will give quite irregular results from pair to pair. That is, the sampling distribution of the average of two numbers will have a high variance. On the other hand, averaging together batches of $n = 1000$ x's will give very consistent results from batch to batch, because the occasional very large x will tend to be balanced by the occasional very small x: a sample of $n = 1000$ will tend to have nearly equally many very large and very small values. The variance of the sampling distribution (i.e., the batch-to-batch variability) of the average of 1000 independent numbers will thus be small.

For small (absolute) values of t in Equation 5.3, the difference in the numerator is small in comparison to the standard deviation of the sampling distribution of the difference, suggesting a quite ordinary sampling fluctuation for the sample mean, which should not trigger rejection of H_0. If the difference in the numerator is more than about twice as large as the denominator in absolute value, the null hypothesis would usually be rejected, corresponding to a two-sided test at the 5% level (cf. Table B.1).

5.2.2. Tests for Differences of Mean Under Independence

Another common statistical test involves the difference between two independent sample means. Plausible atmospheric examples of this situation might be differences of average winter 500 mb heights when one or the other of two synoptic regimes had prevailed, or perhaps differences in average July temperature at a location as represented in a climate model under a doubling vs. no doubling of atmospheric carbon dioxide concentration.

In general, two sample means calculated from different batches of data, even if they are drawn from the same population or generating process, will be different. The usual test statistic in this situation is a function of the difference of the two sample means being compared, and the actual observed difference will almost always be some number other than zero. The null hypothesis is usually that the true difference is zero. The alternative hypothesis is either that the true difference is not zero (the case where no a priori information is available as to which underlying mean should be larger, leading to a two-tailed test), or that one of the two underlying means is larger than the other (leading to a one-tailed test). The problem is to find the sampling distribution of the difference of the two sample means, given the null hypothesis assumption about the difference between their population counterparts. It is in this context that the observed difference of means can be evaluated for unusualness.

Nearly always—and sometimes quite uncritically—the assumption is tacitly made that the sampling distributions of the two sample means being differenced are Gaussian. This assumption will be valid either if the data composing each of the sample means are Gaussian, or if the sample sizes are sufficiently large that the Central Limit Theorem can be invoked. If both of the two sample means have Gaussian sampling distributions their difference will be Gaussian as well, since any linear combination of Gaussian variables will itself follow a Gaussian distribution. Under these conditions the test statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - E[\bar{x}_1 - \bar{x}_2]}{(\hat{\mathrm{Var}}[\bar{x}_1 - \bar{x}_2])^{1/2}} \tag{5.5}$$

will be distributed as standard Gaussian (Equation 4.25) for large samples. Note that this equation has a form similar to both Equations 5.3 and 4.27.

If the null hypothesis is equality of means of the two populations from which values of x_1 and x_2 have been drawn, then

$$E[\bar{x}_1 - \bar{x}_2] = E[\bar{x}_1] - E[\bar{x}_2] = \mu_1 - \mu_2 = 0. \tag{5.6}$$

Thus a specific hypothesis about the magnitude of the two equal means is not required. If some other null hypothesis is appropriate to a problem at hand, that difference of underlying means would be substituted in the numerator of Equation 5.5.

The variance of a difference (or sum) of two independent random quantities is the sum of the variances of those quantities. Intuitively this makes sense since contributions to the variability of the

difference are made by the variability of each the two quantities being differenced. With reference to the denominator of Equation 5.5,

$$\hat{\mathrm{Var}}[\bar{x}_1 - \bar{x}_2] = \hat{\mathrm{Var}}[\bar{x}_1] + \hat{\mathrm{Var}}[\bar{x}_2] = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}, \tag{5.7}$$

where the last equality is achieved using Equation 5.4. Thus if the batches making up the two averages are independent, and the sample sizes are sufficiently large, Equation 5.5 can be transformed to good approximation to the standard Gaussian z by rewriting the test statistic as

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\left[s_1^2/n_1 + s_2^2/n_2\right]^{1/2}}, \tag{5.8}$$

when the null hypothesis is that the two underlying means μ_1 and μ_2 are equal. This expression for the test statistic is appropriate when the variances of the two distributions from which the x_1's and x_2's are drawn are not equal. For relatively small sample sizes its sampling distribution is (approximately, although not exactly) the t distribution, with $\nu = \min(n_1, n_2) - 1$. For moderately large samples the sampling distribution is close to the standard Gaussian, for the same reasons presented in relation to its one-sample counterpart, Equation 5.3.

When it can be assumed that the variances of the distributions from which the x_1's and x_2's have been drawn are equal, that information can be used to calculate a single, *"pooled,"* estimate for that variance. Under this assumption of equal population variances, Equation 5.5 becomes instead

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left\{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right\}\right]^{1/2}}. \tag{5.9}$$

The quantity in curly brackets in the denominator is the pooled estimate of the population variance for the data values, which is just a weighted average of the two sample variances, and has in concept been substituted for both s_1^2 and s_2^2 in Equations 5.7 and 5.8. The sampling distribution for Equation 5.9 is the t distribution with $\nu = n_1 + n_2 - 2$. However, when both n_1 and n_2 are moderately large it is again usually quite acceptable to evaluate probabilities associated with the test statistic in Equation 5.9 using the standard Gaussian distribution.

For small (absolute) values of z in either Equations 5.8 or 5.9, the difference of sample means in the numerator is small in comparison to the standard deviation of the sampling distribution of their difference in the denominator, indicating a quite ordinary value in terms of the null distribution. As before, if the difference in the numerator is more than about twice as large as the denominator in absolute value, and the sample size is moderate or large, the null hypothesis would be rejected at the 5% level for a two-sided test.

As is also the case for one-sample tests, Equation 5.8 or 5.9 for two-sample t-tests can be worked backwards to yield a confidence interval around an observed difference of the sample means, $\bar{x}_1 - \bar{x}_2$. A rejection of H_0:$\{\mu_1 = \mu_2\}$ at the α level would correspond to the $100\cdot(1-\alpha)$% confidence interval for this difference to not include zero. However, counterintuitively, the individual $100\cdot(1-\alpha)$% confidence intervals for $\bar{x}_1$ and $\bar{x}_2$ could very well overlap in that case (Lanzante, 2005; Schenker and Gentleman, 2001). That is, overlapping $100\cdot(1-\alpha)$% confidence intervals for two individual sample statistics can very easily be consistent with the two statistics being significantly different according to an appropriate two-sample α-level test. The discrepancy between the results of this so-called *overlap*

method and a correct two-sample test is greatest when the two sample variances s_1^2 and s_2^2 are equal or nearly so, and progressively diminishes as the magnitudes of the two sample variances diverge. Conversely, nonoverlapping $(1-\alpha)\cdot 100\%$ confidence intervals does imply a significant difference at the α-level, at least.

5.2.3. Tests for Differences of Mean for Paired Samples

Equation 5.7 is appropriate when the x_1's and x_2's are observed independently. An important form of nonindependence occurs when the data values making up the two averages are *paired*, or observed simultaneously. In this case, necessarily, $n_1 = n_2$. For example, the daily temperature data in Table A.1 of Appendix A are of this type, since there is an observation of each variable at both locations on each day. When paired data of this kind are used in a two-sample t-test, the two averages being differenced are generally correlated. When this correlation is positive, as will often be the case, Equation 5.7 or the denominators of Equations 5.8 or 5.9 will overestimate the variance of the sampling distribution of the difference in the numerators. The result is that the test statistic will be too small (in absolute value), on average, so that the calculated p values will be too large, and null hypotheses that should be rejected will not be.

We should expect the sampling distribution of the difference in the numerator of the test statistic to be affected if pairs of x's going into the averages are strongly correlated. For example, the appropriate panel in Figure 3.31 indicates that the daily maximum temperatures at Ithaca and Canandaigua are strongly correlated, so that a relatively warm average monthly maximum temperature at one location would likely be associated with a relatively warm average at the other. A portion of the variability of the monthly averages is thus common to both, and that portion cancels in the difference in the numerator of the test statistic. That cancellation must also be accounted for in the denominator if the sampling distribution of the test statistic is to be approximately standard Gaussian.

The easiest and most straightforward approach to dealing with the t-test for paired data is to analyze differences between corresponding members of the $n_1 = n_2 = n$ pairs, which transforms the problem to the one-sample setting. That is, consider the sample statistic

$$\Delta = x_1 - x_2, \tag{5.10a}$$

with sample mean

$$\overline{\Delta} = \frac{1}{n}\sum_{i=1}^{n} \Delta_i = \overline{x}_1 - \overline{x}_2. \tag{5.10b}$$

The corresponding population mean will be $\mu_\Delta = \mu_1 - \mu_2$, which is often zero under H_0. The resulting test statistic is then of the same form as Equation 5.3,

$$z = \frac{\overline{\Delta} - \mu_\Delta}{\left(s_\Delta^2/n\right)^{1/2}}, \tag{5.11}$$

where s_Δ^2 is the sample variance of the n differences in Equation 5.10a. Joint variation in the pairs making up the difference $\Delta = x_1 - x_2$ is also automatically accounted for in the sample variance s_Δ^2 of those differences.

Equation 5.11 is an instance where positive correlation in the data is beneficial, in the sense that a more sensitive test can be achieved. Here a positive correlation results in a smaller standard deviation for

the sampling distribution of the difference of means being tested, implying less underlying uncertainty. This sharper null distribution produces a more powerful test and allows smaller differences in the numerator to be detected as significantly different from zero.

Intuitively this effect on the sampling distribution of the difference of sample means makes sense as well. Consider again the example of Ithaca and Canandaigua temperatures for January 1987, which will be revisited in Example 5.3. The positive correlation between daily temperatures at the two locations will result in the batch-to-batch (i.e., January-to-January, or interannual) variations in the two monthly averages moving together for the two locations: months when Ithaca is warmer than usual tend also to be months when Canandaigua is warmer than usual. The more strongly correlated are x_1 and x_2, the less likely are the pair of corresponding averages from a particular batch of data to differ because of sampling variations. To the extent that the two sample averages are different, then, the evidence against their underlying means not being the same is stronger, as compared to the situation when their correlation is near zero.

5.2.4. Tests for Differences of Mean Under Serial Dependence

The material in the previous sections is essentially a recapitulation of well-known tests for comparing sample means, presented in almost every elementary statistics textbook. A key assumption underlying these tests is the independence among the individual observations composing each of the sample means in the test statistic. That is, it is assumed that all the x_1 values are mutually independent and that the x_2 values are mutually independent, whether or not the data values are paired. This assumption of independence leads to the expression in Equation 5.4 that allows estimation of the variance of the null distribution.

Atmospheric data often do not satisfy the independence assumption. Frequently the averages to be tested are time averages, and the persistence, or time dependence, often exhibited is the cause of the violation of the assumption of independence. Lack of independence invalidates Equation 5.4. In particular, meteorological persistence implies that the variance of a time average is larger than specified by Equation 5.4. Ignoring the time dependence thus leads to underestimation of the variance of sampling distributions of the test statistics in Sections 5.2.2 and 5.2.3. This underestimation leads in turn to an inflated value of the test statistic, and consequently to p values that are too small, and to overconfidence regarding the significance of the difference in the numerator. Equivalently, properly representing the effect of persistence in the data will require a larger sample size to reject a null hypothesis for a given magnitude of the difference in the numerator.

Figure 5.4 illustrates why serial correlation leads to a larger variance for the sampling distribution of a time average. The upper panel of this figure is an artificial time series of 100 independent Gaussian variates drawn from a generating process with $\mu = 0$, as described in Section 4.7.4. The series in the lower panel also consists of Gaussian variables having $\mu = 0$, but in addition this series has a lag-1 autocorrelation (Equation 3.36) of $\rho_1 = 0.6$. This value of the autocorrelation was chosen here because it is typical of the autocorrelation exhibited by daily temperatures (e.g., Madden, 1979; Wilks and Wilby, 1999). Both panels have been scaled to produce unit (population) variance. The two plots look similar because the autocorrelated series was generated from the independent series according to what is called a first-order autoregressive process (Equation 10.16).

The outstanding difference between the independent and autocorrelated pseudo-data in Figure 5.4 is that the correlated series is smoother, so that adjacent and nearby values tend to be more alike than in the independent series. The autocorrelated series exhibits longer runs of points away from the

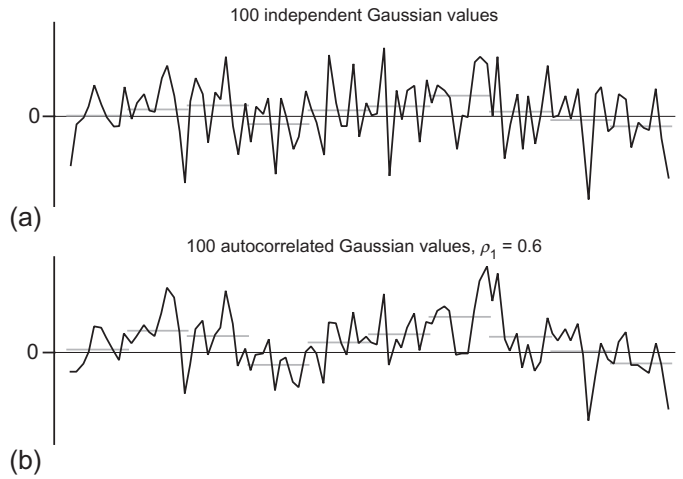

FIGURE 5.4 Comparison of artificial time series of (a) independent Gaussian variates, and (b) autocorrelated Gaussian variates having $\rho_1 = 0.6$. Both series were drawn from a generating process with $\mu = 0$, and the two panels have been scaled to have unit variances for the data points. Nearby values in the autocorrelated series tend to be more alike, with the result that averages over segments with $n = 10$ (*horizontal gray bars*) of the autocorrelated time series are more likely to be further from zero than are averages from the independent series. The sampling distribution of averages computed from the autocorrelated series accordingly has larger variance: the sample variances of the 10 subsample averages in panels (a) and (b) are 0.0825 and 0.2183, respectively.

(generating-process) mean value of zero. As a consequence, averages computed over subsets of the autocorrelated record are less likely to contain compensating points with large absolute value but of opposite sign, and those averages are therefore more likely to be far from zero than their counterparts computed using the independent values. That is, the averages over segments of the autocorrelated series will be less consistent from batch to batch. This is just another way of saying that the sampling distribution of an average of autocorrelated data has a higher variance than that of independent data. The gray horizontal lines in Figure 5.4 are subsample averages over consecutive sequences of $n = 10$ points, and these are visually more variable in Figure 5.4b. The sample variances of the 10 subsample means are 0.0825 and 0.2183 in panels (a) and (b), respectively.

Not surprisingly, the problem of estimating the variance of the sampling distribution of a time average has received considerable attention in the meteorological literature (e.g., Jones, 1975; Katz, 1982; Madden, 1979; Zwiers and Thiébaux, 1987; Zwiers and von Storch, 1995). One convenient and practical approach to dealing with the problem is to think in terms of the *effective sample size*, or *equivalent number of independent samples*, n'. That is, imagine that there is a fictitious sample size, $n' < n$ of independent values, for which the sampling distribution of the average has the same variance as the sampling distribution of the average over the n autocorrelated values at hand. Then, n' could be substituted for n in Equation 5.4, and the classical tests described in the previous section could be carried through as before.

Estimation of the effective sample size is most easily approached if it can be assumed that the underlying data follow a first-order autoregressive process (Equation 10.16). It turns out that first-order autoregressions are often reasonable approximations for representing the persistence of daily meteorological values. This assertion can be appreciated informally by looking at Figure 5.4b. This plot consists of random numbers exhibiting first-order autoregressive dependence, but resembles statistically the day-to-day fluctuations in a meteorological variable like surface temperature.

In general the effective sample size depends on the full autocorrelation function for the data-generating process, according to Equation 10.38. The persistence in a first-order autoregression is completely characterized by the single parameter ρ_1, the lag-1 autocorrelation coefficient, because in that case the autocorrelation function is given by $\rho_k = \rho_1^k$. The lag-1 autocorrelation can be estimated from a data series using the sample estimate, r_1 (Equation 3.36). Using this value, the effective sample size for the sampling variance of a mean can be estimated using the approximation

$$n' \approx n \frac{1-\rho_1}{1+\rho_1}, \tag{5.12}$$

which is valid for moderate and large n. When there is no time correlation, $\rho_1 = 0$ and $n' = n$. As ρ_1 increases the effective sample size becomes progressively smaller. When a more complicated time-series model is necessary to describe the persistence, appropriate but more complicated expressions for the effective sample size can be derived (see Katz, 1982, 1985; and Section 10.3.5).

Note that Equation 5.12 is applicable only to sampling distributions of the mean, and different expressions will be appropriate for use with different statistics (e.g., Ebisuzaki, 1997; Faes et al., 2009; Thiébaux and Zwiers, 1984). For example, the approximate effective sample size for the variance, again assuming an underlying first-order autoregressive generating process, is $n(1-\rho_1^2)/(1+\rho_1^2)$ (e.g., Preisendorfer et al., 1981), and so is relatively less affected by autocorrelation than the corresponding value for the mean in Equation 5.12.

Using Equation 5.12, the counterpart to Equation 5.4 for the variance of a time average over a sufficiently large sample becomes

$$\hat{\mathrm{Var}}[\bar{x}] = \frac{s^2}{n'} \approx \frac{s^2}{n}\left(\frac{1+\rho_1}{1-\rho_1}\right). \tag{5.13}$$

The ratio $(1+\rho_1)/(1-\rho_1)$ acts as a *variance inflation factor*, adjusting the variance of the sampling distribution of the time average upward to reflect the influence of the serial correlation. Sometimes this variance inflation factor is called the *time between effectively independent samples*, T_0 (e.g., Leith, 1973). Equation 5.4 can be seen as a special case of Equation 5.13, with $\rho_1 = 0$. Use of Equation 5.13 in the denominator of the test statistic in Equations 5.11 is in effect a special case of the *Diebold–Mariano test* (Diebold and Mariano, 1995, Hering and Genton 2011), which uses Equation 10.41, and so does not make a restrictive assumption about the correlation structure in the paired data, as the variance inflation factor.

Example 5.2. Unpaired Two-Sample *t*-Test for Autocorrelated Data

Magnitudes of autocorrelation typically exhibited by daily temperature data can have quite strong effects on statistical inferences based on such data. As an example, consider comparing the average 1988 temperatures for June and July at Ithaca, the values for which are listed in the second line of Table 5.2. The June temperature was slightly below average for the month, whereas the July temperature was several degrees warmer.

Considering also that July is climatologically warmer than June, a null hypothesis of equality for the means of the two data-generating processes that produced the monthly means in Table 5.2 seemingly should be easy to reject. However, because of the autocorrelation in the daily data that is near 0.5, Equation 5.12 yields effective sample sizes n' of 9.5 and 10.9 days for June and July respectively, which are reductions of about 2/3. Accordingly Equation 5.8 becomes $z = (83.5 - 75.2)/(7.75 + 11.8)^{1/2} = 1.88$.

TABLE 5.2 Summary Statistics Comparing Temperatures for June and July 1988 at Ithaca

	June	July
n	30	31
$\overline{x}$	75.2°F	83.5°F
s_x	10.6°F	9.2°F
r_1	0.52	0.48
n'	9.5	10.9
$\hat{\text{Var}}(\overline{x})$	11.8°F^2	7.75°F^2

Considering that we know enough about the climatology of the area to expect that if there are differences July will be warmer, then a one-tailed test is appropriate and the p value is 0.030. The null hypothesis of equality for the data-generating process means might be rejected, but not especially strongly. If we did not know in advance that July should be warmer, then the (two-tailed) p value is 0.060 and the inference is weaker still. In contrast, ignoring the autocorrelation and assuming serial independence of the underlying daily data leads to $z = 3.18$ in Equation 5.8, corresponding to the 1-tailed p value of 0.00074, which is wildly overconfident. ◇

Example 5.3. Paired Two-Sample *t*-Test for Autocorrelated Data

Consider testing whether the average maximum temperatures at Ithaca and Canandaigua for January 1987 (Table A.1 in Appendix A) are significantly different. This is equivalent to testing whether the difference of the two sample means is significantly different from zero, so that Equation 5.6 will hold for the null hypothesis. It has been shown previously (see Figure 3.5) that these two batches of daily data are reasonably symmetric and well behaved, so the sampling distributions of the monthly averages should be nearly Gaussian under the Central Limit Theorem.

The data for each location were observed on the same 31 days in January 1987, so the two batches are paired samples. Equation 5.11 is therefore the appropriate choice for the test statistic. Furthermore, we know that the daily data underlying the two time averages exhibit serial correlation (Figure 3.22 for the Ithaca data) so it may be expected that the effective sample size corrections in Equations 5.12 and 5.13 will be necessary as well.

Table A1 also shows the mean January 1987 temperatures, so the difference (Ithaca–Canandaigua) in mean maximum temperature is $29.87 - 31.77 = -1.9$°F. Computing the standard deviation of the differences between the 31 pairs of maximum temperatures yields $s_\Delta = 2.285$°F. The lag-1 autocorrelation for these differences is 0.076, yielding $n' = 31(1 - 0.076)/(1 + 0.076) = 26.6$. Since the null hypothesis is that the two population means are equal, $\mu_\Delta = 0$, and Equation 5.11 (using the effective sample size n' rather than the actual sample size n) yields $z = -1.9/(2.285^2/26.6)^{1/2} = -4.29$. This is a sufficiently extreme value not to be included in Table B.1, although Equation 4.30 estimates $\Phi(-4.29) \approx 0.000002$. The two-tailed p-value would be 0.000004, which is clearly significant. This extremely strong result is possible in part because much of the variability of the two temperature series is shared (the correlation between them is

0.957), and removing shared variance results in a rather small denominator for the test statistic. The magnitude of this shared variability is quantified in Example 5.18.

Notice that the lag-1 autocorrelation for the paired temperature differences is only 0.076, which is much smaller than the autocorrelations in the two individual series: 0.52 for Ithaca and 0.61 for Canandaigua. Much of the temporal dependence is also exhibited jointly by the two series, and so is removed when calculating the differences Δ_i. Here is another advantage of using the series of differences to conduct this test, and another major contribution to the strong result. The relatively low autocorrelation of the difference series translates into an effective sample size of 26.6 rather than only 9.8 (Ithaca) and 7.5 (Canandaigua), which produces an even more sensitive test.

Finally, consider the confidence interval for the mean difference μ_Δ in relation to the confidence intervals that would be calculated for the individual means μ_{Ith} and $\mu_{\text{Can.}}$ The 95% confidence interval around the observed mean difference of $\overline{x}_\Delta = -1.9°\text{F}$ is $-1.9°\text{F} \pm (1.96)(2.285)/(\sqrt{26.6})$, yielding the interval $[-2.77, -1.03]$. Consistent with the extremely low p value for the paired comparison test, this interval does not include zero, and indeed its maximum is well away from zero in standard error (standard deviation of the sampling distribution of $\overline{\Delta}$) units. In contrast, consider the 95% confidence intervals around the individual sample means for Ithaca and Canandaigua. For Ithaca, this interval is $29.9°\text{F} \pm (1.96)(7.71)/(\sqrt{9.8})$, or [25.0, 34.7], whereas for Canandaigua it is $31.8°\text{F} \pm (1.96)(7.86)/(\sqrt{7.5})$, or [26.2, 37.4]. Not only do these two intervals overlap substantially, their length of overlap is greater than the sum of the lengths over which they do not overlap. Thus evaluating the significance of this difference using the so-called overlap method leads to a highly erroneous conclusion. This example provides a nearly worst case for the overlap method, since in addition to the two variances being nearly equal, members of the two data samples are strongly correlated, which also exacerbates the discrepancy with a correctly computed test (Jolliffe, 2007; Schenker and Gentleman, 2001). ◇

Autocorrelation of values in a time series is not the only kind of statistical dependence that may affect statistical inferences. Director and Bornn (2015) address effective sample size adjustments in the context of spatial data.

5.2.5. Goodness-of-Fit Tests

When discussing the fitting of parametric distributions to data samples in Chapter 4, methods for visually and subjectively assessing the *goodness of fit* were presented. Formal, quantitative evaluations of the goodness of fit also exist, and these are carried out within the framework of hypothesis testing. The graphical methods can still be useful when formal tests are conducted, for example, in pointing out where and how a lack of fit is manifested. Many goodness-of-fit tests have been devised, but only a few common ones are presented here.

Assessing goodness of fit presents an atypical hypothesis test setting, in that these tests usually are computed to obtain evidence in favor of H_0, that the data at hand were drawn from a hypothesized distribution. The interpretation of confirmatory evidence is then that the data are "not inconsistent" with the hypothesized distribution, so the power of these tests is an important consideration. Unfortunately, because there are any number of ways in which the null hypothesis can be wrong in this setting, it is usually not possible to formulate a single best (most powerful) test. This problem accounts in part for the large number of goodness-of-fit tests that have been proposed (D'Agostino and Stephens, 1986), and for the ambiguity about which might be most appropriate for a particular problem. Note also that the tests described in this section assume mutually independent data, and as is the case for other

hypothesis tests their uncorrected implementation using autocorrelated data will yield erroneously small p values.

Chi-Square Test

The *chi-square* (χ^2) *test* is a simple and common goodness-of-fit test. It essentially compares a data histogram with the probability distribution (for discrete variables) or probability density (for continuous variables) function. The χ^2 test actually operates more naturally for discrete random variables, since to implement it the range of the data must be divided into discrete classes, or bins. When alternative tests are available for continuous data they are usually more powerful, presumably at least in part because the rounding of data into bins, which may be severe, discards information. However, the χ^2 test is easy to implement and quite flexible, being for example, very straightforward to implement for multivariate data.

For continuous random variables, the probability density function is integrated over each of some number of MECE classes to obtain probabilities for data values in each class. Regardless of whether the data are discrete or continuous, the test statistic involves the counts of data values falling into each class in relation to the computed theoretical probabilities,

$$\begin{aligned}\chi^2 &= \sum_{\text{classes}} \frac{(\#\text{ Observed} - \#\text{ Expected})^2}{\#\text{ Expected}} \\ &= \sum_{\text{classes}} \frac{(\#\text{ Observed} - n\Pr\{\text{data in class}\})^2}{n\Pr\{\text{data in class}\}}.\end{aligned} \tag{5.14}$$

In each class, the number (#) of data values expected to occur, according to the fitted distribution, is simply the probability of occurrence in that class multiplied by the sample size, n. This number of expected occurrences need not be an integer value. If the fitted distribution is very close to the data distribution, the expected and observed counts will be very close for each class, and the squared differences in the numerator of Equation 5.14 will all be very small, yielding a small χ^2. If the fit is not good, at least a few of the classes will exhibit large discrepancies. These will be squared in the numerator of Equation 5.14 and lead to large values of χ^2. It is not necessary for the classes to be of equal width or equal probability, but classes with small numbers of expected counts should be avoided. Sometimes a minimum of five expected events per class is imposed.

Under the null hypothesis that the data were drawn from the fitted distribution, the sampling distribution for the test statistic is the χ^2 distribution with parameter $\nu = (\#\text{ of classes} - \#\text{ of parameters fit} - 1)$ degrees of freedom. The test will be one-sided, because the test statistic is confined to positive values by the squaring process in the numerator of Equation 5.14, and small values of the test statistic support H_0. Right-tail quantiles for the χ^2 distribution are given in Table B.3.

Example 5.4. Comparing Gaussian and Gamma Distribution Fits Using the χ^2 Test

Consider the gamma and Gaussian distributions as candidates for representing the 1933–82 Ithaca January precipitation data in Table A.2. The approximate maximum likelihood estimators for the gamma distribution parameters (Equations 4.48 or 4.50a, and Equation 4.49) are $\alpha = 3.76$ and $\beta = 0.52$ in. The sample mean and standard deviation (i.e., the Gaussian parameter estimates) for these data are 1.96 in. and 1.12 in., respectively. The two fitted distributions are illustrated in relation to the data in Figure 4.16. Table 5.3 contains the information necessary to conduct the χ^2 tests for these two distributions. The precipitation amounts have been divided into six classes, or bins, the limits of which are indicated in the first

TABLE 5.3 The χ^2 Goodness-of-Fit Test Applied to Gamma and Gaussian Distributions for the 1933–82 Ithaca January Precipitation Data

Class	<1″	1–1.5″	1.5–2″	2–2.5″	2.5–3″	>3″
Observed #	5	16	10	7	7	5
Gamma:						
Probability	0.161	0.215	0.210	0.161	0.108	0.145
Expected #	8.05	10.75	10.50	8.05	5.40	7.25
Gaussian:						
Probability	0.195	0.146	0.173	0.178	0.132	0.176
Expected #	9.75	7.30	8.65	8.90	6.60	8.80

Expected numbers of occurrences in each bin are obtained by multiplying the respective probabilities by $n = 50$.

row of the table. The second row indicates the number of years in which the January precipitation total was within each class. Both distributions have been integrated over these classes to obtain probabilities for precipitation in each class. These probabilities were then multiplied by $n = 50$ to obtain the expected number of counts.

Applying Equation 5.14 yields $\chi^2 = 5.05$ for the gamma distribution and $\chi^2 = 14.96$ for the Gaussian distribution. As was also evident from the graphical comparison in Figure 4.16, these test statistics indicate that the Gaussian distribution fits these precipitation data substantially less well. Under the respective null hypotheses, these two test statistics are drawn from a χ^2 distribution with degrees of freedom $\nu = 6 - 2 - 1 = 3$; because Table 5.3 contains six classes, and two parameters (α and β, or μ and σ, for the gamma or Gaussian, respectively) were fit for each distribution.

Referring to the $\nu = 3$ row of Table B.3, $\chi^2 = 5.05$ is smaller than the 90th percentile value of 6.251, so the null hypothesis that the data have been drawn from the fitted gamma distribution would not be rejected even at the 10% level. For the Gaussian fit, $\chi^2 = 14.96$ is between the tabulated values of 11.345 for the 99th percentile and 16.266 for the 99.9th percentile, so this null hypothesis would be rejected at the 1% level, but not at the 0.1% level. ◇

Kolmogorov–Smirnov and Lilliefors Tests

The one-sample *Kolmogorov–Smirnov (K–S) test* is another very frequently used test of the goodness of fit. The χ^2 test essentially compares the histogram and the PDF or discrete distribution function, whereas the K–S test compares the empirical and fitted CDFs. Again, the null hypothesis is that the observed data were drawn from the distribution being tested, and a sufficiently large discrepancy will result in the null hypothesis being rejected. For continuous distributions the K–S test usually will be more powerful than the χ^2 test, and so usually will be preferred.

In its original form, the K–S test is applicable to any distributional form (including but not limited to any of the distributions presented in Chapter 4), provided that the parameters have *not* been estimated from the data sample being compared. In practice this provision can be a serious limitation to the use of the original K–S test, since it is often the correspondence between a fitted distribution and the particular batch of data used to estimate its parameters that is of interest. This may seem like a trivial problem, but it can have serious consequences, as has been pointed out by Crutcher (1975) and Steinskog et al. (2007). Estimating the parameters from the same batch of data used to test the goodness of fit results in the fitted distribution parameters being "tuned" to the data sample. When erroneously using K–S critical values that assume independence between the test data and the estimated parameters, it will often be the case that the null hypothesis (that the distribution fits well) will not be rejected when in fact it should be.

With modification, the K–S framework can be used in situations where the distribution parameters have been fit to the same data used in the test. In this situation, the K–S test is often called the *Lilliefors test*, after the statistician who did much of the early work on the subject (Lilliefors, 1967). Both the original K–S test and the Lilliefors test use the test statistic

$$D_n = \max_x |F_n(x) - F(x)|, \tag{5.15}$$

where $F_n(x)$ is the empirical cumulative probability, estimated as $F_n(x_{(i)}) = i/n$ for the ith smallest data value; and $F(x)$ is the theoretical cumulative distribution function evaluated at x (Equation 4.19). Thus the K–S test statistic D_n looks for the largest difference, in absolute value, between the empirical and fitted cumulative distribution functions. Any real and finite batch of data will exhibit sampling fluctuations resulting in a nonzero value for D_n, even if the null hypothesis is true and the theoretical distribution fits very well. If D_n is sufficiently large, the null hypothesis can be rejected. How large is large enough depends on the level of the test, of course; but also on the sample size, whether or not the distribution parameters have been fit using the test data, and if so also on the particular distribution form being fit.

When the parametric distribution to be tested has been specified completely externally to the data—the data have not been used in any way to fit the parameters—the original K–S test is appropriate. This test is distribution free, in the sense that its critical values are applicable to any distribution. These critical values can be obtained to good approximation (Stephens, 1974) using

$$C_\alpha = \frac{K_\alpha}{\sqrt{n} + 0.12 + 0.11/\sqrt{n}}, \tag{5.16}$$

where $K_\alpha = 1.224$, 1.358, and 1.628, for $\alpha = 0.10$, 0.05, and 0.01, respectively. The null hypothesis is rejected for $D_n \geq C_\alpha$. Alternatively, for n larger than perhaps 50, these critical levels can be obtained as continuous functions of α, using

$$C_\alpha = \sqrt{-\frac{\ln(\alpha/2)}{2n}}, \tag{5.17}$$

which in turn can be inverted to calculate p values for particular D_n.

Usually the original K–S test (and therefore Equations 5.16 and 5.17) is not appropriate because the parameters of the distribution being tested have been fit using the test data. But even in this case bounds

TABLE 5.4 Critical Values for the K–S Statistic D_n Used in the Lilliefors Test to Assess Goodness of Fit of Gamma Distributions, as a Function of the Estimated Shape Parameter, α, When the Distribution Parameters have been Fit Using the Data to be Tested

	20% level			10% level			5% level			1% level		
α	$n=25$	$n=30$	Large n	$n=25$	$n=30$	Large n	$n=25$	$n=30$	Large n	$n=25$	$n=30$	Large n
1	0.165	0.152	$0.84/\sqrt{n}$	0.185	0.169	$0.95/\sqrt{n}$	0.204	0.184	$1.05/\sqrt{n}$	0.241	0.214	$1.20/\sqrt{n}$
2	0.159	0.146	$0.81/\sqrt{n}$	0.176	0.161	$0.91/\sqrt{n}$	0.190	0.175	$0.97/\sqrt{n}$	0.222	0.203	$1.16/\sqrt{n}$
3	0.148	0.136	$0.77/\sqrt{n}$	0.166	0.151	$0.86/\sqrt{n}$	0.180	0.165	$0.94/\sqrt{n}$	0.214	0.191	$1.08/\sqrt{n}$
4	0.146	0.134	$0.75/\sqrt{n}$	0.164	0.148	$0.83/\sqrt{n}$	0.178	0.163	$0.91/\sqrt{n}$	0.209	0.191	$1.06/\sqrt{n}$
8	0.143	0.131	$0.74/\sqrt{n}$	0.159	0.146	$0.81/\sqrt{n}$	0.173	0.161	$0.89/\sqrt{n}$	0.203	0.187	$1.04/\sqrt{n}$
∞	0.142	0.131	$0.736/\sqrt{n}$	0.158	0.144	$0.805/\sqrt{n}$	0.173	0.161	$0.886/\sqrt{n}$	0.200	0.187	$1.031/\sqrt{n}$

The row labeled $\alpha=\infty$ pertains to the Gaussian distribution with parameters estimated from the data.
From Crutcher (1975). © American Meteorological Society. Used with permission.

on the true CDF, whatever its form, can be computed and displayed graphically using $F_n(x) \pm C_\alpha$ as limits covering the actual cumulative probabilities, with probability $1-\alpha$. Values of C_α can also be used in an analogous way to calculate probability bounds on empirical quantiles consistent with a particular theoretical distribution (Loucks et al., 1981). Because the D_n statistic is a maximum over the entire data set, these bounds are valid jointly, for the entire distribution.

When the distribution parameters have been fit using the data at hand, Equation 5.16 is not sufficiently stringent, because the fitted distribution "knows" too much about the data to which it is being compared, and the Lilliefors test is appropriate. Here, however, the critical values of D_n depend on the distribution that has been fit. Table 5.4, from Crutcher (1975), lists critical values of D_n (above which the null hypothesis would be rejected) for four test levels, for the gamma distribution. These critical values depend on both the sample size and the estimated shape parameter, α. Larger samples will be less subject to irregular sampling variations, so the tabulated critical values decline for larger n. That is, smaller maximum deviations from the fitted distribution (Equation 5.15) are tolerated for larger sample sizes. Critical values in the last row of the table, for $\alpha = \infty$, pertain to the Gaussian distribution, since as the gamma shape parameter becomes very large the gamma distribution converges toward the Gaussian.

It is interesting to note that critical values for Lilliefors tests are usually derived through statistical simulation (see Section 4.7). The procedure is that a large number of samples from a known distribution are generated, and estimates of the distribution parameters are calculated from each of these samples. The agreement, for each synthetic data batch, between data generated from the known distribution and the distribution fit to it is then assessed using Equation 5.15. Since the null hypothesis is true in this protocol by construction, the α-level critical value is approximated as the $(1-\alpha)$ quantile of that collection of synthetic D_n's. Thus Lilliefors-test critical values for any distribution that may be of interest can be computed using the methods described in Section 4.7. These methods can also be adapted to represent data having known forms and magnitude of autocorrelation.

Example 5.5. Comparing Gaussian and Gamma Fits Using the K–S Test
Again consider the fits of the gamma and Gaussian distributions to the 1933–82 Ithaca January precipitation data, from Table A.2, shown in Figure 4.16. Figure 5.5 illustrates the Lilliefors tests for these two fitted distributions. In each panel of Figure 5.5, the black dots are the empirical cumulative probability estimates, $F_n(x)$, and the smooth curves are the fitted theoretical CDFs, $F(x)$, both plotted as functions of the observed monthly precipitation. Coincidentally, the maximum differences between the empirical and fitted theoretical cumulative distribution functions occur at the same (highlighted) point, yielding $D_n = 0.068$ for the gamma distribution (a) and $D_n = 0.131$ for the Gaussian distribution (b).

In each of the two tests to be conducted the null hypothesis is that the precipitation data were drawn from the fitted distribution, and the alternative hypothesis is that they were not. These will necessarily be one-sided tests, because the test statistic D_n is the absolute value of the largest difference between the parametric and empirical cumulative probabilities. Therefore values of the test statistic on the far right tail of the null distribution will indicate large discrepancies that are unfavorable to H_0, whereas values of the test statistic on the left tail of the null distribution will indicate $D_n \approx 0$, or near-perfect fits that are very supportive of the null hypothesis.

The critical values in Table 5.4 are the minimum D_n necessary to reject H_0. That is, they are leftmost bounds of the relevant rejection or critical regions. The sample size of $n = 50$ is sufficient to evaluate the tests using critical values from the large-n columns. In the case of the Gaussian distribution, the relevant row of the table is for $\alpha = \infty$. Since $0.886/\sqrt{50} = 0.125$ and $1.031/\sqrt{50} = 0.146$ bound the observed $D_n = 0.131$, the null hypothesis that the precipitation data were drawn from this Gaussian distribution would be rejected at the 5% level, but not the 1% level. For the fitted gamma distribution the nearest row in Table 5.4 is for $\alpha = 4$, where even at the 20% level the critical value of $0.75/\sqrt{50} = 0.106$ is substantially larger than the observed $D_n = 0.068$. Thus these data are quite consistent with the proposition of their having been drawn from this gamma distribution.

Regardless of the distribution from which these data were drawn, it is possible to use Equation 5.16 to calculate confidence intervals on its CDF. Using $K_\alpha = 1.358$, the *gray dots* in Figure 5.5 show the 95% confidence intervals for $n = 50$ as $F_n(x) \pm 0.188$. The intervals defined by these points cover the true CDF with 95% confidence, throughout the range of the data, because the K–S statistic pertains to the largest difference between $F_n(x)$ and $F(x)$, regardless of where in the distribution that maximum discrepancy may occur for a particular sample. ◇

The related two-sample K–S test, or *Smirnov test*, compares two batches of data to one another under the null hypothesis that they were drawn from the same (but unspecified) distribution or generating process. The Smirnov test statistic,

$$D_S = \max_x |F_{n}(x_1) - F_{n}(x_2)|, \tag{5.18}$$

looks for the largest (in absolute value) difference between the empirical cumulative distribution functions of samples of n_1 observations of x_1 and n_2 observations of x_2. Unequal sample sizes can be accommodated by the Smirnov test because the empirical CDFs are step functions (e.g., Figure 3.11), so that this maximum can occur at any of the values of x_1 or x_2. Again, the test is one-sided because of the absolute values in Equation 5.18, and the null hypothesis that the two data samples were drawn from the same distribution is rejected at the $\alpha \cdot 100\%$ level if

$$D_S > \left[-\frac{1}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \ln \left(\frac{\alpha}{2} \right) \right]^{1/2}. \tag{5.19}$$

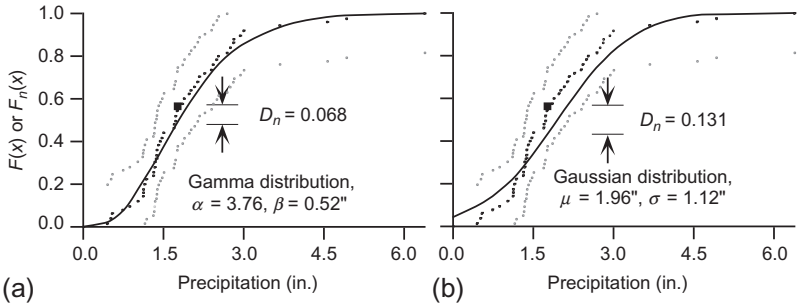

FIGURE 5.5 Illustration of the Kolmogorov–Smirnov D_n statistic in Lilliefors tests applied to the 1933–82 Ithaca January precipitation data fitted to fitted (a) gamma and (b) Gaussian distributions. *Solid curves* indicate cumulative distribution functions, and *black dots* show the corresponding empirical estimates. The maximum difference between the empirical and theoretical CDFs occurs for the *highlighted square point*, and is substantially greater for the Gaussian distribution. *Gray dots* show limits of the 95% confidence interval for the true CDF from which the data were drawn (Equation 5.16).

Filliben Q–Q Test

A good test for Gaussian distribution is often needed, for example, when the multivariate Gaussian distribution (see Chapter 12) will be used to represent the joint variations of (possibly power-transformed, Section 3.4.1) multiple variables. The Lilliefors test (Table 5.4, with $\alpha = \infty$) is an improvement in terms of power over the chi-square test for this purpose, but tests that are generally better (D'Agostino, 1986; Razali and Wah, 2011) can be constructed on the basis of the correlation between the empirical quantiles (i.e., the data), and the Gaussian quantile function based on their ranks. This approach was introduced by Shapiro and Wilk (1965), and both their original test formulation and its subsequent variants are known as *Shapiro–Wilk tests*. A computationally simple variant that is nearly as powerful as the original Shapiro–Wilk formulation was proposed by Filliben (1975). The test statistic is simply the correlation (Equation 3.32) between the empirical quantiles $x_{(i)}$ and the Gaussian quantile function $\Phi^{-1}(p_i)$, with p_i estimated using a plotting position (see Table 3.2) approximating the median cumulative probability for the ith order statistic (e.g., the Tukey plotting position, although Filliben (1975) used Equation 3.19 with $a = 0.3175$). That is, the test statistic is simply the correlation computed from the points on a Gaussian Q–Q plot. If the data are drawn from a Gaussian distribution these points should fall on a straight line, apart from sampling variations.

Table 5.5 presents critical values for the *Filliben test* for Gaussian distribution. The test is one-tailed, because high correlations are favorable to the null hypothesis that the data are Gaussian, so the null hypothesis is rejected if the correlation is smaller than the appropriate critical value. Because the points on a Q–Q plot are necessarily nondecreasing, the critical values in Table 5.5 are much larger than would be appropriate for testing the significance of the linear association between two independent (according to a null hypothesis) variables. Notice that, since the correlation will not change if the data are first standardized (Equation 3.27), this test does not depend in any way on the accuracy with which the distribution parameters may have been estimated. That is, the test addresses the question of whether the data were drawn from a Gaussian distribution but does not address, and is not confounded by, the question of what the parameters of that distribution might be.

TABLE 5.5 Critical Values for the Filliben (1975) Test for Gaussian Distribution, Based on the Q–Q Plot Correlation

n	0.5% Level	1% Level	5% Level	10% Level
10	0.860	0.876	0.917	0.934
20	0.912	0.925	0.950	0.960
30	0.938	0.947	0.964	0.970
40	0.949	0.958	0.972	0.977
50	0.959	0.965	0.977	0.981
60	0.965	0.970	0.980	0.983
70	0.969	0.974	0.982	0.985
80	0.973	0.976	0.984	0.987
90	0.976	0.978	0.985	0.988
100	0.9787	0.9812	0.9870	0.9893
200	0.9888	0.9902	0.9930	0.9942
300	0.9924	0.9935	0.9952	0.9960
500	0.9954	0.9958	0.9970	0.9975
1000	0.9973	0.9976	0.9982	0.9985

H_0 is rejected if the correlation is smaller than the appropriate critical value.

Example 5.6. Filliben Q–Q Correlation Test for Gaussian Distribution

The Q–Q plots in Figure 4.17 showed that the Gaussian distribution fits the 1933–82 Ithaca January precipitation data in Table A.2 less well than the gamma distribution. That Gaussian Q–Q plot is reproduced in Figure 5.6 (x's), with the horizontal axis scaled to correspond to standard Gaussian quantiles, z, rather than to dimensional precipitation amounts. Using the Tukey plotting position (see Table 3.2), estimated cumulative probabilities corresponding to (for example) the smallest and largest of these $n = 50$ precipitation amounts are $0.67/50.33 = 0.013$ and $49.67/50.33 = 0.987$. Standard Gaussian quantiles, z, corresponding to these cumulative probabilities (see Table B.1) are ± 2.22. The correlation for these $n = 50$ untransformed points is $r = 0.917$, which is smaller than all of the critical values in that row of Table 5.5. Accordingly, the Filliben test would reject the null hypothesis that these data were drawn from a Gaussian distribution, at the 0.5% level. The fact that the horizontal scale is the nondimensional z rather than dimensional precipitation (as in Figure 4.17) is immaterial, because the correlation is unaffected by linear transformations of either or both of the two variables being correlated.

Figure 3.14, in Example 3.4, indicated that a logarithmic transformation of these data was effective in producing approximate symmetry. Whether this transformation is also effective at producing a plausibly Gaussian shape for these data can be addressed with the Filliben test. Figure 5.6 also shows the standard Gaussian Q–Q plot for the log-transformed Ithaca January precipitation totals (o's). This relationship

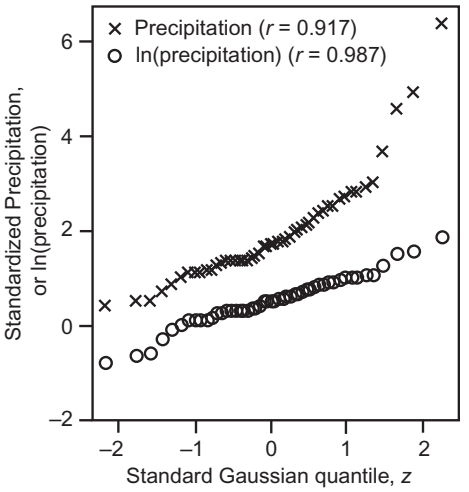

FIGURE 5.6 Standard Gaussian Q–Q plots for the 1933–82 Ithaca January precipitation in Table A.2 (x's), and for the log-transformed data (o's). Using Table 5.5, null hypotheses that these data were drawn from Gaussian distributions would be rejected for the original data ($p < 0.005$), but not rejected for the log-transformed data ($p > 0.10$).

is substantially more linear than for the untransformed data and is characterized by a correlation of $r = 0.987$. Again looking on the $n = 50$ row of Table 5.5, this correlation is larger than the 10% critical value, so the null hypothesis of Gaussian distribution would not be rejected.

Notice that Example 5.4 found that these data were also not inconsistent with a fitted gamma distribution. The goodness-of-fit tests cannot tell us whether these data were drawn from a gamma distribution, a lognormal distribution, or some other distribution that is similar to both the gamma and lognormal distributions that have been fit to these data. This ambiguity illustrates the situation that logically weaker statistical inferences result from nonrejection of null hypotheses. ◇

Using statistical simulation (see Section 4.7), tables of critical Q–Q correlations can be obtained for other distributions, by generating large numbers of batches of size n from the distribution of interest, computing Q–Q plot correlations for each of these batches, and defining the critical value as that delineating the extreme $\alpha \cdot 100\%$ smallest of them. Results of this approach have been tabulated for the Gumbel distribution (Vogel, 1986), the uniform distribution (Vogel and Kroll, 1989), the GEV distribution (Chowdhury et al., 1991), and the Pearson III distribution (Vogel and McMartin, 1991). Heo et al. (2008) present regression-based critical values for Q–Q correlation tests pertaining to Gaussian, Gumbel, gamma, GEV, and Weibull distributions. The underlying simulations can also be adapted to represent data having known forms and magnitudes of autocorrelation.

5.2.6. Likelihood Ratio Tests

Sometimes we need to construct a test in a parametric setting, but the hypothesis is sufficiently complex that the simple, familiar parametric tests cannot be brought to bear. A flexible alternative, known as the *likelihood ratio test*, can be used if two conditions are satisfied. First, it must be possible to cast the problem in such a way that the null hypothesis pertains to some number, k_0 of free (i.e., fitted) parameters, and the alternative hypothesis pertains to some larger number, $k_A > k_0$, of parameters. Second, it must be possible to regard the k_0 parameters of the null hypothesis as a special case of the full parameter

set of k_A parameters. That is, the null hypothesis is "nested" within the alternatives. Examples of this second condition on H_0 could include forcing some of the k_A parameters to have fixed values (often zero), or imposing equality between two or more of them. As the name implies, the likelihood ratio test compares the likelihoods associated with H_0 vs. H_A, when the k_0 and k_A parameters, respectively, have been fit using the method of maximum likelihood (see Section 4.6).

Even if the null hypothesis is true, the likelihood associated with H_A will always be at least as large as that for H_0. This is because the greater number of parameters $k_A > k_0$ allows the maximized likelihood function for the former greater freedom in accommodating the observed data. The null hypothesis is therefore rejected only if the likelihood associated with the alternative is sufficiently large that the difference is unlikely to have resulted from sampling variations.

The test statistic for the likelihood ratio test is

$$\Lambda^* = 2 \ln\left[\frac{\Lambda(H_A)}{\Lambda(H_0)}\right] = 2\,[L(H_A) - L(H_0)]. \tag{5.20}$$

This quantity is also known as the *deviance*. Here $\Lambda(H_0)$ and $\Lambda(H_A)$ are the likelihood functions (see Section 4.6) associated with the null and alternative hypothesis, respectively. The second equality, involving the difference of the log-likelihoods $L(H_0) = \ln[\Lambda(H_0)]$ and $L(H_A) = \ln[\Lambda(H_A)]$, is used in practice since it is generally the log-likelihoods that are maximized (and thus computed) when fitting the parameters.

Under H_0, and given a large sample size, the sampling distribution of the statistic in Equation 5.20 is χ^2, with degrees of freedom $\nu = k_A - k_0$. That is, the degrees-of-freedom parameter is given by the difference between H_A and H_0 in the number of empirically estimated parameters. Since small values of Λ^* are not unfavorable to H_0, the test is one-sided and H_0 is rejected only if the observed Λ^* is in a sufficiently improbable region on the right tail.

Example 5.7. Testing for Climate Nonstationarity Using the Likelihood Ratio Test

Suppose there is a reason to suspect that the first 25 years (1933–57) of the Ithaca January precipitation data in Table A.2 have been drawn from a different gamma distribution than the second half (1958–82). This question can be tested against the null hypothesis that all 50 precipitation totals were drawn from the same gamma distribution using a likelihood ratio test. To perform the test it is necessary to fit gamma distributions separately to the two halves of the data, and compare these two distributions with the single gamma distribution fit using the full data set.

The relevant information is presented in Table 5.6, which indicates some differences between the two 25-year periods. For example, the average January precipitation ($=\alpha\beta$) for 1933–57 was 1.87 in., and the

TABLE 5.6 Gamma Distribution Parameters (MLEs) and Log-Likelihoods for Fits to the First and Second Halves of the 1933–82 Ithaca January Precipitation Data in Table A.2, and to the Full Data Set

	Dates	α	β	$\Sigma_i L(\alpha, \beta; x_i)$
H_A	1933–57	4.525	0.4128	−30.2796
	1958–82	3.271	0.6277	−35.8965
H_0	1933–82	3.764	0.5209	−66.7426

corresponding average for 1958–82 was 2.05 in. The year-to-year variability ($=\alpha\beta^2$) of January precipitation was greater in the second half of the period as well. Whether the extra two parameters required to represent the January precipitation using two gamma distributions rather than one are justified by the data can be evaluated using the test statistic in Equation 5.20. For this specific problem the test statistic is

$$\Lambda^* = 2\left\{\left[\sum_{i=1933}^{1957} L(\alpha_1, \beta_1; x_i)\right] + \left[\sum_{i=1958}^{1982} L(\alpha_2, \beta_2; x_i)\right] - \left[\sum_{i=1933}^{1982} L(\alpha_0, \beta_0; x_i)\right]\right\}, \tag{5.21}$$

where the subscripts 1, 2, and 0 on the parameters refer to the first half, the second half, and the full period (null hypothesis), respectively, and the log-likelihood for the gamma distribution given a single observation, x_i, is (compare Equation 4.45)

$$L(\alpha, \beta; x_i) = (\alpha - 1)\,\ln(x_i/\beta) - x_i/\beta - \ln(\beta) - \ln[\Gamma(\alpha)]. \tag{5.22}$$

The three terms in square brackets in Equation 5.21 are given in the last column of Table 5.6.

Using the information in Table 5.6, $\Lambda^* = 2\,(-30.2796 - 35.8965 + 66.7426) = 1.130$. Since there are $k_A = 4$ parameters under H_A (α_1, β_1, α_2, β_2) and $k_0 = 2$ parameters under H_0 (α_0, β_0), the null distribution is the χ^2 distribution with $\nu = 2$. Looking on the $\nu = 2$ row of Table B.3, we find $\chi^2 = 1.130$ is smaller than the median value, leading to the conclusion that the observed Λ^* is quite ordinary in the context of the null hypothesis that the two data records were drawn from the same gamma distribution, which would not be rejected. More precisely, recall that the χ^2 distribution with $\nu = 2$ is itself a gamma distribution with $\alpha = 1$ and $\beta = 2$, which in turn is the exponential distribution with $\beta = 2$. The exponential distribution has the closed-form CDF in Equation 4.53, which yields the right-tail probability (p value) $1 - F(1.130) = 0.5684$. ◇

Example 5.8. Likelihood Ratio Tests comparing Simpler Versus More Elaborate Fitted Distributions

The annual maximum daily precipitation amounts in Table 4.6 can be represented by a GEV distribution (Equation 4.63) with parameters $\zeta = 3.49$, $\beta = 1.18$, and $\kappa = -0.32$, for which the minimized log-likelihood is $L_{\text{GEV}} = -30.273$. Since the fitted shape parameter κ is relatively small in absolute value it may be of interest to investigate whether the three-parameter GEV is justified by these data, in preference to fitting a two-parameter Gumbel distribution (Equation 4.66), effectively forcing $\kappa = 0$. Maximum likelihood estimates for the two Gumbel distribution parameters for these data are $\zeta = 3.33$ and $\beta = 1.05$, which yield the log-likelihood $L_{\text{Gumbel}} = -31.519$. Of course $L_{\text{GEV}} > L_{\text{Gumbel}}$ since the GEV has an additional free parameter, but whether this additional complexity is supported by the data can be investigated using a likelihood ratio test because the Gumbel distribution is a special case of the GEV. Here the null hypothesis is that the Gumbel distribution is adequate, and the alternative hypothesis is that these data support use of the GEV. The resulting test statistic (Equation 5.20) is therefore $\Lambda^* = 2\,(-30.273 + 31.519) = 2.49$. If the null hypothesis is true then this statistic follows the χ^2 distribution with $\nu = 3 - 2 = 1$ degree of freedom. Consulting the $\nu = 1$ row of Table B.3 we find that 2.49 is a bit smaller than the 90th percentile of this distribution, so that the right-tail p value is larger than 0.10. It appears that additional complexity of the GEV is not justified by these data, since the null hypothesis that they are Gumbel distributed has not been rejected.

Figure 4.14 shows a two-component Gaussian mixture distribution (Equation 4.75) representing the Guayaquil temperature data in Table A.3. The final line of Table 4.8 shows the five parameters that have

been estimated using maximum likelihood, together with the maximized log-likelihood $L_{\text{mix}} = -20.48$. Might these data be represented adequately by a single Gaussian distribution? The mean and standard deviation (Equation 4.84a and b) of the 20 temperature values in Table A.3 are 24.76°C and 0.93°C, respectively, corresponding to the maximized log-likelihood (sum of 20 terms of the form of Equation 4.82) $L_{\text{Gauss}} = -27.47$. Since this latter distribution is a special case of the two-component Gaussian mixture, a likelihood ratio test can be computed in order to gauge how well the data justify the more elaborate model. The null hypothesis is that the single Gaussian distribution is adequate, and the alternative is that the five-parameter Gaussian mixture is supported by the data. The test statistic is $\Lambda^* = 2(-20.48 + 27.47) = 13.98$. The appropriate χ^2 null distribution has $\nu = 5 - 2 = 3$ degrees of freedom, and the $\nu = 3$ row of Table B.3 shows that the null hypothesis is rejected, with $0.001 < p < 0.01$, supporting the use of the mixture model. ◇

5.3. NONPARAMETRIC TESTS

Not all formal hypothesis tests rest on assumptions involving specific parametric distributions for the data or for the sampling distributions of the test statistics. Tests not requiring such assumptions are called *nonparametric* or *distribution free*. Nonparametric methods are appropriate if either or both of the following conditions apply:

1. We know or suspect that the parametric assumption(s) required for a particular test are not met, for example, grossly non-Gaussian data in conjunction with the *t*-test for the difference of means in Equation 5.5.
2. A test statistic that is suggested or dictated by the scientific problem at hand is a complicated function of the data, and its sampling distribution is unknown and/or cannot be derived analytically.

The same hypothesis testing ideas apply to both parametric and nonparametric tests. In particular, the five elements of the hypothesis test presented at the beginning of this chapter apply also to nonparametric tests. The difference between parametric and nonparametric tests is in the way the null distribution is obtained in Step 4.

There are two branches of nonparametric testing. The first, called *classical nonparametric testing* in the following, consists of tests based on mathematical analysis of selected hypothesis test settings. These are older methods, devised before the advent of cheap and widely available computing. They employ analytic mathematical results (formulas) that are applicable to data drawn from any distribution. Only a few classical nonparametric tests will be presented here, although the range of classical nonparametric methods is much more extensive (e.g., Conover, 1999; Daniel, 1990; Sprent and Smeeton, 2001).

The second branch of nonparametric testing includes procedures collectively called *resampling tests*. A resampling test builds up a discrete approximation to the null distribution using a computer, by repeatedly operating on (resampling) the data set at hand. Since the null distribution is arrived at empirically, the analyst is free to use virtually any computable test statistic that may be relevant, regardless of how mathematically complicated it may be.

5.3.1. Classical Nonparametric Tests for Location

Two classical nonparametric tests for the difference in location between two data samples are especially common and useful. These are the *Wilcoxon–Mann–Whitney*, or *rank-sum test* for two independent

samples (analogous to the parametric test in Equation 5.8) and the *Wilcoxon signed-rank test* for paired samples (corresponding to the parametric test in Equation 5.11).

Rank-Sum Test (Unpaired Data)

The Wilcoxon–Mann–Whitney rank-sum test was devised independently in the 1940s by Wilcoxon, and by Mann and Whitney, although in different forms. The notations from both forms of the test are commonly used, and this can be the source of some confusion. However, the fundamental idea behind the test is not difficult to understand. The test is resistant, in the sense that a few wild data values that would completely invalidate the t-test of Equation 5.8 will have little or no influence. It is robust in the sense that, even if all the assumptions required for the t-test in Equation 5.8 are met, the rank-sum test is almost as good (i.e., nearly as powerful). However, unlike the t-test, it is not invertible in a way that can yield a confidence interval computation.

Given two samples of independent (i.e., both serially independent and unpaired) data, the aim is to test for a possible difference in location. It is often erroneously stated that the null hypothesis for this test pertains to the difference between the two sample medians. However, the actual null hypothesis is that the two data samples have been derived from the same distribution, so that a random draw from the population underlying the first sample is equally likely to be larger or smaller than a counterpart from the second (Devine et al., 2018). Both one-sided (the center of one sample is expected in advance to be larger or smaller than the other if the null hypothesis is not true) and two-sided (no prior information on which sample should be larger) alternative hypotheses are possible. Importantly, the effect of serial correlation on the Wilcoxon–Mann–Whitney test is qualitatively similar to its effect on the t-test: the variance of the sampling distribution of the test statistic is inflated by serial correlation in the data, possibly leading to unwarranted rejection of H_0 if the problem is ignored (Yue and Wang, 2002). The same effect occurs in other classical nonparametric tests as well (Von Storch, 1995).

Under the null hypothesis that the two data samples are from the same distribution, the labeling of each data value as belonging to one group or the other is entirely arbitrary. That is, if the two data samples have really been drawn from the same population, each observation is as likely as the next to have been placed in one sample or the other by the process that generated the data. Under the null hypothesis, then, there are not n_1 observations in Sample 1 and n_2 observations in Sample 2, but rather $n = n_1 + n_2$ observations making up a single empirical distribution. The notion that the data labels are arbitrary because all the data have all been drawn from the same distribution under H_0 is known as the principle of *exchangeability*, which also underlies permutation tests, as discussed in Section 5.3.4.

The rank-sum test statistic is a function not of the data values themselves, but of their ranks within the n observations that are pooled under the null hypothesis. It is this feature that makes the underlying distribution(s) of the data irrelevant. Define R_1 as the sum of the ranks held by the members of Sample 1 in this pooled distribution, and R_2 as the sum of the ranks held by the members of Sample 2. Since there are n members of the pooled empirical distribution implied by the null distribution, $R_1 + R_2 = 1+2+3+4 +\ldots+ n = (n)(n+1)/2$. Therefore the mean of this pooled distribution of ranks is $(n+1)/2$, and its variance is the variance of n consecutive integers $= n(n+1)/12$. If the two samples really have been drawn from the same distribution (i.e., if H_0 is true), then R_1 and R_2 will be similar in magnitude if $n_1 = n_2$. Regardless of whether or not the sample sizes are equal, however, R_1/n_1 and R_2/n_2 should be similar in magnitude if the null hypothesis is true.

The null distribution for R_1 and R_2 is obtained in a way that exemplifies the approach of nonparametric tests more generally. If the null hypothesis is true, the observed partitioning of the data into two

groups of size n_1 and n_2 is only one of very many equally likely ways in which the n values could have been split and labeled. Specifically, there are $(n!)/[(n_1!)(n_2!)]$ such equally likely partitions of the data under the null hypothesis. For example, if $n_1 = n_2 = 10$, this number of possible distinct pairs of samples is 184,756. Conceptually, imagine the statistics R_1 and R_2 being computed for each of these 184,756 possible arrangements of the data. It is simply this very large collection of (R_1, R_2) pairs, or, more specifically, the collection of 184,756 scalar test statistics computed from these pairs, that constitutes the null distribution. If the observed test statistic characterizing the closeness of R_1 and R_2 falls comfortably near the middle this large empirical distribution, then that particular partition of the n observations is quite consistent with H_0. If, however, the observed R_1 and R_2 are more different from each other than under most of the other possible partitions of the data, H_0 would be rejected.

It is not actually necessary to compute the test statistic for all $(n!)/[(n_1!)(n_2!)]$ possible arrangements of the data. Rather, the Mann–Whitney U-statistic,

$$U_1 = R_1 - \frac{n_1}{2}(n_1 + 1) \tag{5.23a}$$

or

$$U_2 = R_2 - \frac{n_2}{2}(n_2 + 1), \tag{5.23b}$$

is computed for one or the other of the two Wilcoxon rank-sum statistics, R_1 or R_2. Both U_1 and U_2 carry the same information, since $U_1 + U_2 = n_1 n_2$, although some tables of null distribution probabilities for the rank-sum test evaluate unusualness of only the smaller of U_1 and U_2.

A little thought shows that the rank-sum test is a test for location in a way that is analogous to the conventional t-test. The t-test sums the data and equalizes the effects of different sample sizes by dividing by the sample size. The rank-sum test operates on sums of the ranks of the data, and the effects of possible differences in the sample sizes n_1 and n_2 are equalized using the Mann–Whitney transformation in Equation 5.23. This comparison is developed more fully by Conover and Iman (1981).

For even moderately large values of n_1 and n_2 (both larger than about 10), a simple method for evaluating null distribution probabilities is available. In this case, the null distribution of the Mann–Whitney U-statistic is approximately Gaussian, with

$$\mu_U = \frac{n_1 n_2}{2} \tag{5.24a}$$

and

$$\sigma_U = \left[\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}\right]^{1/2}. \tag{5.24b}$$

Equation 5.24b is valid if all n data values are distinct and is approximately correct when there are few repeated values. If there are many tied values, Equation 5.24b overestimates the sampling variance, and a more accurate estimate is provided by

$$\sigma_U = \left[\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2}{12(n_1 + n_2)(n_1 + n_2 - 1)} \sum_{j=1}^{J} \left(t_j^3 - t_j\right)\right]^{1/2}, \tag{5.25}$$

where J indicates the number of groups of tied values, and t_j indicates the number of members in group j.

For samples too small for application of the Gaussian approximation to the sampling distribution of U, tables of critical values (e.g., Conover, 1999), or an exact recurrence relation (Mann and Whitney, 1947; Mason and Graham, 2002) can be used.

Example 5.9. Evaluation of a Cloud Seeding Experiment Using the Wilcoxon–Mann–Whitney Test
Table 5.7 contains data from a weather modification experiment investigating the effect of cloud seeding on lightning strikes (Baughman et al., 1976). It was suspected in advance that seeding the storms would reduce lightning. The experimental procedure involved randomly seeding or not seeding candidate thunderstorms, and recording a number of characteristics of the lightning, including the counts of strikes presented in Table 5.7. There were $n_1 = 12$ seeded storms, exhibiting an average of 19.25 cloud-to-ground lightning strikes; and $n_2 = 11$ unseeded storms, with an average of 69.45 strikes.

Inspecting the data in Table 5.7, it is apparent that the distribution of lightning counts for the unseeded storms is distinctly non-Gaussian. In particular, the set contains one very large outlier of 358 strikes. We suspect, therefore, that uncritical application of the t-test (Equation 5.8) to test the significance of the difference in the observed mean numbers of lightning strikes could produce misleading results. This is because the single very large value of 358 strikes leads to a sample standard deviation for the unseeded storms of 98.93 strikes, which is larger even than the mean number. This large sample standard deviation would lead us to attribute a very large spread to the assumed t-distributed sampling distribution of the difference of means, so that even rather large values of the test statistic would be judged to be fairly ordinary.

TABLE 5.7 Counts of Cloud-to-Ground Lightning for Experimentally Seeded and Unseeded Storms

Seeded		Unseeded	
Date	Lightning Strikes	Date	Lightning Strikes
7/20/65	49	7/2/65	61
7/21/65	4	7/4/65	33
7/29/65	18	7/4/65	62
8/27/65	26	7/8/65	45
7/6/66	29	8/19/65	0
7/14/66	9	8/19/65	30
7/14/66	16	7/12/66	82
7/14/66	12	8/4/66	10
7/15/66	2	9/7/66	20
7/15/66	22	9/12/66	358
8/29/66	10	7/3/67	63
8/29/66	34		

From Baughman et al. (1976). © American Meteorological Society. Used with permission.

The mechanics of applying the rank-sum test to the data in Table 5.7 are presented in Table 5.8. In the left-hand portion of the table, the 23 data points are pooled and ranked, consistent with the null hypothesis that all the data came from the same population, regardless of the labels S (for seeded) or N (for not seeded). There are two observations of 10 lightning strikes, and as is conventional each has been assigned the average rank $(5+6)/2=5.5$. In the right-hand portion of the table, the data are segregated according to their labels, and the sums of the ranks of the two groups are computed. It is clear from this portion of Table 5.8 that the smaller numbers of strikes tend to be associated with the seeded

TABLE 5.8 Illustration of the Procedure of the Rank-Sum Test Using the Cloud-to-Ground Lightning Data in Table 5.7

Pooled Data			Segregated Data			
Strikes	**Seeded**	**Rank**	**Seeded**	**Rank**	**Seeded**	**Rank**
0	N	1			N	1
2	S	2	S	2		
4	S	3	S	3		
9	S	4	S	4		
10	N	5.5			N	5.5
10	S	5.5	S	5.5		
12	S	7	S	7		
16	S	8	S	8		
18	S	9	S	9		
20	N	10			N	10
22	S	11	S	11		
26	S	12	S	12		
29	S	13	S	13		
30	N	14			N	14
33	N	15			N	15
34	S	16	S	16		
45	N	17			N	17
49	S	18	S	18		
61	N	19			N	19
62	N	20			N	20
63	N	21			N	21
82	N	22			N	22
358	N	23			N	23
		Sums of ranks:	$R_1=108.5$		$R_2=167.5$	

In the left portion of this table, the $n_1+n_2=23$ counts of lightning strikes are pooled and ranked. In the right portion of the table, the observations are segregated according to their labels of seeded (S) or not seeded (N) and the sums of the ranks for the two categories (R_1 and R_2) are computed.

storms, and the larger numbers of strikes tend to be associated with the unseeded storms. These differences are reflected in the differences in the sums of the ranks: R_1 for the seeded storms is 108.5, and R_2 for the unseeded storms is 167.5. The null hypothesis that seeding does not affect the number of lightning strikes can be rejected if this difference between R_1 and R_2 is sufficiently unusual against the backdrop of all possible $(23!)/[(12!)(11!)] = 1{,}352{,}078$ distinct arrangements of these data under H_0.

The Mann–Whitney U-statistic, Equation 5.23, corresponding to the sum of the ranks of the seeded data, is $U_1 = 108.5 - (6)(12+1) = 30.5$. The null distribution of all 1,352,078 possible values of U_1 for this data is closely approximated by the Gaussian distribution having (Equation 5.24) $\mu_U = (12)(11)/2 = 66$ and $\sigma_U = [(12)(11)(12+11+1)/12]^{1/2} = 16.2$. Within this Gaussian distribution, the observed $U_1 = 30.5$ corresponds to a standard Gaussian $z = (30.5 - 66)/16.2 = -2.19$. Table B.1 shows the (one-tailed) p value associated with this z to be 0.014, indicating that approximately 1.4% of the 1,352,078 possible values of U_1 under H_0 are smaller than the observed U_1. Accordingly, H_0 usually would be rejected. ◇

Signed-Rank Test (Paired Data)

There is also a classical nonparametric test, the *Wilcoxon signed-rank test*, analogous to the paired two-sample parametric test of Equation 5.11. As is the case for its parametric counterpart, the signed-rank test takes advantage of positive correlation between the members of data pairs in assessing possible differences in location. In common with the unpaired rank-sum test, the signed-rank test statistic is based on ranks rather than the numerical values of the data. Therefore this test also does not depend on the distribution of the underlying data and is resistant to outliers.

Denote the data pairs (x_i, y_i), for $i = 1, \ldots, n$. The signed-rank test is based on the set of n differences, Δ_i, between the n data pairs. If the null hypothesis is true, and the two data sets represent paired samples from the same population or generating process, roughly equally many of these differences will be positive and negative, and the overall magnitudes of the positive and negative differences should be comparable. The comparability of the positive and negative differences is assessed by ranking them in absolute value. That is the n differences Δ_i are transformed to the series of ranks,

$$T_i = \text{rank}\,|\Delta_i| = \text{rank}\,|x_i - y_i|. \tag{5.26}$$

Data pairs for which $|\Delta_i|$ are equal are assigned the average rank of the tied values of $|\Delta_i|$, and pairs for which $x_i = y_i$ (implying $\Delta_i = 0$) are not included in the subsequent calculations. Denote as n^* the number of pairs for which $x_i \neq y_i$.

If the null hypothesis is true, the labeling of a given data pair as (x_i, y_i) could just as well have been reversed, so that the ith data pair is been just as likely to have been labeled (y_i, x_i). Changing the ordering reverses the sign of Δ_i, but yields the same $|\Delta_i|$. The unique information in the pairings that actually were observed is captured by separately summing the ranks, T_i, corresponding to pairs having positive or negative values of Δ_i, denoting as T either the statistic

$$T^+ = \sum_{\Delta_i > 0} T_i \tag{5.27a}$$

or

$$T^- = \sum_{\Delta_i < 0} T_i, \tag{5.27b}$$

respectively. Tables of null distribution probabilities sometimes require choosing the smaller of Equations 5.27a and 5.27b. However, knowledge of one is sufficient for the other, since $T^+ + T^- = n^*(n^*+1)/2$.

The null distribution of T is arrived at conceptually by considering again that H_0 implies the labeling of one or the other of each datum in a pair as x_i or y_i is arbitrary. Therefore under the null hypothesis there are $2n^*$ equally likely arrangements of the $2n^*$ data values at hand, and the resulting $2n^*$ possible values of T constitute the relevant null distribution. As before, it is not necessary to compute all possible values of the test statistic, since for moderately large n^* (greater than about 20) the null distribution is approximately Gaussian, with parameters

$$\mu_T = \frac{n^*(n^*+1)}{4} \tag{5.28a}$$

and

$$\sigma_T = \left[\frac{n^*(n^*+1)(2n^*+1)}{24}\right]^{1/2}. \tag{5.28b}$$

For smaller samples, tables of critical values for T^+ (e.g., Conover, 1999) can be used. Under the null hypothesis, T ($=T^+$ or T^-) will be close to μ_T because the numbers and magnitudes of the ranks T_i will be comparable for the negative and positive differences Δ_i. If there is a substantial difference between the x and y values in location, most of the large ranks will correspond to either the negative or positive Δ_i's, implying that T will be either very large or very small.

Example 5.10. Comparing Thunderstorm Frequencies Using the Signed Rank Test

The procedure for the Wilcoxon signed-rank test is presented in Table 5.9. Here the paired data are counts of thunderstorms reported in the northeastern United States (x) and the Great Lakes states (y) for the $n = 21$ years 1885–1905. Since the two areas are relatively close geographically, we expect that large-scale flow conducive to thunderstorm formation in one of the regions would be generally conducive in the other region as well. It is thus not surprising that the reported thunderstorm counts in the two regions are substantially positively correlated.

For each year the difference in reported thunderstorm counts, Δ_i, is computed, and the absolute values of these differences are ranked. None of the $\Delta_i = 0$, so $n^* = n = 21$. Years having equal differences, in absolute value, are assigned the average rank (e.g., 1892, 1897, and 1901 have the eighth, ninth, and tenth smallest $|\Delta_i|$, and are all assigned the rank 9). The ranks for the years with positive and negative Δ_i, respectively, are added in the final two columns, yielding $T^+ = 78.5$ and $T^- = 152.5$.

If the null hypothesis that the reported thunderstorm frequencies in the two regions are equal is true, then labeling of counts in a particular year as being Northeastern or Great Lakes is arbitrary and thus so is the sign of each Δ_i. Consider, arbitrarily, the test statistic T as the sum of the ranks for the positive differences, $T^+ = 78.5$. Its unusualness in the context of H_0 is assessed in relation to the $2^{21} = 2{,}097{,}152$ values of T^+ that could result from all the possible permutations of the data under the null hypothesis. This null distribution is closely approximated by the Gaussian distribution having $\mu_T = (21)(22)/4 = 115.5$ and $\sigma_T = [(21)(22)(42+1)/24]^{1/2} = 28.77$. The p value for this test is then obtained by computing the standard Gaussian $z = (78.5 - 115.5)/28.77 = -1.29$. If there is no reason to expect one or the other of the two regions to have had more reported thunderstorms, the test is two-tailed (H_A is simply "not H_0"), so the p value is $\Pr\{z \le -1.29\} + \Pr\{z > +1.29\} = 2\ \Pr\{z \le -1.29\} = 0.197$. The null hypothesis would not be rejected in this case. Note that the same result would be obtained if the test statistic $T^- = 152.5$ had been chosen instead. ◇

TABLE 5.9 Illustration of the Procedure of the Wilcoxon Signed-Rank Test Using Data for Counts of Thunderstorms Reported in the Northeastern United States (x) and the Great Lakes States (y) for the Period 1885–1905

	Paired Data		Differences		Segregated Ranks	
Year	x	y	Δ_i	Rank $\lvert\Delta_i\rvert$	$\Delta_i > 0$	$\Delta_i < 0$
1885	53	70	−17	20		20
1886	54	66	−12	17.5		17.5
1887	48	82	−34	21		21
1888	46	58	−12	17.5		17.5
1889	67	78	−11	16		16
1890	75	78	−3	4.5		4.5
1891	66	76	−10	14.5		14.5
1892	76	70	+6	9	9	
1893	63	73	−10	14.5		14.5
1894	67	59	+8	11.5	11.5	
1895	75	77	−2	2		2
1896	62	65	−3	4.5		4.5
1897	92	86	+6	9	9	
1898	78	81	−3	4.5		4.5
1899	92	96	−4	7		7
1900	74	73	+1	1	1	
1901	91	97	−6	9		9
1902	88	75	+13	19	19	
1903	100	92	+8	11.5	11.5	
1904	99	96	+3	4.5	4.5	
1905	107	98	+9	13	13	
				Sums of ranks:	$T^+ = 78.5$	$T^- = 152.5$

Analogously to the procedure of the rank-sum test (see Table 5.8), the absolute values of the annual differences, $|\Delta_i|$, are ranked and then segregated according to whether Δ_i is positive or negative. The sum of the ranks of the segregated data constitutes the test statistic.
From Brooks and Carruthers (1953).

5.3.2. Mann–Kendall Trend Test

Investigating a possible trend through time of the central tendency of a data series is of interest in the context of a changing underlying climate, among other settings. The usual parametric approach to this kind of question is through conventional least-squares regression analysis (Section 7.2) with a time index as the predictor, and the associated test for the null hypothesis is that a regression slope is zero. The regression slope itself is proportional to the correlation between the time-series variable and the time index.

The *Mann–Kendall trend test* is a popular nonparametric alternative for testing for the presence of a trend, or nonstationarity of the central tendency, of a time series. In a parallel to the alternative parametric regression approach, the Mann–Kendall test arises as a special case of Kendall's τ (Equation 3.35), reflecting a tendency for monotone association between two variables. Accordingly it can accommodate nonlinear as well as linear trends. In the context of examining the possibility of trend underlying a time series x_i, $i = 1, \ldots, n$, the time index i (e.g., the year of observation of each datum) is by definition monotonically increasing, which simplifies the calculations.

The test statistic for the Mann–Kendall trend test is

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}(x_j - x_i) = \sum_{i<j} \text{sgn}(x_j - x_i), \tag{5.29a}$$

where

$$\text{sgn}(\Delta x) = \begin{cases} +1, & \Delta x > 0 \\ 0, & \Delta x = 0 \\ -1, & \Delta x < 0 \end{cases}. \tag{5.29b}$$

That is, the statistic in Equation 5.29a counts the number of all possible data pairs in which the first value is smaller than the second, and subtracts the number of data pairs in which the first is larger than the second. If the data x_i are serially independent and drawn from the same distribution (in particular, if the generating process has the same mean throughout the time series), then the numbers of data pairs for which $\text{sgn}(\Delta x)$ is positive and negative should be nearly equal.

For moderate (n about 10) or larger series lengths, the sampling distribution of the test statistic in Equation 5.29 is approximately Gaussian, and if the null hypothesis of no trend is true this Gaussian null distribution will have zero mean. The variance of this distribution depends on whether all the x's are distinct, or if some are repeated values. If there are no ties, the variance of the sampling distribution of S is

$$\text{Var}(S) = \frac{n(n-1)(2n+5)}{18}, \tag{5.30a}$$

otherwise the variance is

$$\text{Var}(S) = \frac{n(n-1)(2n+5) - \sum_{j=1}^{J} t_j(t_j-1)(2t_j+5)}{18}. \tag{5.30b}$$

Analogously to Equation 5.25, J indicates the number of groups of repeated values, and t_j is the number of repeated values in the jth group. The test p value is evaluated using the standard Gaussian value

$$z = \begin{cases} \dfrac{S-1}{[\text{Var}(S)]^{1/2}}, & S > 0 \\ \dfrac{S+1}{[\text{Var}(S)]^{1/2}}, & S < 0 \end{cases}. \tag{5.31}$$

The mean of S under the null hypothesis is zero; however, the ± 1 in the numerator of Equation 5.31 represents a continuity correction. Alternatively for small sample sizes, Hamed (2009) has suggested use of beta sampling distributions for S.

If all n data values in the series are distinct, the relationship of Equation 5.29a to the value of Kendall's τ characterizing the relationship between the x's and the time index i is

$$S = \binom{n}{2}\tau = \frac{n(n-1)}{2}\tau. \tag{5.32}$$

Example 5.11. Testing for Climate Change Using the Mann–Kendall Test

In Example 5.7, the possibility of a change in the distribution of Ithaca January precipitation between the 1933–57 and 1958–82 periods was examined using a likelihood ratio test. A similar question can be addressed by examining the data in Table A.2 against a null hypothesis of no trend, using the Mann–Kendall Test.

Of the $(50)(49)/2 = 1225$ distinct pairs of individual data values, the earlier of the two is smaller for 580 of the pairs, and it is larger in 638 pairs, yielding $S = -58$. There are $J = 5$ groups of repeated precipitation values in Table A.2, four of which consist of pairs ($t_j = 2$) and one of which is a triple ($t_j = 3$). The subtracted correction term in the numerator of Equation 5.30b is therefore 138, yielding $\mathrm{Var}(S) = [(50)(49)(105) - 138]/18 = 14{,}284$. Using the lower option in Equation 5.31, $z = (-58+1)/(14{,}284)^{1/2} = -0.477$, which is quite ordinary in the context of the null distribution, and associated with a rather large p value. Neglecting the effect of the repeated values, the corresponding Kendall τ characterizing the association between the precipitation data and the year labels in Table A.2 would be, according to Equation 5.32, $\tau = -57/[(50)(49)/2] = -0.0465$, which also indicates a very weak degree of association.

This result is quite consistent with the likelihood ratio test in Example 5.7, which also provided only extremely weak evidence against a null hypothesis of no climate change. However, it is important to keep in mind that the tests do not and cannot prove that no changes are occurring. Because of the logical structure of the hypothesis testing paradigm one can only conclude that any changes in Ithaca January precipitation would be occurring too slowly over the 50 years of data to be discerned against the very considerable year-to-year background variability in precipitation. ◇

Although the Mann–Kendall test can reject a null hypothesis of no trend, it does not return an estimate of the magnitude of any trend so detected. An approach for doing so is provided by the nonparametric regression method of Sen (1968, see also Section 7.7.2), which estimates the linear trend slope as the median of slopes between all $n(n-1)/2$ distinct data pairs. This procedure is thus naturally aligned with Kendall's τ, and with the Mann–Kendall trend test.

When annual time series values are the subject of a trend analysis the Mann–Kendall assumption of serial independence will usually be well met, but other series may exhibit sufficiently strong serial dependence that the results are adversely affected. As is the case in hypothesis testing more generally, positive serial correlation in the data series leads to underestimation of the sampling variance, with the result that the statistic in Equation 5.31will be too large in absolute value, yielding p values that are too small. Lettenmaier (1976) proposed use of an effective-sample-size correction for first-order autoregressive time series that is approximately equivalent to those in Equations 5.12 and 5.13 for the t-test. In particular, assuming first-order autoregressive structure (Section 10.3.1) for the time dependence, $\mathrm{Var}[S]$ would be replaced in Equation 5.31 with $\mathrm{Var}^*[S] = \mathrm{Var}[S](1+r_1)/(1-r_1)$. Yue and Wang (2004) obtained good results with this modification. Here r_1 is the lag-1 autocorrelation in the data

apart from contributions to positive autocorrelation induced by any trend, and so is generally estimated after detrending the data series, often by subtracting a linear regression function (see Section 7.2.1). Alternatively, relaxing the assumption of first-order autoregressive time dependence, Equation 10.41 rather than Equation 10.40 could be used to estimate the variance inflation factor. Cabilio et al. (2013) use the block bootstrap (Section 5.3.5) to estimate the sampling variance of the Mann–Kendall statistic for autocorrelated series.

5.3.3. Introduction to Resampling Tests

Since the advent of inexpensive and fast computing, another approach to nonparametric testing has become practical. This approach is based on the construction of artificial data sets from a given collection of real data, by resampling the observations in a manner consistent with the null hypothesis. Sometimes such methods are also known as *resampling tests, randomization tests*, *rerandomization tests*, or *Monte Carlo tests*. Resampling methods are highly adaptable to different testing situations, and there is considerable scope for the analyst to creatively design new tests to meet particular needs.

The basic idea behind resampling tests is to build up a collection of artificial data batches of the same size as the actual data at hand using a procedure that is consistent with the null hypothesis, and then to compute the test statistic of interest for each artificial batch. The result is as many artificial values of the test statistic as there are artificially generated data batches. Taken together, these reference test statistics constitute an estimated null distribution against which to compare the test statistic computed from the original data.

As a practical matter, a computer is programmed to do the resampling. Fundamental to this process are the uniform [0,1] random number generators described in Section 4.7.1. These algorithms produce streams of numbers that resemble independent values drawn independently from the probability density function $f(u) = 1,\ 0 \le u \le 1$. The synthetic uniform variates are used to draw random samples from the data to be tested.

In general, resampling tests have two very appealing advantages. The first is that no assumptions regarding underlying parametric distributions for the data or the sampling distribution for the test statistic are necessary, because the procedures consist entirely of operations on the data themselves. The second is that any statistic that may be suggested as important by the scientific context of the problem can form the basis of the test, so long as it can be computed from the data. For example, when investigating location (i.e., overall magnitudes) of a sample of data, we are not confined to the conventional tests involving the arithmetic mean or sums of ranks; because it is just as easy to use alternative measures such as the median, the geometric mean, or more exotic statistics if any of these are more meaningful to the problem at hand. The data being tested can be scalar (each data point is one number) or vector-valued (data points are composed of pairs, triples, etc.), as dictated by the structure of each particular problem. Resampling procedures involving vector-valued data can be especially useful when the effects of spatial correlation must be captured by a test, in which case each element in the data vector corresponds to a different location, so that each data vector can be thought of as a "map."

Any computable statistic (i.e., any function of the data) can be used as a test statistic in a resampling test, but not all will be equally good. In particular, some choices may yield tests that are more

powerful than others. Good (2000) suggests the following desirable attributes for candidate test statistics.

1. *Sufficiency*: All the relevant information about the distribution attribute or physical phenomenon of interest contained in the data is also reflected in the chosen statistic. Given a sufficient statistic, the data have nothing additional to say about the question being addressed.
2. *Invariance*: A test statistic should be constructed in a way that the test result does not depend on arbitrary transformations of the data, for example, from °F to °C.
3. *Loss*: The mathematical penalty for discrepancies that is expressed by the test statistic should be consistent with the problem at hand, and the use to which the test result will be put. Often squared-error losses are assumed in parametric tests because of mathematical tractability and connections with the Gaussian distribution, although squared-error loss is disproportionately sensitive to large differences relative to an alternative like absolute error. In a resampling test there is no reason to avoid absolute-error loss or other loss functions if these make more sense in the context of a particular problem.

In addition, Hall and Wilson (1991) point out that better results are obtained when the resampled statistic does not depend on unknown quantities, for example, unknown parameters. Such statistics are called *pivotal*. Equation 5.8 is an example of a pivotal statistic, since the differencing eliminates its dependence on an unknown location parameter, and the division by the estimated standard deviation removes dependence on an unknown scale parameter.

5.3.4. Permutation Tests

Two- (or more-) sample problems can often be approached using *permutation tests*. Early examples of permutation tests in the atmospheric science literature were provided by Mielke Jr. et al. (1981) and Preisendorfer and Barnett (1983). The concept behind permutation tests is not new (Fisher, 1935; Pitman, 1937), but the approach did not become practical until the advent of fast and abundant computing. (Fisher (1935) noted "The arithmetical procedure of such an examination is tedious …").

Permutation tests are a natural generalization of the Wilcoxon–Mann–Whitney test described in Section 5.3.1, and also depend on the principle of exchangeability. Exchangeability implies that, according to the null hypothesis, all the data were drawn from the same distribution. Therefore the labels identifying particular data values as belonging to one sample or another are arbitrary. Under H_0 these data labels are exchangeable.

The key difference between permutation tests generally, and the Wilcoxon–Mann–Whitney test as a special case, is that any test statistic that may be meaningful can be employed, including but certainly not limited to the particular function of the ranks given in Equation 5.23. Among other advantages, lifting restrictions on the mathematical form of possible test statistics expands the range of applicability of permutation tests to vector-valued data. For example, Mielke Jr. et al. (1981) provide a simple illustrative example using two batches of bivariate data ($\boldsymbol{x} = [x, y]$) and the Euclidian distance measure to examine the tendency of the two batches to cluster in the $[x, y]$ plane. Zwiers (1987a) gives an example of a permutation test that uses higher-dimensional multivariate Gaussian variates.

The exchangeability principle leads logically to the construction of the null distribution using samples drawn by computer from a pool of the combined data. As was the case for the Wilcoxon–Mann–Whitney test, if two batches of size n_1 and n_2 are to be compared, the pooled set to be resampled contains $n = n_1 + n_2$ points. However, rather than computing the test statistic using all possible $n!/(n_1!n_2!)$ groupings (i.e., permutations) of the pooled data, the pool is merely sampled some large number (perhaps 10,000) of times.

(An exception can occur when n is small enough for a *full enumeration* of all possible permutations to be practical, and some authors reserve the term "permutation test" for this situation.) For permutation tests the samples are drawn *without replacement*, so that on a given iteration each of the individual n observations is represented once and once only in one or the other of the artificial samples of size n_1 and n_2. In effect, the data labels are randomly permuted for each resample. For each of these pairs of synthetic samples the test statistic is computed, and the resulting distribution of (perhaps 10,000) outcomes forms the null distribution against which the observed test statistic can be compared.

An efficient permutation algorithm can be implemented in the following way. Assume for convenience that $n_1 \geq n_2$. The data values (or vectors) are first arranged into a single array of size $n = n_1 + n_2$. Initialize a reference index $m = n$. The algorithm proceeds by implementing the following steps n_2 times:

- Randomly choose x_i, $i = 1, \ldots, m$; using Equation 4.100 (i.e., randomly draw from the first m array positions).
- Exchange the array positions of (or, equivalently, the indices pointing to) x_i and x_m (i.e., each of the chosen x's will be placed in the bottom section of the n-dimensional array).
- Decrement the reference index by 1 (i.e., $m = m - 1$).

At the end of this process there will be a random selection of the n pooled observations in the first n_1 positions, which can be treated as Sample 1. The remaining n_2 data values at the end of the array can be treated as Sample 2. The scrambled array can be operated upon directly for subsequent random permutations—it is not necessary first to restore the data to their original ordering.

Example 5.12. Two-Sample Permutation Test for a Complicated Statistic
Consider again the lightning data in Table 5.7. Assume that their dispersion is best (from the standpoint of some criterion external to the hypothesis test) characterized by the *L-scale* statistic (Hosking, 1990),

$$\lambda_2 = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |x_i - x_j|. \tag{5.33}$$

Equation 5.33 amounts to half the average difference, in absolute value, between all possible pairs of points in the sample of size n. For a tightly clustered sample of data each term in the sum will be small, and therefore λ_2 will be small. For a data sample that is highly variable, some of the terms in Equation 5.33 will be very large, and λ_2 will be correspondingly large.

To compare sample values of λ_2 from the seeded and unseeded storms in Table 5.7, we probably would use either the ratio or the difference of λ_2 for the two samples. A resampling test procedure provides the freedom to choose the one (or some other) making more sense for the problem at hand. Suppose the most relevant test statistic is the ratio $[\lambda_2(\text{seeded})]/[\lambda_2(\text{unseeded})]$. Under the null hypothesis that the two samples have the same L-scale, this statistic should be near one. If the seeded storms are more variable with respect to the numbers of lightning strikes, the ratio statistic should be greater than one. If the seeded storms are less variable, the statistic should be less than one. The ratio of L-scales has been chosen for this example arbitrarily, to illustrate that any computable function of the data can be used as the basis of a permutation test, regardless of how unusual or complicated it may be.

The null distribution of the test statistic is built up by sampling some (say 10,000) of the $23!/(12!\,11!) = 1{,}352{,}078$ distinct partitions, or permutations, of the $n = 23$ data points into two batches of size $n_1 = 12$ and $n_2 = 11$. For each partition, λ_2 is computed according to Equation 5.33 for each of the two

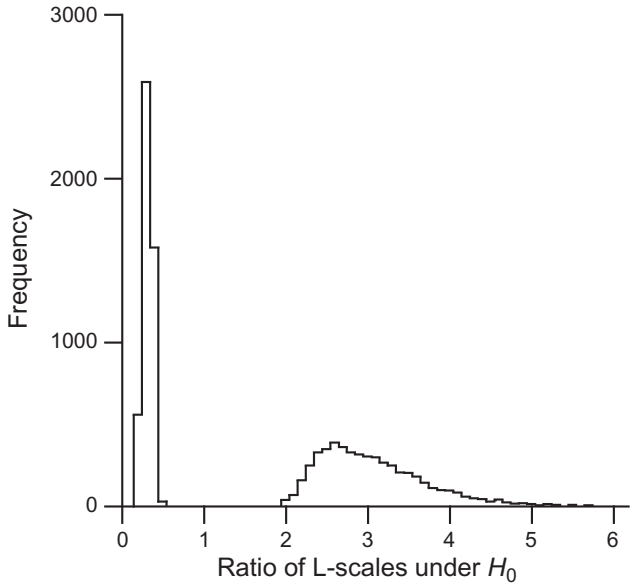

FIGURE 5.7 Histogram for the null distribution of the ratio of the *L*-scales for lightning counts of seeded versus unseeded storms in Table 5.7. The observed ratio of 0.188 is smaller than all but 49 of the 10,000 permutation realizations of the ratio, which provides very strong evidence that the lightning production by seeded storms was less variable than by unseeded storms. This null distribution is bimodal because the one outlier (353 strikes on 9/12/66) produces a very large *L*-scale in whichever of the two partitions it has been randomly assigned.

synthetic samples, and their ratio (with the value for the $n_1 = 12$ batch in the numerator) is computed and stored. The observed value of the ratio of the *L*-scales, 0.188, is then evaluated with respect to this empirically generated null distribution.

Figure 5.7 shows a histogram for the null distribution of the ratios of the *L*-scales constructed from 10,000 permutations of the original data. The observed value of 0.188 is smaller than all except 49 of these 10,000 values, which would lead to the null hypothesis being soundly rejected. Depending on whether a one-sided or two-sided test would be appropriate on the basis of prior external information, the p value would be either 0.0049 or 0.0098, respectively. Notice that this null distribution has the unusual feature of being bimodal, having two humps. This characteristic results from the large outlier in Table 5.7, 358 lightning strikes on 9/12/66, producing a very large *L*-scale in whichever partition it has been assigned. Partitions for which this observation has been assigned to the unseeded (n_2) group are in the left hump, and those for which the outlier has been assigned to the seeded (n_1) group are in the right hump.

The conventional test for differences in dispersion involves the ratio of sample variances, the null distribution for which would be the *F distribution* if the two underlying data samples are both Gaussian, but the *F* test is not robust to violations of the Gaussian assumption. Computing a permutation test on the basis the variance ratio s^2(seeded)/s^2(unseeded) would be as easy if not easier than computing the *L*-scale ratio permutation test, and in that case the permutation null distribution would also be bimodal (and probably exhibit larger variance because the sample variance is not resistant to outliers). It is likely that the results of such a test would be similar to those for the permutation test based on the *L*-scale ratio. However, the

corresponding parametric test, which would examine the observed $s^2(\text{seeded})/s^2(\text{unseeded}) = 0.0189$ in relation to the F distribution rather than the corresponding resampling distribution would likely be misleading, since the F distribution would not resemble the counterpart of Figure 5.7 for resampled variance ratios. ◇

Permutation tests can also be applied in paired-data situations, analogously to the parametric paired t-test described in Section 5.2.3, even though the pairing induces reduction to a one-sample test. Indeed, Fisher's (1935) original description of the permutation method was applied in exactly this setting. Again consistent with the null hypothesis that the labels for each pair as belonging to one group or the other are arbitrary, the permutations are achieved by randomly (with probability 1/2) switching those labels for each pair. After each randomization of labels has been implemented, the test statistic of interest (which may be, but is not limited to, the difference of means) is computed. The procedure is repeated a large number of times, and the resulting collection of randomized test statistics provides the needed null distribution.

5.3.5. The Bootstrap

Permutation schemes are very useful in multiple-sample settings where the exchangeability principle applies. In one-sample settings permutation procedures are useless because there is nothing to permute: there is only one way to resample a single data batch with replacement, and that is to replicate the original sample by choosing each of the original n data values exactly once. When the exchangeability assumption cannot be supported, the justification for pooling multiple samples before permutation disappears, because the null hypothesis no longer implies that all data, regardless of their labels, were drawn from the same population.

In either of these situations an alternative computer-intensive resampling procedure called the *bootstrap* is available. The bootstrap is a newer idea than permutation, dating from Efron (1979). The idea behind the bootstrap is known as the *plug-in principle*, under which we estimate any function of the underlying (population) distribution by using (plugging into) the same function, but using the empirical distribution, which puts probability $1/n$ on each of the n observed data values. Put another way, the idea behind the bootstrap is to treat a finite sample at hand as similarly as possible to the unknown distribution from which it was drawn. This perspective leads to resampling *with replacement*, since an observation of a particular value from an underlying distribution does not preclude subsequent observation of an equal data value. In general the bootstrap is less accurate than the permutation approach when permutation is appropriate, but can be used in instances where permutation cannot. Fuller exposition of the bootstrap than is possible here can be found in Efron and Gong (1983), Efron and Tibshirani (1993), and Leger et al. (1992), among others. Some examples of its use in climatology are given in Downton and Katz (1993) and Mason and Mimmack (1992).

Resampling with replacement is the primary distinction in terms of mechanics between the bootstrap and the permutation approach, where the resampling is done without replacement. Conceptually, the resampling process is equivalent to writing each of the n data values on separate slips of paper and putting all n slips of paper in a hat. To construct one bootstrap sample, n slips of paper are drawn from the hat and their data values recorded, but each slip is put back in the hat and mixed (this is the meaning of "with replacement") before the next slip is drawn. Generally some of the original data values will be drawn into a given bootstrap sample multiple times, and others will not be drawn at all. On average, the fraction of the original n values that are omitted from a bootstrap sample is $(1–1/n)^n$, which for large

n is about $1/e \approx 0.368$. If n is small enough, all possible distinct bootstrap samples can be fully enumerated. However, the number of possible bootstrap samples increases rapidly with n, according to

$$\#\text{possible bootstrap samples} = \binom{2n-1}{n}. \tag{5.34}$$

For example, for $n = 5$, Equation 5.34 yields only 126 possible bootstrap samples, but the number increases to 92,378 for $n = 10$, and to nearly 69×10^9 for $n = 20$.

In practice, we usually program a computer to perform the resampling, using Equation 4.100 in conjunction with a uniform random number generator (Section 4.7.1). This process is repeated a large number, perhaps $n_B = 10{,}000$ times, yielding n_B bootstrap samples, each of size n. The statistic of interest is computed for each of these n_B bootstrap samples. The resulting frequency distribution is then used to approximate the true sampling distribution of that statistic.

Example 5.13. One-Sample Bootstrap: Confidence Interval for a Complicated Statistic
The bootstrap is often used in one-sample settings to estimate confidence intervals around observed values of a test statistic. Because we do not need to know the analytical form of its sampling distribution, the procedure can be applied to any test statistic, regardless of how mathematically complicated it may be. To take a hypothetical yet not especially complicated example, consider the standard deviation of the logarithms, $s_{\ln(x)}$, of the 1933–82 Ithaca January precipitation data in Table A.2 of Appendix A. This statistic has been chosen for this example arbitrarily, to illustrate that any computable sample statistic can be bootstrapped. Here scalar data are used, but Efron and Gong (1983) illustrate the bootstrap using vector-valued (paired) data, for which a confidence interval around the sample Pearson correlation coefficient was estimated.

The value of $s_{\ln(x)}$ computed from the $n = 50$ data values is 0.537, but in order to make inferences about the true value, we need to know or estimate its sampling distribution. Figure 5.8 shows a histogram of the sample standard deviations computed from $n_B = 10{,}000$ bootstrap samples of size $n = 50$ from the logarithms of this data set. This empirical distribution approximates the sampling distribution of $s_{\ln(x)}$ for these data.

Confidence regions for $s_{\ln(x)}$ are most easily approached using the straightforward and intuitive *percentile method* (Efron and Gong, 1983; Efron and Tibshirani, 1993). To form a $(1-\alpha)\cdot 100\%$ confidence interval using this approach, we simply find the values of the estimates defining largest and smallest $n_B \cdot \alpha/2$ of the n_B bootstrap estimates. These values also define the central $n_B \cdot (1-\alpha)$ of the estimates, which is the region of interest. In Figure 5.8, for example, the estimated 95% confidence interval for $s_{\ln(x)}$ using the percentile method is between 0.410 and 0.648. ◇

The previous example illustrates use of the bootstrap in a one-sample setting where permutations are not possible. Bootstrapping is also applicable in multiple-sample situations where the data labels are not exchangeable, so that pooling and permutation of data is not consistent with the null hypothesis. These kinds of data can still be resampled with replacement using the bootstrap, while maintaining the separation of samples having meaningfully different labels.

To illustrate, consider investigating differences of means using the test statistic in Equation 5.5. Depending on the nature of the underlying data and the available sample sizes, we might not trust the Gaussian approximation to the sampling distribution of this statistic, in which case an attractive alternative would be to approximate it through resampling. If the data labels were exchangeable, it would be natural to compute a pooled estimate of the variance and use Equation 5.9 as the test statistic, estimating its sampling distribution through a permutation procedure because both the means and variances would

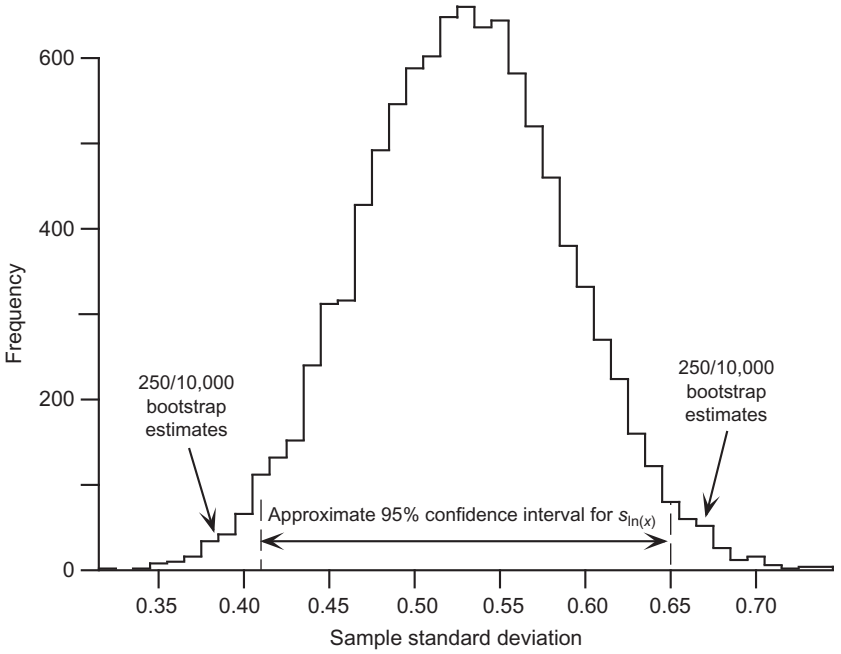

FIGURE 5.8 Histogram of $n_B = 10{,}000$ bootstrap estimates of the standard deviation of the logarithms of the 1933–82 Ithaca January precipitation data in Table A.2. The sample standard deviation computed directly from the data is 0.537. The 95% confidence interval for the statistic, as estimated using the percentile method, is also shown.

be equal under the null hypothesis. On the other hand, if the null hypothesis did not include equality of the variances, Equation 5.8 would be the correct test statistic, but it would not be appropriate to estimate its sampling distribution through permutation, because in this case the data labels would be meaningful, even under H_0. However, the two samples could be separately resampled with replacement (i.e., bootstrapped individually) to build up a bootstrap approximation to the sampling distribution of Equation 5.8. We would need to be careful in generating the bootstrap distribution for Equation 5.8 to construct the bootstrapped quantities consistent with the null hypothesis of equality of means. In particular, we could not bootstrap the raw data directly, because they have different means (whereas the two population means are equal according to the null hypothesis). One option would be to center each of the data batches at the overall mean (which would equal the estimate of the common, pooled mean, according to the plug-in principle). A more straightforward approach would be to estimate the sampling distribution of the test statistic directly, and then exploit the duality between hypothesis tests and confidence intervals to address the null hypothesis. This second approach is illustrated in the following example.

Example 5.14. Two-Sample Bootstrap Test for a Complicated Statistic

Consider again the situation in Example 5.12, in which we were interested in the ratio of *L*-scales (Equation 5.33) for the numbers of lightning strikes in seeded vs. unseeded storms in Table 5.7. The permutation test in Example 5.12 was based on the assumption that, under the null hypothesis, *all* aspects of the distribution of lightning strikes were the same for the seeded and unseeded storms. But pooling and permutation would not be appropriate if we wish to allow for the possibility that, even if the *L*-spread does not depend on seeding, other aspects of the distributions (e.g., the median numbers of lightning strikes) may be different.

Less restrictive null hypotheses can be accommodated by separately and repeatedly bootstrapping the $n_1 = 12$ seeded and $n_2 = 11$ unseeded lightning counts, and forming $n_B = 10{,}000$ samples of the ratio of one bootstrap realization of each, yielding bootstrap realizations of the test statistic λ_2(seeded)/λ_2(unseeded). The result, shown in Figure 5.9 is a bootstrap estimate of the sampling distribution of this ratio for the data at hand. Its center is near the observed ratio of 0.188, which is the $q_{.4835}$ quantile of this bootstrap distribution. Even though this is not the bootstrap null distribution—which would be the sampling distribution if λ_2(seeded)/λ_2(unseeded) = 1—it can be used to evaluate the null hypothesis by examining the unusualness of λ_2(seeded)/λ_2(unseeded) = 1 with respect to this sampling distribution. The horizontal gray arrow indicates the 95% confidence interval for the *L*-scale ratio, estimated using the percentile method, which ranges from 0.08 to 0.75. Since this interval does not include 1, H_0 would be rejected at the 5% level (two-sided). The bootstrap *L*-scale ratios are >1 for only 33 of the $n_B = 10{,}000$ resamples, so the actual p value would be estimated as either 0.0033 (one-sided) or 0.0066 (two-sided), and so H_0 could be rejected at the 1% level as well. ◇

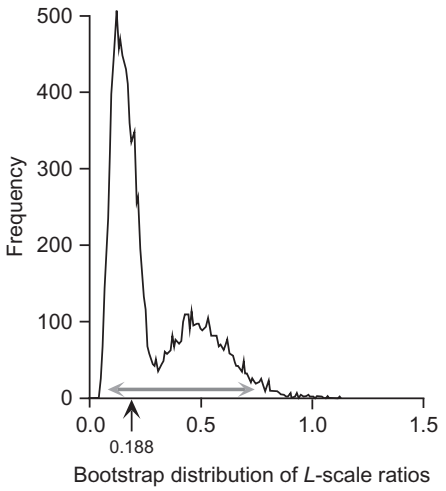

FIGURE 5.9 Bootstrap distribution for the ratio of *L*-scales for lightning strikes in seeded and unseeded storms, Table 5.7. The ratio is >1 for only 33 of 10,000 bootstrap samples, indicating that a null hypothesis of equal *L*-scales would be rejected. Also shown (gray arrows) is the 95% confidence interval for the ratio computed using the percentile method, which ranges from 0.08 to 0.75.

The percentile method is straightforward and easy to use, and gives generally good results in large-sample situations where the statistic being considered is unbiased and its sampling distribution is symmetrical or nearly so. For more moderate sample sizes and more general statistics, a better and more sophisticated method of bootstrap confidence interval construction is available, called bias-corrected and accelerated, or *BC$_a$ intervals* (Efron, 1987; Efron and Tibshirani, 1993). BC$_a$ intervals are more accurate than bootstrap confidence intervals based on the percentile method in the sense that the fraction of $(1 - \alpha)\cdot 100\%$ confidence intervals including the true value of the underlying statistic will be closer to $(1 - \alpha)$ for BC$_a$ intervals.

In common with the percentile method, BC$_a$ confidence intervals are based on quantiles of the bootstrap distribution. Denote the sample estimate of the statistic of interest, around which a confidence interval is to be constructed, as S. In Example 5.13, $S = s_{\ln(x)}$, the sample standard deviation of the

log-transformed data. Denote the ith order statistic of the n_B bootstrap resamples of S as $S^*_{(i)}$. The percentile method estimates the lower and upper bounds of the $(1-\alpha)\cdot 100\%$ confidence interval as $S^*_{(L)}$ and $S^*_{(U)}$, where $L = n_B \cdot \alpha_L = n_B \cdot \alpha/2$ and $U = n_B \cdot \alpha_U = n_B \cdot (1-\alpha/2)$. $\mathrm{BC_a}$ confidence intervals are computed similarly, except that different quantiles of the bootstrap distribution are chosen, typically yielding $\alpha_L \neq \alpha/2$ and $\alpha_U \neq (1-\alpha/2)$. Instead, the estimated confidence interval limits are based on

$$\alpha_L = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z(\alpha/2)}{1-\hat{a}(\hat{z}_0 + z(\alpha/2))}\right] \tag{5.35a}$$

and

$$\alpha_U = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z(1-\alpha/2)}{1-\hat{a}(\hat{z}_0 + z(1-\alpha/2))}\right]. \tag{5.35b}$$

Here $\Phi[\cdot]$ denotes the CDF of the standard Gaussian distribution, the parameter $\hat{z}_0$ is the bias correction, and the parameter $\hat{a}$ is the "acceleration." For $\hat{z}_0 = \hat{a} = 0$, Equations 5.35a and 5.35b reduce to the percentile method since, for example, $\Phi[z(\alpha/2)] = \alpha/2$.

The bias correction parameter $\hat{z}_0$ reflects median bias of the bootstrap distribution, or the difference between the estimated statistic S and the median of the bootstrap distribution, in units of Gaussian standard deviations. It is estimated using

$$\hat{z}_0 = \Phi^{-1}\left[\frac{\#\{S_i^* < S\}}{n_B}\right], \tag{5.36}$$

where the numerator inside the square brackets denotes the number of bootstrap estimates S_i^* that are smaller than the estimate computed using each of the n data values exactly once, S. Equation 5.36 is thus the normal quantile transform (Equation 4.57) of the relative frequency of bootstrap samples smaller than S. If exactly half of the S^* estimates are smaller than S, then the median bias is zero, because $\hat{z}_0 = \Phi^{-1}[1/2] = 0$.

The acceleration parameter $\hat{a}$ is conventionally computed using a statistic related to the *jackknife* estimate of the skewness of the sampling distribution of S. The jackknife (e.g., Efron and Hastie, 2016; Efron and Tibshirani, 1993) is a relatively early and therefore less computationally intensive resampling algorithm, in which the statistic S of interest is recomputed n times, each time omitting one of the n data values that were used to compute the original S. Denote the ith jackknife estimate of the statistic S, which has been computed after removing the ith data value, as S_{-i}; and denote the average of these n jackknife estimates as $\overline{S}_{\mathrm{jack}} = (1/n)\Sigma_i S_{-i}$. The conventional estimate of the acceleration is then

$$\hat{a} = \frac{-\sum_{i=1}^{n}\left(S_{-i} - \overline{S}_{\mathrm{jack}}\right)^3}{6\left[\sum_{i=1}^{n}\left(S_{-i} - \overline{S}_{\mathrm{jack}}\right)^2\right]^{3/2}}. \tag{5.37}$$

Typical magnitudes for both $\hat{z}_0$ and $\hat{a}$ are on the order of $n^{-1/2}$ (Efron, 1987), so that as the sample size increases the $\mathrm{BC_a}$ and percentile methods yield increasingly similar results.

Example 5.15. A BC_a Confidence Interval: Example 5.13 Revisited
In Example 5.13, a 95% confidence interval for the standard deviation of the log-transformed Ithaca January precipitation data from Table A.2 was computed using the straightforward percentile method. A 95% BC_a confidence interval is expected to be more accurate, although it is more difficult to compute. The same $n_B = 10{,}000$ bootstrap samples of $S = s_{\ln(x)}$ are used in each case, but the difference will be that for the BC_a confidence interval Equation 5.35 will be used to compute α_L and α_U, which will differ from $\alpha/2 = 0.025$ and $(1 - \alpha/2) = 0.975$, respectively.

The particular $n_B = 10{,}000$-member bootstrap distribution computed for these two examples contained 5552 bootstrap samples with $S_i^* < S = 0.537$. Using Equation 5.36, the bias correction is therefore estimated as $\hat{z}_0 = \Phi^{-1}[0.5552] = 0.14$. The acceleration in Equation 5.37 requires computation of the $n = 50$ jackknife values of the sample statistic, S_{-i}. The first three of these (i.e., standard deviations of the batches of 49 log-transformed precipitation omitting in turn the data for 1933, 1934, and 1935) are $S_{-1} = 0.505$, $S_{-2} = 0.514$, and $S_{-3} = 0.516$. The average of the 50 jackknifed values is 0.537, the sum of squared deviations of the jackknife values from this average is 0.004119, and the sum of cubed deviations is -8.285×10^{-5}. Substituting these values into Equation 5.37 yields $\hat{a} = (8.285 \times 10^{-5})/[6(0.004119)^{3/2}] = 0.052$. Using these values with $z(0.025) = -1.96$ in Equation 5.35a yields $\alpha_L = \Phi[-1.52] = 0.0643$, and similarly Equation 5.35b with $z(0.975) = +1.96$ yields $\alpha_U = \Phi[2.50] = 0.9938$.

The lower endpoint for the BC_a estimate of the 95% confidence interval around $S = 0.537$ is thus the bootstrap quantile corresponding to $L = n_B\,\alpha_L = (10{,}000)(0.0643)$, or $S^*_{(643)} = 0.437$, and the upper endpoint is the bootstrap quantile corresponding to $U = n_B\,\alpha_U = (10{,}000)(0.9938)$, or $S^*_{(9938)} = 0.681$. This interval [0.437, 0.681] is slightly wider and shifted upward, relative to the interval [0.410, 0.648] computed in Example 5.13. ◇

Parametric Bootstrap

Use of the bootstrap relies on having enough data for the underlying population or generating process to have been reasonably well sampled. A small sample may exhibit too little variability for bootstrap samples drawn from it to adequately represent the variability of the generating process that produced the data. For such relatively small data sets, an improvement over ordinary nonparametric bootstrapping may be provided by the *parametric bootstrap*, if a good parametric model can be identified for the data. The parametric bootstrap operates in the same way as the ordinary, nonparametric, bootstrap except that each of the n_B the bootstrap samples is a synthetic sample of size n that has been generated (Section 4.7) through random draws from a parametric distribution that has been fit to the size-n data sample. Kysely (2008) has compared the performance of parametric and nonparametric bootstrap confidence intervals in settings simulating extreme-value distributions for precipitation and temperature, and reports better results for the parametric bootstrap when $n \leq 40$.

Bootstrapping Autocorrelated Data

It is important to note that direct use of either bootstrap or permutation methods only makes sense when the underlying data to be resampled are independent. If the data are mutually correlated (exhibiting, e.g., time correlation or persistence) the results of these approaches will be misleading (Zwiers, 1987a, 1990), in the same way and for the same reason that autocorrelation adversely affects parametric tests. The random sampling used in either permutation or the bootstrap shuffles the original data, destroying the ordering that produces the autocorrelation.

One approach to respecting data correlation in bootstrapping, called *nearest-neighbor bootstrapping* (Lall and Sharma, 1996), accommodates serial correlation by resampling according to probabilities that depend on similarity to the previous few data points, rather the unvarying $1/n$ implied by the independence assumption. Essentially, the nearest-neighbor bootstrap resamples from relatively close analogs rather than from the full data set. The closeness of the analogs can be defined for both scalar and vector (multivariate) data.

The *sieve bootstrap* (Bickel and Bühlmann, 1999; Bühlmann, 1997; Choi and Hall, 2000) extends the idea of the parametric bootstrap to time-series data. In this method, the time dependence of a data series is modeled with a fitted autoregressive process (see Section 10.3), the complexity of which depends on the underlying data series. A large number n_B synthetic time series of length n are then generated (Section 10.3.7) from the fitted model, from which the desired inferences are then computed. Bühlmann (2002) concludes that the sieve bootstrap is the preferred method for resampling time-series data if it is known that the underlying data-generating process is a linear time-series process.

Time-series data are most commonly bootstrapped using the modification known as the *moving-blocks bootstrap* (Efron and Tibshirani, 1993; Künsch, 1989; Lahiri, 2003; Leger et al., 1992; Wilks, 1997b). Instead of resampling individual data values or data vectors, contiguous sequences of length L are resampled in order to build up a synthetic sample of size n. Figure 5.10 illustrates resampling a data series of length $n = 12$ by choosing $b = 3$ contiguous blocks of length $L = 4$, with replacement. The resampling works in the same way as the ordinary bootstrap, except that instead of resampling from a collection of n individual, independent values, the objects to be resampled with replacement are all the $n - L + 1$ contiguous subseries of length L.

The idea behind the moving-blocks bootstrap is to choose the blocklength L to be large enough for data values separated by this time period or more to be nearly independent. A good choice for the blocklength b is important to the success of the procedure, but a single best method for defining it has not emerged. Theoretical work (Bühlmann, 2002) suggests choosing $b \approx n^{1/3}$, although the optimality of this choice for the smaller samples encountered in practice is not clear. An alternative approach is to choose the blocklength from the middle of a range in which the results (e.g., a confidence interval width) change little, which is called the *minimum volatility method* (Politis et al., 1999). Intuitively, in addition to the blocklength increasing with increasing sample size n, it should also increase as the strength of dependence in the underlying time series increases, and should tend toward $b = 1$ (i.e., ordinary bootstrapping) as the time dependence progressively weakens (Braverman et al., 2011; Wilks, 1997b). If it can be assumed that the data follow a first-order autoregressive process (Equation 10.16), good results are achieved by choosing the blocklength according to the implicit equation (Wilks, 1997b)

$$L = (n - L + 1)^{(2/3)(1-n'/n)}, \tag{5.38}$$

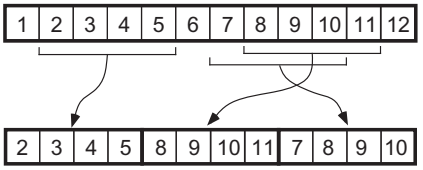

FIGURE 5.10 Schematic illustration of the moving-blocks bootstrap. Beginning with a time series of length $n = 12$ (above), $b = 3$ blocks of length $L = 4$ are drawn with replacement. The resulting time series (below) is one of $(n - L + 1)^b = 729$ equally likely bootstrap samples. *From Wilks (1997b).*

where n' is defined by Equation 5.12. Regardless of how L has been chosen, if n/L is not an integer, the number of blocks b is chosen as the next larger integer, but the final block in each bootstrap sample is truncated so each bootstrap series has length n.

One weakness of the block-bootstrap procedure just described is that data values contained in blocks originating near the beginning and end of the original data series will be undersampled. This undersampling can be avoided through use of the *circular block bootstrap* (Politis and Romano, 1992). Here the first $b - 1$ values are appended to the end of the underlying data series, so that n blocks of length b can be resampled with equal probability.

5.4. MULTIPLICITY AND "FIELD SIGNIFICANCE"

Special problems occur when the results of multiple statistical tests must be evaluated simultaneously, which is known as the problem of test *multiplicity*. The multiplicity problem arises in many settings, but in meteorology and climatology it is most usually confronted when analyses involving atmospheric fields must be performed. Accordingly the multiplicity problem is sometimes conceptualized in terms of assessing *"field significance"* (Livezey and Chen, 1983). In this context the term atmospheric field often connotes a two-dimensional (horizontal) array of geographical locations at which data are available. It may be, for example, that two atmospheric models (one, perhaps, reflecting an increase of the atmospheric carbon dioxide concentration) both produce realizations of surface temperature at each of many gridpoints, and the question is whether the average temperatures portrayed by the two models are significantly different.

In principle, multivariate methods of the kind described in Section 12.5 would be preferred for this kind of problem, but often in practice the data are insufficient to implement them effectively if at all. Accordingly, statistical inference for these kinds of data is often approached by first conducting individual tests at each of the gridpoints, computing perhaps a collection of two-sample t-tests (Equation 5.8). If appropriate, a correction for serial correlation of the underlying data such as that in Equation 5.13 would be part of each of these local tests. Having conducted the local tests, however, it still remains to evaluate, collectively, the overall significance of the differences between the fields, or the field significance. This evaluation of overall significance is sometimes called determination of *global significance* or *pattern significance*. There are two major difficulties associated with this step. These derive from the problems of test multiplicity and from spatial correlation of the underlying data.

5.4.1. The Multiplicity Problem for Independent Tests

Consider first the problem of evaluating the collective significance of N independent hypothesis tests. If all their null hypotheses are true then the probability of falsely rejecting any one of them, picked at random, will be α. But we naturally tend to focus attention on the tests with the smallest p values, which would be a distinctly nonrandom sample of these N tests, and the bias deriving from this mental process needs to be accounted for in the analysis procedure.

The issue has been amusingly framed by Taleb (2001) in terms of the so-called *infinite monkeys theorem*. This is actually an allegory and not a theorem at all. If we could somehow put an infinite number of monkeys in front of keyboards and allow them to type random characters, it is virtually certain that one would eventually reproduce the *Iliad*. But it would not be reasonable to conclude that this particular monkey is special, in the sense, for example, that it would have a higher chance than any of the others of subsequently typing the *Odyssey*. Given the limitless number of monkeys typing, the fact that

one has produced something recognizable does not provide strong evidence against a null hypothesis that this is just an ordinary monkey, whose future literary output will be as incoherent as that of any other monkey. In the realistic and less whimsical counterparts of this kind of setting, we must be careful to guard against *survivorship bias*, or focusing attention on the few instances or individuals surviving some test, and regarding them as typical or representative. That is, when we cherry-pick the (nominally) most significant results from a collection of tests, we must hold them to a higher standard (e.g., require smaller p values) than would be appropriate for any single test, or for a randomly chosen test from the same collection.

It has been conventional in the atmospheric sciences since publication of the Livezey and Chen (1983) paper to frame the multiple testing problem as a meta-test, where the data being tested are the results of N individual or "local" tests, and the "global" null hypothesis is that all the local null hypotheses are true. The Livezey–Chen approach was to compute the number of local tests exhibiting significant results, sometimes called the *counting norm* (Zwiers, 1987a), necessary to reject the global null hypothesis at a level α_{global}. Usually this global test level is chosen to be equal to the local test level, α. If there are $N = 20$ independent tests it might be naively supposed that, since 5% of 20 is 1, finding that any one of the 20 tests indicated a significant difference at the 5% level would be grounds for declaring the two fields to be significantly different, and that by extension, three significant tests out of 20 would be very strong evidence.

Although this reasoning sounds superficially plausible, because of survivorship bias it is really only even approximately true if there are very many, perhaps 1000, tests, and these tests are statistically independent (Livezey and Chen, 1983; Von Storch, 1982). Recall that declaring a significant difference at the 5% level means that, if the null hypothesis is true and there are really no significant differences, there is a probability no larger than 0.05 that evidence against H_0 as strong as or stronger than observed would have appeared by chance. For a single test conducted at the 5% level, the situation is analogous to rolling a 20-sided die (Figure 5.11), and observing that the side with the "20" on it has come up. However,

FIGURE 5.11 An icosahedron, or regular 20-sided geometrical solid, with distinct integer labels on each face. Participants in role-playing board games know this object as a "d20."

conducting $N = 20$ independent tests is probabilistically equivalent to rolling this die 20 times: there is a substantially higher chance than 5% that the side with "20" on it comes up at least once in 20 throws, and it is this latter situation that is analogous to the evaluation of the results from $N = 20$ independent hypothesis tests. If the die is rolled hundreds of times, it is virtually certain that the outcome that is rare for a single throw will occur many times.

Thinking about this analogy between multiple tests and multiple rolls of the 20-sided die suggests that we can quantitatively analyze the multiplicity problem for independent tests in the context of the binomial distribution and conduct a global hypothesis test based on the number of the N individual independent hypothesis tests that are nominally significant. Recall that the binomial distribution specifies probabilities for X successes out of N independent trials if the probability of success on any one trial is p. In the testing multiplicity context, X is the number of significant individual tests out of N tests conducted, and p is the level of the local tests.

Example 5.16. Illustration of the Livezey–Chen Approach for Independent Tests
In the hypothetical example just discussed, there are $N = 20$ independent tests, and $\alpha = 0.05$ is the level of each of these tests. Suppose the local tests pertain to inferences about means at N spatial locations, and $x = 3$ of the 20 tests have yielded significant differences. The question of whether the differences are (collectively) significant at the $N = 20$ gridpoints thus reduces to evaluating $\Pr\{X \geq 3\}$, given that the null distribution for the number of significant tests is binomial with $N = 20$ and $p = 0.05$. Using the binomial probability distribution function (Equation 4.1) with these two parameters, we find $\Pr\{X = 0\} = 0.358$, $\Pr\{X = 1\} = 0.377$, and $\Pr\{X = 2\} = 0.189$. Thus $\Pr\{X \geq 3\} = 1 - \Pr\{X < 3\} = 0.076$, and the null hypothesis that the two mean fields, as represented by the $N = 20$ gridpoints, are equal would not be rejected at the $\alpha_{\text{global}} \cdot 100\% = 5\%$ level. Since $\Pr\{X = 3\} = 0.060$, finding four or more significant local tests would result in a declaration of global (field) significance, at the (global) 5% level.

Even if there are no real differences, the chances of finding at least one significant test result out of 20 are almost 2 out of 3, since $\Pr\{X = 0\} = 0.358$. Until we are aware of and accustomed to the issue of test multiplicity, results such as these seem counterintuitive. ◇

Livezey and Chen (1983) pointed out some instances in the literature of the atmospheric sciences where a lack of awareness of the multiplicity problem had led to conclusions that were not supported by the data. Wilks (2016b) found that this unfortunate situation has not improved in the intervening years, counting more than one-third of a representative sample of papers published in the *Journal of Climate* that ignored the problem of test multiplicity, while only 1% of these papers had addressed it (the remaining papers contained no multiple statistical inferences). Such results are usually displayed on maps that locate positions of nominally significant tests with black dots, leading to the visual impression of stippling over portions of the domain. By not addressing statistical test multiplicity and thereby producing unwarranted rejections of valid null hypotheses, this naive *stippling method* overstates research results even if the multiple tests are mutually independent (which is rarely the case). The result is that this naive procedure may lead researchers to overinterpret their data, and possibly to construct fanciful rationalizations for nonreproducible sampling fluctuations (Wilks, 2016b).

5.4.2. Field Significance and the False Discovery Rate

The Livezey–Chen approach to addressing test multiplicity by counting numbers of rejected null hypotheses is straightforward and attractive in its simplicity, but suffers from two important drawbacks.

The first is that the binary view of the local test results can reduce the global test power (i.e., can yield poor sensitivity for rejecting false null hypotheses). Local null hypotheses that are very strongly rejected (local p values that are very much smaller than α) carry no greater weight in the global test than do local tests for which the p values are only slightly smaller than α. That is, no credit is given for rejecting one or more local null hypotheses with near certainty when evaluating the plausibility of the global null hypothesis that all local null hypotheses are valid. The poor power of Livezey–Chen test is especially problematic if only a small fraction of the N local null hypotheses are not valid.

The second major problem is that the Livezey–Chen test is very sensitive to the effects of positive correlations among the data underlying the different tests, and therefore to positive correlation among the local test results. This situation occurs commonly when the local tests pertain to data at a collection of correlated spatial locations. The issue of spatially correlated tests will be taken up more fully in Section 5.4.3. The naive stippling approach shares this problematic attribute.

These shortcomings of the Livezey–Chen approach to addressing test multiplicity can in general be improved upon through the use of a global test statistic that depends on the magnitudes of the individual p values of the N local tests, rather than on simply counting the numbers of local tests having p values smaller than the chosen α. An attractive choice is to jointly analyze the results of the N multiple tests in a way that controls the *false discovery rate*, or FDR (Benjamini and Hochberg, 1995; Ventura et al., 2004), which is the expected fraction of nominally significant tests whose null hypotheses are actually valid. This terminology derives from the medical statistics literature, where rejection of an individual null hypothesis might correspond to a medical discovery, and survivorship bias in multiple testing is accounted for by controlling the maximum expected rate of erroneous null hypothesis rejection. Within the field significance paradigm, this ceiling on the FDR is numerically equal to the global test level, α_{global} (Wilks, 2006a).

The FDR approach to evaluating multiple hypothesis tests begins with the order statistics for the p values of the N tests, $p_{(1)}, p_{(2)}, p_{(3)}, \ldots, p_{(N)}$. The smallest (nominally most significant) of these is $p_{(1)}$, and the largest (least significant) is $p_{(N)}$. Results of individual tests are regarded as significant if the corresponding p value is no greater than

$$p_{\text{FDR}} = \max_{j=1,\cdots,N} \left\{ p_{(j)} : \; p_{(j)} \le \frac{j}{N} \, \alpha_{\text{global}} \right\}. \tag{5.39}$$

That is, the sorted p values are evaluated with respect to a sliding scale, so that if the largest of them, $p_{(N)}$, is no greater than $\alpha_{\text{global}} = \text{FDR}$, then all N tests are regarded as statistically significant at that level. If $p_{(N)} > \alpha_{\text{global}}$ (i.e., if the largest p value does not lead to rejection of the corresponding null hypothesis) then survivorship bias is compensated by requiring $p_{(N-1)} \le (N-1)\alpha_{\text{global}}/N$ in order for the second-least-significant test and all others with smaller p values to have their null hypotheses rejected. The general rule is that the null hypothesis for the test having the largest p value satisfying Equation 5.39 is rejected, as are the null hypotheses for all other tests with smaller p values. Survivorship bias is addressed by requiring a more stringent standard for declaring statistical significance as progressively smaller p values are considered in turn. If p values for none of the N tests satisfy Equation 5.39, then none are deemed to be statistically significant at the α_{global} level, and in effect the global null hypothesis that all N local tests have valid null hypotheses is not rejected at the α_{global} level. Equation 5.39 is valid regardless of the form of the hypothesis tests that produce the p values, provided that those tests operate at the correct level.

Example 5.17. Illustration of the FDR Approach to Multiple Testing

Consider again the hypothetical situation in Example 5.16, where $N = 20$ independent tests have been computed, of which 3 have p values smaller than $\alpha = 0.05$. The FDR approach accounts for how much smaller than $\alpha = 0.05$ each of these p values is, and thus provides greater test power than does the Livezey–Chen approach. Figure 5.12 plots the magnitudes of $N = 20$ hypothetical ranked p values of which, consistent with the calculations in Example 5.16, three are smaller than $\alpha = 0.05$ (*dotted horizontal line*). According to the calculations in Example 5.16, none of the corresponding hypothesis tests would be regarded as significant according to the Livezey–Chen approach, because $\Pr\{X \geq 3\} = 0.076 > 0.05 = \alpha_{\text{global}}$.

However, the Livezey–Chen approach does not consider how much smaller each of these p values is relative to $\alpha = 0.05$. As drawn in Figure 5.12, these smallest p values are $p_{(1)} = 0.001$, $p_{(2)} = 0.004$, and $p_{(3)} = 0.034$. Since all of the remaining p values are larger than $\alpha_{\text{global}} = 0.05$, none can satisfy Equation 5.39. Neither does $p_{(3)} = 0.034$ satisfy Equation 5.39, since $0.034 > (3/20)(0.05) = 0.0075$. However, $p_{(2)} = 0.004$ does satisfy Equation 5.39, so that the individual tests corresponding to both of the smallest two p values would be regarded as significant. Both would also be declared significant even if it were the case that $p_{(1)} > \alpha_{\text{global}}/N = 0.05/20 = 0.0025$. Both $p_{(1)}$ and $p_{(2)}$ are small enough that it is unlikely that they arose by chance from valid null hypotheses, even after accounting for the survivorship bias inherent in focusing on the most nominally significant of the N tests. Geometrically, the FDR approach rejects the null hypothesis for the largest ranked p value below the sloping dashed line in Figure 5.12, corresponding to $p_{\text{FDR}} = \alpha_{\text{global}}(\text{rank}(p_{(j)})/N)$, and likewise for any other tests having smaller p values.

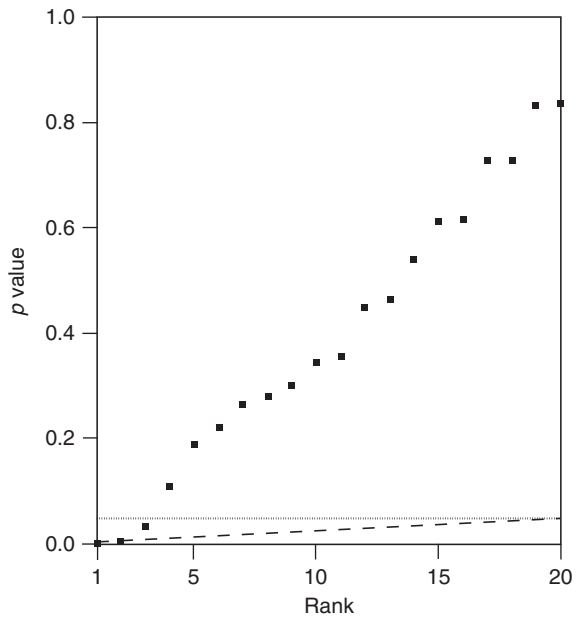

FIGURE 5.12 A hypothetical collection of 20 ranked p values, three of which are smaller than $\alpha = 0.05$ (*dotted horizontal line*). As detailed in Example 5.16, none would be considered statistically significant according to the Livezey–Chen procedure. Using the FDR approach, the largest p value below the sloping dashed line, and any other smaller p values, (in this case, the smallest two p values) would correspond to significant tests. All three would be declared significant according to the naive stippling procedure.

In contrast, the naive stippling approach ignores the effects of test multiplicity altogether, and would consider all three p values falling below the dotted horizontal line to be significant at the 5% level. ◇

5.4.3. Field Significance and Spatial Correlation

When a collection of multiple tests are performed using data from spatial fields, the positive spatial correlation of the underlying data induces statistical dependence among the local tests. Informally, we can imagine that positive correlation between data at two locations would result in the probability of a Type I error (falsely rejecting H_0) at one location being larger if a Type I error had occurred at the other location. This is because a hypothesis test statistic is a statistic like any other—a function of the data—and, to the extent that the underlying data are correlated, the statistics calculated from them will be also. Thus false rejections of the null hypothesis tend to cluster in space, possibly leading (if we are not careful) to the erroneous impression that a spatially coherent and physically meaningful feature may exist.

The binomial distribution underpinning the traditional Livezey–Chen procedure is very sensitive to positive correlation among the outcomes of the N tests, yielding too many spurious rejections of null hypotheses that are true. The same is true of the naive stippling approach, and indeed this is a usual response of conventional testing procedures to correlated data, as was illustrated in a different setting in Section 5.2.4. One approach suggested by Livezey and Chen (1983) was to hypothesize and estimate some number $N' < N$ of effectively independent gridpoint tests, as a spatial analog of Equation 5.12. A variety of approaches for estimating these "spatial degrees of freedom" have been proposed. Some of these have been reviewed by Van den Dool (2007), although even if this quantity can be estimated meaningfully the Livezey–Chen procedure still lacks power.

Another virtue of the FDR approach is that its sensitivity to correlation among the multiple tests is modest, and if anything results in tests that are more conservative (i.e., rejection levels that are smaller than the nominal α). Figure 5.13 illustrates this characteristic of the FDR approach for synthetic data generated on a $2° \times 3°$ latitude–longitude grid (Figure 5.14) mimicking much of the northern hemisphere, when each of the local null hypotheses on the 3720-point grid is valid. The horizontal axis in Figure 5.13 corresponds to increasing spatial correlation, with the vertical arrow at 1.54×10^3 km indicating a typical value for the 500 mb height field. The figure shows that the FDR approach is somewhat conservative when the underlying data are spatially correlated, but that approximately correct results can be achieved by choosing $\alpha_{\mathrm{FDR}} = 2\alpha_{\mathrm{global}}$.

By placing a control limit on the fraction of nominally significant gridpoint test results that are spurious, the FDR method enhances scientific interpretability of the patterns of significant results by focusing attention only on those results unlikely to have been inflated by survivorship bias. Figure 5.14 illustrates this point for the 3720-point grid of synthetic hypothesis tests to which Figure 5.13 also pertains, constructed with the level of spatial correlation indicated by the arrow in Figure 5.13. Here the 156 gridpoint null hypotheses in the central octagonal region are not true, and the remaining 3654 null hypotheses are valid. Figure 5.14a shows results for the FDR method with $\alpha_{\mathrm{FDR}} = 0.10$, which is expected to yield FDR control at approximately the 5% level as indicated by Figure 5.13. The octagonal "signal" is clearly detected even though p values for 12 of the gridpoint tests in its interior (circles) fall above the threshold defined by Equation 5.39. Only six gridpoints near the leftmost edge of the figure yield rejections of valid null hypotheses, yielding the false discovery proportion $6/(144+6) = 0.04$. In contrast, Figure 5.14b shows corresponding results for the naive stippling method, where any gridpoint test with a p value smaller than 0.05 is flagged as nominally significant. In

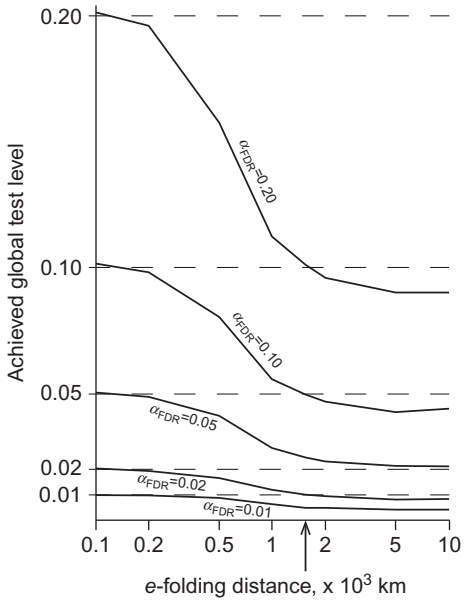

FIGURE 5.13 Achieved global test levels, which are probabilities of rejecting at least one gridpoint null hypothesis using the FDR sliding scale of Equation 5.39 when all null hypotheses are valid, for 3720 gridpoin *t*-tests as functions of the strength of the spatial correlation. For moderate and strong spatial correlation, approximately correct results can be achieved by choosing $\alpha_{FDR} = 2\alpha_{global}$. Arrow locates spatial correlation corresponding to northern hemisphere 500mb height fields. *From Wilks (2016b).*

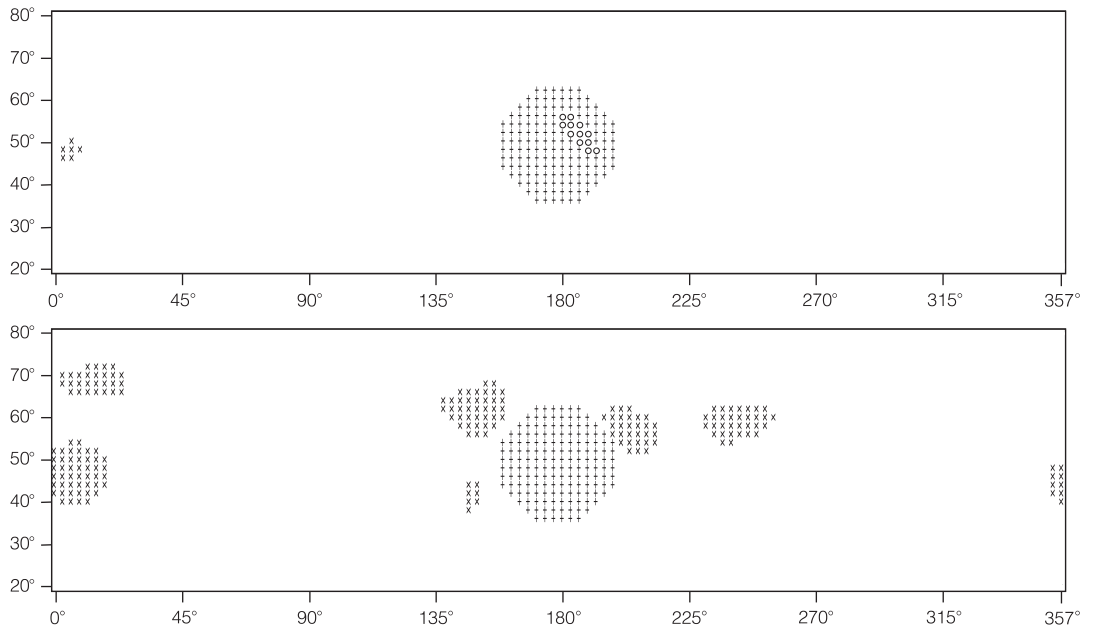

FIGURE 5.14 Maps of local test decisions made by (a) the FDR procedure with $\alpha_{FDR} = 0.10$ and (b) the naive stippling approach using $\alpha = 0.05$. Correct gridpoint null hypotheses are indicated by plus signs, failures to reject false null hypotheses are indicated by circles, and erroneous rejections of valid null hypotheses are indicated by x's. *From Wilks (2016b).*

addition to detecting the octagonal region of 156 incorrect null hypotheses, this method also indicates several regions of nominally but falsely significant tests, composed of 189 gridpoints in total. The false discovery proportion for the naive stippling method is therefore $189/(156+189) = 0.55$, so that a majority of the gridpoint rejections are erroneous. The incorrectly rejected null hypotheses cluster spatially because of the spatial correlation of the underlying data, and accordingly this result might tempt an investigator to waste effort in constructing misleading rationales for their existence.

5.5. ANALYSIS OF VARIANCE AND COMPARISONS AMONG MULTIPLE MEANS

Section 5.2.1 reviewed the traditional approach to statistical inferences about a single mean, and Sections 5.2.2 and 5.2.3 considered the conventional methods for comparing two means. It is possible to extend these procedures to comparisons among multiple means, using a procedure known as Analysis of Variance (ANOVA). ANOVA was developed early in the last century (Fisher, 1925, 1935), mainly to support analysis of data from scientific (mainly agricultural and biological) experiments, and so different forms of ANOVA relate to different experimental designs. Such controlled-experiment applications in meteorology and climatology have been relatively few (e.g., Hingray et al., 2007; Räisänen, 2001; Sain et al., 2011; Sansom et al., 2013), because observational meteorological and climatological data cannot be manipulated in an experimental setting, and comprehensive computational experiments with dynamical weather and climate models are lengthy and expensive. However with increasing computing capacity and multi-institutional coordination such studies are becoming more common (e.g., Taylor et al., 2012).

5.5.1. One-Way ANOVA, and the Completely Randomized Experimental Design

One-way ANOVA is the simplest approach to multimean comparisons. It was developed to support the simple and straightforward experimental approach known as the completely randomized experimental design. In this design there are some number $J \geq 2$ experimental treatments, to which n_j experimental units have been assigned at random. The subscript on the sample sizes indicates that these need not be equal across the treatments. The ith experimental unit in the jth treatment exhibits the quantitative response $x_{i,j}$, and the scientific interest is comparison of the mean responses $\overline{x}_j$, $j = 1, \ldots, J$, under the null hypothesis that all J of these have been drawn from a population or generating process with the same mean, μ. For $J = 2$ the analysis and inference are equivalent to the (two-tailed) two-sample t-test in Equation 5.9.

The underlying statistical model for the one-way ANOVA is

$$X_{i,j} = \mu + \alpha_j + \varepsilon_{i,j}. \tag{5.40}$$

Here the coefficients $\alpha_j = \mu_j - \mu$ are the treatment effects, so that the null hypothesis implies $\alpha_1 = \alpha_2 = \ldots = \alpha_J = 0$. The treatment effect coefficients are constrained according to

$$\sum_{j=1}^{J} n_j \alpha_j = 0 \tag{5.41}$$

in order for the sum of the individual treatment means μ_j to yield the overall mean μ. The errors, or residuals, $\varepsilon_{i,j}$ are assumed to be independently Gaussian distributed, with mean zero and variance σ^2.

The ANOVA proceeds by deriving two independent variance estimates, which will both estimate the true common variance σ^2 if the null hypothesis is true. These are derived from computation of the total sum of squares around the overall mean

$$\mathrm{SST} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(x_{i,j} - \overline{x}\right)^2, \tag{5.42a}$$

the (weighted) sum of the squares of the treatment means around the overall mean

$$\mathrm{SSA} = \sum_{j=1}^{J} n_j \left(\overline{x}_j - \overline{x}\right)^2 \tag{5.42b}$$

and the sum of the squares of the data values around their respective treatment means

$$\mathrm{SSE} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(x_{i,j} - \overline{x}_j\right)^2. \tag{5.42c}$$

These three quantities are related according to SST = SSA + SSE. Here the J treatment means are

$$\overline{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{i,j} \tag{5.43a}$$

and the overall mean is simply

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} x_{i,j}, \tag{5.43b}$$

where the total sample size $n = n_1 + n_2 + \ldots + n_J$. The two variance estimates are computed from equations 5.42b and 5.42c, by dividing by their respective degrees of freedom,

$$\mathrm{MSA} = \frac{\mathrm{SSA}}{J-1} \tag{5.44a}$$

and

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-J}. \tag{5.44b}$$

The degrees of freedom are $J-1$ in the denominator of Equation 5.44a because only $J-1$ of the treatment means $\overline{x}_j$ need to be defined in order for their weighted sum to yield the overall mean $\overline{x}$. Similarly, only $n-J$ of the data values $x_{i,j}$ need be known if the J treatment means $\overline{x}_j$ in Equation 5.42c are given.

If the previously stated assumptions and the null hypothesis are true, both of the quantities in Equations 5.44 estimate the common variance σ^2. To the extent that one or more of the α_j in Equation 5.40 are not zero, then SSA in Equation 5.42b (and therefore also MSA in Equation 5.44a) will be inflated, and SSE in Equation 5.42c (and therefore also MAE in Equation 5.44b) will be deflated. Possible violations of the null hypothesis are accordingly tested by computing the ratio

$$F = \mathrm{MSA}/\mathrm{MSE}. \tag{5.45}$$

TABLE 5.10 Generic One-Way ANOVA Table

Source	*df*	SS	MS	F
Total	$n-1$	SST (Equation 5.42a)		
Treatment	$J-1$	SSA (Equation 5.42b)	MSA (Equation 5.44a)	F (Equation 5.45)
Error	$n-J$	SSE (Equation 5.42c)	MSE (Equation 5.44b)	

Typically these results for a one-way ANOVA are presented in tabular form, which is illustrated generically in Table 5.10. If the null hypothesis that all μ_j are equal is true, then the sampling distribution of the statistic in Equation 5.45 is the F distribution, with degrees-of-freedom parameters $\nu_1 = J-1$ and $\nu_2 = n-J$. Violations of the null hypothesis will tend to lead to larger values of this ratio, so that H_0 is rejected if Equation 5.45 is sufficiently large. Right-tail critical values for the F distribution can be found in most introductory statistics textbooks.

Rejecting the null hypothesis in the one-way ANOVA indicates that at least one of the α_j in Equation 5.40 is unlikely to be zero, but this does not specify how many or which of these are responsible for the rejection result. A one-sample test addressing the significance of one of the treatment means can be computed using the conventional t test using the statistic

$$t = \frac{\bar{x}_j - \bar{x}}{\sqrt{\text{MSE}/n_j}}, \tag{5.46}$$

corresponding to Equation 5.3; and similarly a two-sample t-test addressing the equality of two of the treatment means can be computed using

$$t = \frac{\bar{x}_j - \bar{x}_k}{\sqrt{\text{MSE}\left(\frac{1}{n_j} + \frac{1}{n_k}\right)}}, \tag{5.47}$$

corresponding to Equation 5.9. In both of these cases the appropriate null distribution is the t distribution with $\nu = n - J$ degrees of freedom, and confidence intervals can be computed in the usual way.

The tests defined by Equations 5.46 and 5.47 are valid if the question being investigated has been defined in advance of seeing the data underlying the ANOVA calculations. If the means being tested are instead chosen on the basis of "interesting" post hoc results (e.g., the largest $\bar{x}_j$ or the largest difference between two treatment means) the inferential results will be optimistic due to survivorship bias (Section 5.4). Similarly, computing two or more of the J one-sample tests in Equation 5.46, or two or more of the $J(J-1)/2$ possible two-sample tests in Equation 5.47, will require adjustments for test multiplicity. Possible approaches to dealing with these problems are to use the Bonferroni adjustment to the tail probabilities (i.e., require $p \leq \alpha/m$, where m is the number of multiple tests), employ false discovery rate control (Section 5.4.2), or compute a range test such as Tukey's "honest significant difference" (e.g., Steel and Torrie, 1960, pp. 109–110; available as the R routine TukeyHSD).

5.5.2. Two-Way ANOVA, and the Randomized Block Experimental Design

Just as better inferential precision can be achieved for paired data using the paired two-sample t-test outlined in Section 5.2.3, more powerful ANOVA testing can be achieved when the data can be grouped a priori into blocks within which the data are sufficiently strongly correlated. In the classical setting of agricultural trials, where the experimental treatments could be different plant varieties or different fertilization regimes, results from treatments imposed at multiple locations would form natural blocks because of the correlations in crop yield induced by the same weather sequences and similar soil conditions. In computer experiments that might look at responses of dynamical climate models to different boundary conditions and/or parameterization schemes, results obtained with models from different research centers would likely be good candidates for grouping into blocks. Such experiments would be examples of what are called randomized block designs.

In the simplest implementation of the randomized block design, there are J treatments as before, with experimental units grouped into K blocks that each contain one of the J experimental units. Thus the total sample size is $n = JK$. The underlying statistical model, corresponding to Equation 5.40, is

$$X_{j,k} = \mu + \alpha_j + \beta_k + \varepsilon_{j,k}, \tag{5.48}$$

where as before the α_j are the treatment effects, and now in addition the β_k, $k = 1, 2, \ldots, K$, are the block effects. Both the block and treatment effects specify possible adjustments to the overall mean μ so that now $\Sigma_j \alpha_j = \Sigma_k \beta_k = 0$. As before, the null hypothesis of no treatment effects implies $\alpha_1 = \alpha_2 = \ldots = \alpha_J = 0$.

Again the ANOVA table is based on sums of squared deviations around the overall mean $\overline{x}$, which are the total sum of squares

$$\mathrm{SST} = \sum_{k=1}^{K} \sum_{j=1}^{J} \left(x_{j,k} - \overline{x}\right)^2, \tag{5.49a}$$

the sum of squares of the block means around the overall mean is

$$\mathrm{SSB} = J \sum_{k=1}^{K} \left(\overline{x}_k - \overline{x}\right)^2, \tag{5.49b}$$

the sum of squares of the treatment means around the overall mean is

$$\mathrm{SSA} = K \sum_{j=1}^{J} \left(\overline{x}_j - \overline{x}\right)^2, \tag{5.49c}$$

and the error sum of squares is most easily computed as the residual

$$\mathrm{SSE} = \mathrm{SST} - \mathrm{SSB} - \mathrm{SSA}. \tag{5.49d}$$

Because there is a single experimental unit for each combination of treatment and block, the block and treatment means in Equations 5.49b and 5.49c are simply

$$\overline{x}_k = \frac{1}{J} \sum_{j=1}^{J} x_{j,k} \tag{5.50a}$$

TABLE 5.11 Generic Two-Way ANOVA Table

Source	*df*	SS	MS	F
Total	$n-1$	SST (Equation 5.49a)		
Blocks	$K-1$	SSB (Equation 5.49b)	MSB (Equation 5.51a)	MSB/MSE
Treatment	$J-1$	SSA (Equation 5.49c)	MSA (Equation 5.44a)	F (Equation 5.45)
Error	$n-K-J+1$	SSE (Equation 5.49d)	MSE (Equation 5.51b)	

and

$$\bar{x}_j = \frac{1}{K}\sum_{K=1}^{K} x_{j,k}, \tag{5.50b}$$

respectively. Equation 5.49d shows that, to the extent that there are strong block effects (i.e., the block means differ strongly from the overall mean in Equation 5.49b), the error sum of squares SSE will be reduced when this variation is subtracted, leading to more precise inferences.

The two-way ANOVA table is an extension of the one-way ANOVA table in Table 5.10, but with an additional line for the block effects, as shown in Table 5.11. The sums of squares in Equations 5.49 are arranged in the SS column. Under the null hypothesis of no treatment effects and no block effects on the mean (all $\alpha_j = 0$ and all $\beta_k = 0$ in Equation 5.48), SSB, SSA, and SSE are all independent estimates of the assumed common variance σ^2. Therefore the presence of such effects should be detectable in the ratios of the respective MS entries, given by the corresponding SS values divided by their respective degrees of freedom in the *df* column. These entries are given by Equation 5.44a for MSA,

$$\text{MSB} = \frac{\text{SSB}}{K-1} \tag{5.51a}$$

and

$$\text{MSE} = \frac{\text{SSE}}{n-K-J+1}. \tag{5.51b}$$

The primary interest is generally in possible treatment effects, examined using the F ratio MSA/MSE (-Equation 5.45), the null distribution for which is F, now with degrees-of-freedom parameters $\nu_1 = J-1$ and $\nu_2 = n-K-J+1$. Similarly the null hypothesis of no block effects can be examined using the statistic $F = \text{MSB/MSE}$, the null distribution for which is F with parameters $\nu_1 = K-1$ and $\nu_2 = n-K-J+1$.

Example 5.18. Comparing One-Way and Two-Way ANOVA, and the Respective *t*-Tests
Table 5.12 contains maximum temperature data extracted from Table A.1, but sampled every third day in order to suppress the serial correlation. Considering the two locations as $J = 2$ "treatments," Table 5.13a presents the one-way ANOVA table for examining the null hypothesis that the mean temperatures are equal. The treatment MSA (Equation 5.44a) is of comparable magnitude to the MSE (Equation 5.44b),

TABLE 5.12 Ithaca and Canandaigua Maximum Temperature Data (°F) from Table A.1, Sampled Every Third Day of January 1987 to Suppress Serial Correlation

Date	Ithaca T_{max}	Canandaigua T_{max}
1	33	34
4	29	29
7	37	44
10	30	33
13	33	34
16	45	44
19	32	36
22	28	29
25	9	11
28	26	26
31	34	38
Average	30.54	32.54
Std. Dev.	8.78	9.17

suggesting that the null hypothesis of equal means is plausible according to these calculations. Comparing the computed $F = 0.273$ to the F distribution with $\nu_1 = 1$ and $\nu_2 = 20$ degrees of freedom yields the p value 0.607, so that the null hypothesis is not rejected by this test.

Because the temperatures at these two nearby locations are strongly positively correlated (Table 3.5), much of the variance estimated by the MSE in Table 5.13a is shared by the two temperature data sets, and so does not contribute to uncertainty regarding their average difference. Accordingly the two-way ANOVA in Table 5.13b, with each of the $K = 11$ days regarded as a block, is expected to provide more precise and powerful inferences about any difference in the $J = 2$ means. Here SSB is quite large because the two temperature values in each block are strongly correlated. This shared variability is quantified by MSB, and is subtracted in Equation 5.49d to yield the much smaller MSE in Table 5.13b, even though the blocking has consumed an additional 10 degrees of freedom relative to the one-way analysis in Table 5.13a. These lost 10 degrees of freedom are therefore not available in the denominator of the computation of MSE in Table 5.13b. The primary interest is usually in the treatment effect, reflected by $F = 8.148$. Comparing this value with the F distribution having $\nu_1 = 1$ and $\nu_2 = 10$ degrees of freedom yields the $p = 0.017$, providing strong evidence against the null hypothesis. Because of the strong spatial correlation for the data at these two locations, the null hypothesis of no block effect is rejected even more strongly, since $F = 58.69$ with $\nu_1 = \nu_2 = 10$ degrees of freedom yields $p = 1.6 \times 10^{-7}$.

Since there are $J = 2$ treatment means being compared, the two ANOVAs in Table 5.13 are equivalent to the corresponding two-sample t-tests. In particular, quantiles of F distributions with $\nu_1 = 1$ degree of freedom are equal to the squares of (two-tailed) t distribution quantiles having ν_2 degrees of freedom. The ANOVA in Table 5.13a, which ignores the spatial correlation that defines the paired

TABLE 5.13 One-Way (a) and Two-Way (b) ANOVA Tables for the Temperature data in Table 5.12, Considering the Locations as $J = 2$ Treatments

(a)

Source	*df*	SS	MS	*F*
Total	21	1633.45		
Treatment	1	22.00	22.00	0.273
Error	20	1611.45	80.57	

(b)

Source	*df*	SS	MS	*F*
Total	21	1633.45		
Blocks	10	1588.45	158.45	58.69
Treatment	1	22.00	22.00	8.148
Error	10	27.00	2.70	

The two-way ANOVA in (b) has been constructed using individual days as the $K = 11$ blocks because of the spatial correlation between these two nearby locations.

nature of these data, is therefore equivalent to the two-sample t-test in Equation 5.9, because equal variances for the two samples has been assumed. The resulting test statistic is

$$t = \frac{30.54 - 32.54}{\left[\left(\frac{1}{11} + \frac{1}{11}\right)\left(\frac{8.78^2 + 9.17^2}{2}\right)\right]^{1/2}} = -0.5226, \tag{5.52}$$

and $t^2 = -0.5226^2 = 0.273$ matches the F ratio in Table 5.13a.

Similarly, the two-way ANOVA with individual days as blocks is equivalent to the paired two-sample t-test in Equation 5.11 and Example 5.3, which operate on the daily temperature differences and thereby subtract the shared variance. Using the smaller data set in Table 5.12, the resulting test statistic is

$$t = \frac{-2}{\left(2.324^2/11\right)^{1/2}} = -2.8545. \tag{5.53}$$

Again $t^2 = -2.8545^2 = 8.148$, matching the corresponding F ratio in Table 5.13b. The inference here is weaker than in the corresponding test in Example 5.3 because of the reduced sample size. ◇

One of the assumptions used for ANOVA is that the data are independent realizations from their population or generating process. For atmospheric data separated by sufficiently long sampling times, such as annual values, this assumption is generally valid. Zwiers (1987b) has proposed an approach to ANOVA when the data are autocorrelated, based on frequency-domain calculations (Section 10.5). Lund et al. (2016) describe an alternative approach, based on time-domain time series methods (Section 10.3).

The ANOVAs and corresponding experimental designs presented here are the simplest examples from a much larger class of possibilities. One such extension allows multiple experimental units for each treatment-block combination in the two-way ANOVA, in which case it is possible to include interaction coefficients $\gamma_{j,k}$ in Equation 5.48. Another variation is use of two-way ANOVA for a completely randomized design, examining different levels of two treatment types, which is suggested by the symmetry of SSB and SSA in Equations 5.49b and 5.49c, so that the "blocks" are just levels of the second treatment type. Much more on this subject can be found in such sources as Casella (2008), Montgomery (2013), and Wu and Hamada (2009).

5.5.3. Partitioning of Variance Contributions

One use to which ANOVA has been put in the atmospheric sciences literature does not involve hypothesis testing for means, but rather focuses on the variance partition within the ANOVA table. Zwiers (1987b) and Zwiers and Kharin (1998) consider potential predictability in dynamical climate models in the one-way ANOVA setting, with the error variance representing unpredictable internal variability that the analysis distinguishes from potentially predictable interannual variability.

Wang and Zwiers (1999) describe a similar analysis using two-way ANOVA, where responses in different years, and from different dynamical climate model formulations, represent different treatments. Similarly, Yip et al. (2011) and Northrop and Chandler (2014) use two-way ANOVA to partition variance attributable to different dynamical models, different climate-change trajectory scenarios, and internal climate variability, over a century of projected future climate evolution.

5.6. EXERCISES

5.1 For the June temperature data in Table A.3,
   a. Use a two-sample t-test to investigate whether the average June temperatures in El Niño and non-El Niño years are significantly different. Assume the variances are unequal and that the Gaussian distribution is an adequate approximation to the distribution of the test statistic.
   b. Construct a 95% confidence interval for the difference in average June temperature between El Niño and non-El Niño years.

5.2 Calculate n', the equivalent number of independent samples, for the two sets of minimum air temperatures in Table A.1.

5.3 Consider a data series for which $n = 100$, $\Sigma x = 25.00$, $\Sigma x^2 = 2481.25$, and for which the lag-1 autocorrelation is +0.6. Using an appropriately modified t-test, investigate the null hypothesis that the true mean is zero, versus the alternative that the true mean is greater than zero. Assume the Gaussian distribution is an adequate approximation to the true (t) distribution, and report a p value.

5.4 Use the data set in Table A.1 to test the null hypothesis that the average minimum temperatures for Ithaca and Canandaigua in January 1987 are equal. Compute p values, assuming the Gaussian distribution is an adequate approximation to the null distribution of the test statistic, and
   a. H_A = the minimum temperatures are different for the two locations.
   b. H_A = the Canandaigua minimum temperatures are warmer.

5.5 Suppose you are testing for significantly different means between a pair of time series. Assume that the standard deviations for both of these series are 2.7 units, the lag-1 autocorrelation for each is +0.37, and the sample sizes for the two series are equal but the data are not paired. How big do these samples need to be to reject a null hypothesis of no difference at the 10% level, if the observed difference in the means is 1.0 units? You have no prior reason to expect one mean to be larger than the other. Assume the standard Gaussian distribution is an adequate approximation for your test statistic.

5.6 You have 30 years' of daily January temperature data at a location of interest. The standard deviation of this set of $30 \times 31 = 930$ daily temperature values is 14.14°F. When you compute the 30 monthly mean temperatures, and then calculate the standard deviation of these 30 values, you find it is 4.40°F. Assuming the statistical properties of the climatology of temperature do not change appreciably from one part of January to another or from year to year,
   a. Estimate the lag-1 autocorrelation for the daily temperature data.
   b. Determine the standard deviation of 5-day means of the daily data.

5.7 The Fisher Z-transform (Equation 3.25) can be used for statistical inferences regarding correlation coefficients. For a null hypothesis of zero correlation, the sampling distribution of Z is Gaussian, with mean zero and variance $1/(n-3)$. How large a correlation coefficient computed from $n=30$ independent data pairs is required to reject a null hypothesis of zero correlation at the 5% level, versus the alternative that $\rho \neq 0$?

5.8 Test the fit of the Gaussian distribution to the July precipitation data in Table 4.8, using.
- a. A K–S (i.e., Lilliefors) test.
- b. A Chi-square test.
- c. A Filliben Q–Q correlation test.

5.9 Test whether the 1951–80 July precipitation data in Table 4.8 might have been drawn from the same distribution as the 1951–80 January precipitation comprising part of Table A.2, using a likelihood ratio test, assuming gamma distributions.

5.10 Use the Wilcoxon–Mann–Whitney test to investigate whether the magnitudes of the pressure data in Table A.3 are lower in El Niño years,
- a. Using the exact one-tailed critical values 18, 14, 11, and 8 for tests at the 5%, 2.5%, 1%, and 0.5% levels, respectively, for the smaller of U_1 and U_2.
- b. Using the Gaussian approximation to the sampling distribution of U.

5.11 Discuss how the sampling distribution of the skewness coefficient (Equation 3.9) of June precipitation at Guayaquil could be estimated using the data in Table A.3, by bootstrapping. How could the resulting bootstrap distribution be used to estimate a 95% confidence interval for this statistic? If the appropriate computing resources are available, implement your algorithm.

5.12 Discuss how to construct a resampling test to investigate whether the variance of June precipitation at Guayaquil is different in El Niño versus non-El Niño years, using the data in Table A.3.
- a. Assuming that the precipitation distributions are the same under H_0.
- b. Allowing other aspects of the precipitation distributions to be different under H_0.

If the appropriate computing resources are available, implement your algorithms.

5.13 Consider the following sorted p values from $N=10$ independent hypothesis tests: 0.007, 0.009, 0.052, 0.057, 0.072, 0.089, 0.119, 0.227, 0.299, 0.533.
- a. Do these results support a conclusion of "field significance" (i.e., at least one of the 10 local null hypotheses can be rejected) at the $\alpha_{\text{global}}=0.05$ level using either the Livezey–Chen binomial "counting" test with $\alpha=0.05$, or the FDR approach?
- b. Which if any of the p values would lead to rejection of the respective local null hypotheses according to the calculations in part a, using each of the methods?

5.14 Augment the data in Table 5.12 with the minimum temperatures for Ithaca and Canandaigua from Table A.1, for the same 11 days. Then, considering the four temperature variables as "treatments,"
- a. Construct the one-way ANOVA table for testing equality of the four means.
- b. Construct the two-way ANOVA table for testing equality of the four means, considering the 11 days as blocks.