Chapter 6
Hypothesis Testing

> The temptation to form premature theories upon insufficient data is the bane of our profession—Sherlock Holmes, in Arthur Conan Doyle, *The Valley of Fear* (1914).

> A phenomenon having been observed, or a group of phenomena having been established by empiric classification, the investigator invents an hypothesis in explanation. He then devises and applies a test of the validity of the hypothesis. If it does not stand the test he discards it and invents a new one. If it survives the test, he proceeds at once to devise a second test. And thus he continues.—G. K. Gilbert (1886), The inculcation of scientific method by example with an illustration drawn from the Quaternary geology of Utah, *Amer. J. Sci*, **136**, 284-299.

6. Introduction

We now turn from estimating parameters of probability density functions, to testing hypotheses, specifically testing **statistical hypotheses**. For science in general, a hypothesis is some assertion we make about the way the world is; a statistical hypothesis is something much more restricted, namely an assertion about how a dataset relates to some kind of probability model. The idea of testing applies to both. Some examples of scientific hypotheses and the statistical hypotheses they give rise to would be:

- We may hypothesize that there was a change in the core dynamo between the time of the Cretaceous Superchron (Section 1.2) and the subsequent period of frequent reversals. The statistical hypothesis to go with this would be that a point process that models all the other reversals (that is, a pdf $\phi(t)$ for the inter-reversal times) would be very unlikely to produce so long a time without reversals.
- We may hypothesize that earthquakes are triggered by earth tides. The statistical hypothesis to go with this would be that earthquakes occur more often than not, on average, at times related to (say) high and low tides—as opposed to occurring "at random" relative to the tides.
- We may want to claim that a new model for seismic velocity in the Earth is better than an existing one. The statistical hypothesis to go with this would be that the mismatch between some data (say times of propagation of seismic waves) and the new model is smaller than it was for the old model, by an amount "much greater than" the errors in the measurements.
- And sometimes, we start with a statistical hypothesis; as, for example, that the GPS data in Section 1.1 are normally distributed, as a prelude to further analysis.

In each case, we start by formulating a probability model to go with the actual hypothesis we are considering. How to make this step is not a matter of statistical analysis, but of informed judgment, both about the particular problem and about the methods that might be available for deciding if a probability model is supported by the data or not. This chapter gives some general principles about testing statistical hypotheses, and describes some of the methods that are used most often.

An important function of statistical tests is to keep us from being fooled into thinking that what we have observed indicates something important, when it might as well be expected to happen by chance. Long experience shows that we can be easily fooled: the normal human propensity is to find patterns even when none are present.

6.1. Problems and Caveats

To begin, we offer some general remarks about this branch of statistics. There are many different hypothesis tests, in part because of the range of questions we may try to answer; but also because is that there are long-standing and fundamental disagreements about the basic principles of testing. In many cases different principles end up leading to similar results, but these disagreements make this subject more difficult to learn. Technical issues aside, it may be that these disagreements have been so hard to resolve because different approaches are appropriate to different areas of reasoning; methods that are appropriate in an economic context (where costs and benefits are clear) are less obviously useful for making inferences about scientific theories. We shall select what seems most useful while admitting that it may have less of a logical basis than we would like.

Leaving aside Bayesian inference, which is a separate approach to almost all of statistics, there are two approaches to hypothesis testing:

1. The procedures developed by R.A. Fisher, which focus on the use of tests to determine if data are consistent with some assumption; as we will see, this is often done by showing that the data are in fact inconsistent with the opposite assumption.
2. The **Neyman-Pearson** approach to hypothesis testing, which sought to formalize and justify some of Fisher's methods by expressing hypothesis testing as a choice between hypotheses. In this framework it is possible to define tests which are in some sense "best": this is rigorous, but may not be applicable to the kinds of inference we may wish to make.

A third approach has been called the "hybrid" method, though "bastardized" might be better; this is what is usually taught to non-statisticians—and this course will be no exception. This combines parts of both the Fisher and Neyman-Pearson procedures to produce a methodology that is easier to describe, even though it is not fully consistent. But it does satisfy our aim to infer no less and (especially) no more from the data than we should.

6.2. A Framework for Tests

If we *knew* that our data were in fact modeled by a random variable X, whose density function we also knew, we would know all that we could: for example, that the data can be described by Normal random variables with known mean and variance. Estimation was about finding the "best values" for parameters in the density function. **Hypothesis testing** is more general, in that it is about testing statements about the density function that produces the random variables that are believed to describe the data. One such statement (fundamental to the Neyman-Pearson approach) involves a choice between hypotheses. The Fisherian approach is

to say that we can see if a particular hypothesis is inconsistent with the data: often this can be quite useful.

Up to a point, the basic schema for hypothesis testing is the same in either method, and in part resembles the schema used for estimation. First, we create a statistical hypothesis that relates to the actual thing we are interested in; as noted above, this is a matter of judgment, not something that can be done mechanically. Often, what we actually do is set up a statistical hypothesis contrary to what we want to show; this is called the **null hypothesis**, conventionally denoted as H_0. Whatever hypothesis we choose stipulates that the data can be modeled, by some kind of random variable. In hypothesis testing we can allow much more general pdf's than we could in estimation; for example, in the class of tests called "distribution-free" we assume only that the data come from a pdf of some kind, otherwise unspecified.

Having set up a statistical model for the data, we now proceed as follows:

A. Compute a **test statistic** $T(\vec{x})$ from the data $\vec{x}$; that is, we take the data and produce a relevant number (or numbers), just as we did in estimation.

B. Determine the pdf for the random variable equivalent to T under the assumption that the data are modeled in the way that the null hypothesis H_0 assumes. We call this pdf $\hat{T} = T(\vec{X}) = \phi(t)$.

C. Given $\phi(t)$, compute

$$\alpha = \int_{-\infty}^{T=-T(\vec{x})} \phi(t)\,dt + \int_{T=T(\vec{x})}^{\infty} \phi(t)\,dt \qquad (1)$$

This is the area under the tails of the pdf (or some cases, the area under only one tail) for values of the test statistic greater in magnitude than the value actually found from the data. This integral, like any other integral of a pdf, is a probability. It is conventional to call the quantity $1 - \alpha$ the **confidence coefficient**, while α is called the **significance level**. What we do next is described in a later section.

6.2.1. An Example: the Schuster Test

We illustrate this by working through a test of a particular hypothesis—one with geophysical relevance.

A common opinion about California earthquakes is that they seem to often occur in the early morning: this was true for the 1906 San Francisco earthquake, and more recently for the San Fernando (1971), Landers (1992), Northridge (1994), and Hector Mine (1999) events; this has been a good thing, since it meant that most people were at home, which in California is a relatively safe place to be.

What we have here is an anecdote; to go beyond this, we want to see if such temporal clustering is true if we look at all earthquakes. In terms of our understanding of how earthquakes work, having large earthquakes correlate with local time would be very odd indeed; so it is not unreasonable to argue that our anecdotal evidence is "just coincidence".

We begin by setting up the null hypothesis, which is that the times of large earthquakes are uniformly distributed throughout the day; that is, that the pdf of the random variable that we claim model the times of earthquakes is uniform over [0, 24) (in hours). More formally, if X is the time of day, we have H_0: $X \sim U(x)$.

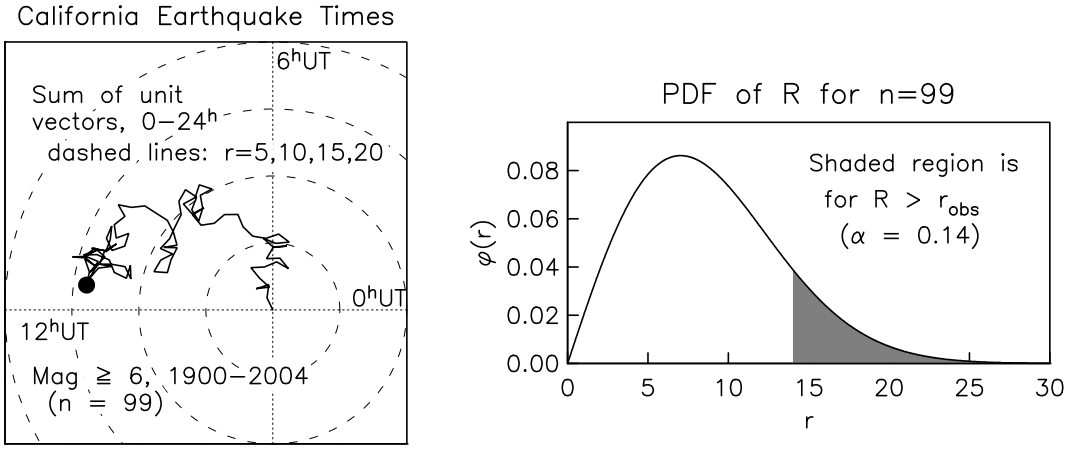

Figure 6.1

One possible a test statistic (not the only one) can be formed from the n observed times by taking

$$r_1 = \sum_{i=1}^{n} \cos(2\pi x/24) \qquad r_2 = \sum_{i=1}^{n} \sin(2\pi x/24)$$

$$R = \sqrt{r_1^2 + r_2^2}$$

That is, we represent the time of each earthquake, x_i, by a unit vector (r_1, r_2), whose direction corresponds to the time on a 24-hour clock; then we add these vectors, and take the distance from the origin to be the test statistic. Obviously, the more clustered the times, the bigger R will be. The left panel of Figure 6.1 shows this procedure applied to all large earthquakes in California since 1900, with the individual unit vectors shown head-to-tail; the large dot is the sum, which turns out to be $r_{obs} = 14.02$.[1] In geophysics this is called the **Schuster test**, after the person who introduced it (for this very problem); in statistics this test is more often named for Rayleigh, who determined the pdf of R when n is large, and the distribution of the X_i's is uniform:

$$\phi(r) = \frac{2r}{n} e^{-r^2/n} \tag{2}$$

which we plot on the right-hand side of Figure 6.1. Because of the central limit theorem, this is the same as the Rayleigh distribution of Chapter 3, even for steps of a fixed length. The shaded region shows the part of the pdf (equation 1) for which $r > r_{obs}$; probability of observing R in this region (supposing the null hypothesis) is a

[1] The earthquakes used are all California events with magnitude 6 or above, for 1900 through 1989 from W. L. Ellsworth (1990), The San Andreas Fault: earthquake history, 1769–1989, *U.S. Geol. Surv. Prof. Pap.*, **1515**, 153-188, with later events from the regional catalogs. Immediate aftershocks in both catalogs were omitted.

probability of 0.14, making $1 - \alpha$ equal to 0.86.

For completeness we note that a better approximation, for $n < 50$, to the probability is

$$\alpha = e^{-z}\left[1 + \frac{2z - z^2}{4n} - \frac{24z - 132z^2 + 76z^3 - 9z^4}{288n^2}\right]$$

where $z = R^2/n$; for n large this just becomes $\alpha = e^{-z}$, consistent with integrating (2) from R to ∞.

6.2.2. What Do We Do With the Results?

The statistical interpretation of the result is simple: if the null hypothesis were true, and we could run the test many times, we would get a value of the test statistic as large as we see, or larger, 14% of the time; we say that we have a significance level of 0.14.

But this is the point at which simplicity, and consensus, end. Here are some things we might do:

1. Report the value of α as a summary of what we got: suggestive (in that one chance in 6 is not that likely), but not really conclusive.
2. Say that, since α does not reach some value α_0 (decided on before we did the test), that we cannot reject the null hypothesis, so that it is reasonable to say that the observed result could have occurred "by chance." This is the original claim that the apparent clustering in time is "just coincidence:" the data do not support this claim.
3. Make the stronger statement that the hypothesis of interest is false (in this case) because the α we observed was above α_0. Conversely, argue that the hypothesis is true if $\alpha \leq \alpha_0$.
4. Take some action depending on whether α exceeds α_0 or not, without prejudice, as it were, regarding the truth or falsity of the hypothesis. In an industrial setting, for example, if we were using testing to evaluate the quality of our manufacturing, we might stop a production line or reject a batch of products. Parallels to this in research are not always obvious, but might be what data to collect, or what other ideas to entertain, next.

We have laid out these options because (in our view) many (but not all) are acceptable in some way. All too often only one is viewed as correct. In particular, it is quite common to pick conventional values for $1 - \alpha_0$ (say 0.95 or 0.99) and then to exercise only option (3), saying that if this value is reached or exceeded, then H_0 is rejected at (say) a 95% confidence level—and further that the alternative that we set out to establish is true. This is a very dubious procedure. As a form of words it may be acceptable to say that a hypothesis has been rejected (option 2), but we should realize that

- There is nothing special about a particular value of α_0. In particular, to view $1 - \alpha = 0.94$ as being a very different outcome from $1 - \alpha = 0.96$ is nonsense.
- We should not confuse having shown that the data do or do not support a hypothesis at some level as having proved anything about its truth (option 3)—such a result simply makes a strong case. But perhaps not that strong: remember that one in twenty times we would reach $\alpha = 0.05$.

Confusion about the meaning of values of α has been exacerbated by the custom (fortunately not present in geophysics) of declaring that if results did not reach some level of significance, they should not be published. This is much too rigid a rule for applying statistical results to scientific inference.

For our example of the times of large California earthquakes, perhaps the best we can say is that anyone arguing for temporal clustering does not have a strong case, by the usual standards of the field. Note that we cannot rule the proposal out, but only say that it is not supported by the data in any convincing way.

6.2.3. The Perils of Going Fishing

This is an appropriate place to discuss a different abuse of the testing procedure, and one that all too easily can happen in geophysics because of the difficulty of often not being able to produce more data just by doing experiments. There is a natural tendency to look through the data to find a pattern—and then, having found it, perform a test. But this destroys the assumption, built into the procedure described above, that we are doing only one test.

An example[2] may make this clearer. Suppose we decide to test the idea that earthquakes occur when the time-varying stress from earth tides is favorable to the kind of faulting that occurred. We collect source mechanisms for a number of earthquakes, and apply a test to see if this is so. (We can use a modified version of the Schuster test for this). And we decide, quite reasonably, to try the test for different types of stress and different types of faulting, ending up with 12 possible combinations. We get a result with $\alpha = 0.04$, and decide that we have established tidal triggering for that particular class.

But we have done no such thing. Suppose the null hypothesis is true. If we choose a significance level of 0.04, the probability of *not* getting a significant result becomes 0.96. Then not seeing a result in 12 independent trials has a probability of $(0.96)^{12} = 0.61$, which means that the probability of getting one such result would be 0.39; this is hardly unlikely. There is nothing wrong with using significance tests to go fishing for a possible result, so long as we do not claim that whatever result we get is in fact significant. It matters in what order we try different things: applying a test to the data, and then stopping, is not the same as trying a number of tests, and finally doing the same test as the original.[3]

In an experimental science, we can (usually) decide to collect more data to see if the significance remains with a new dataset; in geophysics we often cannot. One way around this problem is to divide the data, in advance and at random, into two sets, one for fishing in and one for testing when you are done fishing. Another procedure, called the **Bonferroni method**, is to set the significance level chosen in advance to α/k, where k is the number of tests and α the conventional level for one test. For the tidal-triggering example, this would set the significance level to 0.0033

[2] See T. H. Heaton (1982). Tidal triggering of earthquakes, *Bull. Seismol. Soc. Amer.*, **72**, 2181-2200, updating, correcting, and apologizing for T. H. Heaton (1975). Tidal triggering of earthquakes, *Geophys. J. Roy. Astron. Soc.*, **43**, 307-326.

[3] Likewise, we have to decide on the number of data in advance, and not alter this as we get results—unless, that is, we are using a **sequential test**, which is designed to cover exactly this case.

(0.04/12).

6.3. Tests for Differing Means

We now leave aside the philosophical complexities of what statistical testing does, and move to how to do some common tests. We start with tests for differences in means, followed by tests of whether data conform to a particular pdf. Another category of tests—those involving variances—we will discuss in the context of least-squares fitting, in a later chapter.

6.3.1. Means and Variances Known

To start with a somewhat artificial case, which however we can use to make an important point, suppose our null hypothesis is that the data come from a normal distribution with known mean μ_0 and variance σ^2; and we are going to test the part of this that is about the mean. Though there are few geophysical examples of such a hypothesis, it can arise in other settings, for example, in checking that the dimension of some manufactured item is within a specified tolerance. The usual shorthand for this test would be

$$H_0\text{:}\ \ \mu = \mu_0$$

The conventional phrase to describe such a statistical hypothesis, in which all the parameters of the pdf are known, is that it is a **simple hypothesis**. The Schuster test was also of a simple hypothesis, since the pdf we were testing against was completely specified by the statement that it was uniform over $[0, 24)$.

For this H_0, the test statistic is the difference between the sample mean, $\bar{x}$, and the assumed mean:

$$t = T(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \mu_0 \tag{3}$$

We know, from previous discussions (Section 5.2.1) that the pdf of the statistic, $\hat{t}$, for random variables from the assumed normal distribution, would be

$$\phi(t) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}}\, e^{-nt^2/2\sigma^2} \qquad \text{or} \qquad \hat{t} \sim \mathrm{N}(0, \sigma^2/n) \tag{4}$$

Since H_0 would be invalidated if $\bar{x}$ were either much larger or much smaller than μ_0, we need to include both tails of $\phi(t)$ in determining the significance level; for a given α_0 the level is t_0 such that (from equation (1)):

$$\alpha_0 = \int_{-\infty}^{-t_0} \phi(t)\, dt + \int_{t_0}^{\infty} \phi(t)\, dt \tag{5}$$

But this means that we can write the level t_0 in terms of the cumulative distribution function $\Phi(t)$ for the distribution given by equation (3)—or rather, in terms of its inverse, $\Phi^{-1}(\alpha)$; equation (5) will be satisfied for

$$t_0 = \Phi^{-1}(\alpha_0/2) \tag{6}$$

where the $\alpha_0/2$ comes from the inclusion of both tails of the pdf in (4). H_0 would thus be rejected, with a confidence of $1 - \alpha_0$, if

$$|\bar{x} - \mu_0| \geq t_0 \tag{7}$$

Another way to write this result is to say that H_0 would be rejected if μ_0 fell outside the $1 - \alpha$ confidence interval for $\bar{x}$, because this interval is

$$[\bar{x} + \Phi^{-1}(\alpha/2),\ \bar{x} + \Phi^{-1}(1 - \alpha/2)] \tag{8}$$

There is thus a close relationship between confidence intervals on a statistic, and a test applied to that statistic: since both specify intervals that include a specified amount of probability inside them, they give rise to equivalent limits, though these limits are used differently.

An easy extension of the above is to the case when we have two data sets, which we shall call $\vec{x}_A$ and $\vec{x}_B$, with assumed variances σ^2 for both, and assumed means that differ by $\Delta\mu$. The test statistic to see if the difference in sample means is in fact this value, is then

$$t = (\bar{x}_A - \bar{x}_B) - (\Delta\mu) \tag{9}$$

which reduces to $(\bar{x}_A - \bar{x}_B)$ if we are testing to see if the means are equal. The test statistic $\hat{t}$, assuming a normal distribution, is the convolution of the two distributions for $\hat{\mu}_A$ and $\hat{\mu}_B$, and so is distributed as $N(0, \sigma^2(n_A^{-1} + n_b^{-1}))$. If we use the cdf $\Phi(t)$ appropriate to this distribution, the critical value for the test, t_0 is again given by equation (6), with equations (7) and (8) following as before.

6.3.2. Testing Against a Known Mean, with Unknown Variance

A more interesting case is

$$H_0\text{: } \mu = \mu_0 \qquad \sigma^2 \text{ unknown}$$

Because this involves unknown parameters (other than the one being tested for) this is called a **composite hypothesis**. The test statistic is then

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{10}$$

where s^2 is the sample variance defined in Section 5.3.1. As we showed in that section, the statistic $\hat{t}$, which is

$$\hat{t} = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \left[\frac{\hat{\sigma}}{\sigma}\right]^{-1}$$

is the ratio between an rv with a normal distribution, and an rv distributed as the square root of a χ^2_{n-1} random variable. Such a ratio, and hence the test statistic t, are distributed as Student's t distribution with $n - 1$ degrees of freedom. So we can use the pdf $\phi(t)$ for that distribution to find critical values t_0 for given significance levels α, again using equation (6). (By now we hope you appreciate that all of these cases have the same fundamental structure; only the pdf changes).

6.3.3. Means Unknown, Equal but Unknown Variances

Finally, we consider testing for the difference in means, assuming equal but unknown variances. Our test statistic is the combination of (9) and (10):

$$t = \frac{\bar{x}_A - \bar{x}_B - \Delta\mu}{s_p}$$

where the sample means are $\bar{x}_A$ and $\bar{x}_B$ for the two data sets A and B. This is normalized by the **pooled variance**

$$s_p^2 = \frac{\sum_i (x_i - \bar{x}_A)^2 + \sum_i (x_i - \bar{x}_B)^2}{n_A + n_B - 2}\left[\frac{1}{n_A} + \frac{1}{n_B}\right]$$

where n_A and n_B are the number of data in datasets A and B. The statistic $\hat{t}$ is distributed as Student's t, with $n_A + n_B - 2$ degrees of freedom. Of course, this test, like the others above, assumes that the data are normally distributed.

Note that this test assumes that σ^2 is the same for both datasets, though we do not know it. While it would clearly be desirable to drop this restriction, doing so complicates the problem substantially—indeed, there is no test for this specific case (known as the Fisher-Behrens problem). But all is not lost, as we will now proceed to show.

6.3.4. A Nonparametric Test for Differences in Location

All the tests we have described so far assume some form for the pdf, and in most cases assume we know or can estimate parameters associated with that pdf. But it is possible to have tests that make no such assumptions; these are called **nonparametric** or **distribution-free** to indicate their independence from a specific pdf. One such test allows us to test whether or not two data sets can be described by the same pdf, whatever it is:

$$H_0: \quad \phi_A = \phi_B, \qquad \phi \text{ unknown}$$

This is about as general a test for equality of distributions between two data sets as we could ask for.

It is fairly obvious that if we had (say) 100 data values from set A, all falling between 0 and 1, and 100 values from B, all between 99 and 100, that they are very unlikely to be modelable as random variables from a single pdf: if the pdf peaked in these two regions, we would expect to get about 50:50 distribution of points in each region for each dataset—and if the pdf was nonzero anywhere else, we would expect to get some data outside these regions. What is important is that all the x_A's are below the x_B's, a behavior we can quantify using a **rank-sum test**, also called the **Mann-Whitney** or **Wilcoxon** test.

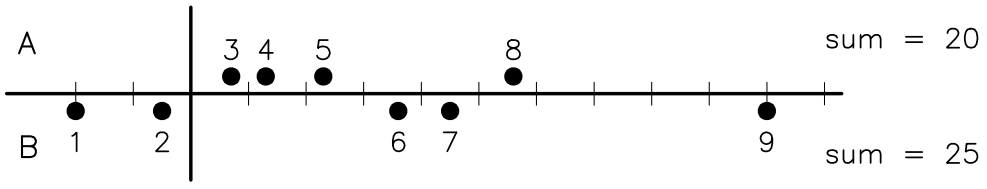

Figure 6.2

How this works is best shown with an example, and a figure (Figure 6.2). Suppose we have four data values in A (for which we use x's): $x_1 = 2.3$ $x_2 = 5.6$ $x_3 = 0.7$

$x_4 = 1.0$ and five data values in B (for which we use y's): $y_1 = -2.0$ $y_2 = 4.5$ $y_3 = 3.6$ $y_4 = 0.0$ $y_5 = 10.0$. Then, after sorting all the data together, we get the arrangement shown in the figure. The x's are the dots above the axis, and the y's the ones below it; the numbers next to each dot are the **ranks** of the data. The test statistic is formed by summing the rank values for one dataset; the sum for the other is related to it because both must sum to the sum of the first n integers, $\frac{1}{2}n(n+1)$ where $n = n_A + n_B$ is the total number of data. In the example the two rank sums are 20 for the x's and 25 for the y's.[4]

Given n and (say) the smaller of the rank sums, we can find the probability of getting this small a value or smaller, which becomes our significance level for a test of the hypothesis. If we denote the ranks by r_i, then the two statistics in common use are

$$S = \sum_{i=1}^{n_A} r_i \qquad \text{and} \qquad U = \sum_{i=1}^{n_A} r_i - i$$

where we have supposed A to have the smaller rank sum. For small n the distribution of these statistics is complicated, but for n larger than about 20 a good approximation is a normal distribution:

$$U \sim N(\mu_U, \sigma^2) \qquad S \sim N(\mu_S, \sigma^2)$$

where

$$\mu_U = \tfrac{1}{2} n_A n_B \qquad \mu_S = \tfrac{1}{2} n_A (n+1) \qquad \sigma^2 = \frac{n_A n_B (n+1)}{12}$$

This can be used to find significance levels for a one-sided test (one distribution less than another) or a two-sided one (one distribution different from another).

6.4. Testing for a Given PDF

For the tests above (the last, nonparametric, one aside), we assumed a normal distribution; can we test this assumption? Yes, this is just another hypothesis test, one to which we now turn. As we have seen, a lot of statistical theory assumes that we know the pdf. In spite of the central limit theorem it can be dangerous to assume that everything is Gaussian. A histogram of the values gives an empirical idea of the pdf. But we would like a more rigorous method of deciding whether a group of data can properly be modeled by random variables with a specified pdf. It turns out that there is a test statistic for this question, one that, somewhat surprisingly, does not depend on the underlying distribution having some particular form—though we do have to know what it is.

6.4.1. Kolmogorov-Smirnov Test

Suppose that we have a set of n numbers $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ and we want to test whether they can be modeled as independent random variables $\vec{X}$, each element of

[4] If values are tied, they all get the average of the ranks assigned to them; for example, if there were three identical values that had ranks 3, 4, and 5, they would each be assigned a rank of 12/3 = 4. Alternatively, if the data have only finite precision (that is, are not intrinsically integer), simply apply small random perturbations to apparently tied data, at a level one or two decimal places below the last significant digit.

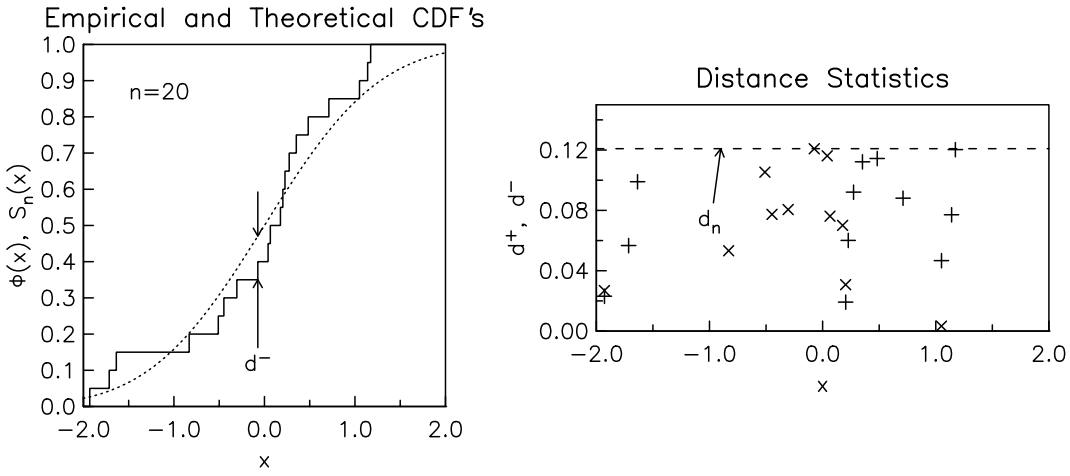

Figure 6.3

which is distributed with a pdf $\phi(x)$. There are a number of tests for this, which make use of the order statistics $(x_{(1)}, x_{(2)}, \ldots, x_{(N)})$ that we discussed in Section 5.1.2. From the order statistics we can create the **sample distribution function**, $S_n(x)$, which is a "stairstep" version of the cumulative distribution function; like the cdf it increases monotonically (though discontinuously) from 0 to 1. It is defined as

$$S_N(x) = \begin{cases} 0 & x < x_{(1)} \\ i/n & x_{(i)} < x < x_{(i+1)},\ i = 1, .., n-1 \\ 1 & x_{(n)} < x \end{cases}$$

if S_n is derived from n random variables with cdf Φ, the law of large numbers guarantees that as n gets larger and larger $S_n(x)$ approaches $\Phi(x)$. We want a means of deciding how different $S_n(x)$ and $\Phi(x)$ are; this is our test statistic for deciding if a data set can be represented by random variables with pdf ϕ.

At each value of i, we can define two distances between S_n and Φ: d^+ measured from Φ to the "top of the step", and d^- measured from Φ to the "bottom of the step". These are defined by

$$d^+(i) = \frac{i}{n} - \Phi(x_{(i)}) \qquad d^-(i) = \Phi(x_{(i)}) - \frac{i-1}{n}$$

Figure 6.3 shows a possible S_n and Φ in the left-hand panel, with one value of d^- indicated. The right-hand panel shows all the positive values of d^+ (pluses) and d^- (crosses). The **Kolmogorov statistic**, d, is the maximum deviation between S_n and Φ; it is given in two steps, first by taking the maximum value over all the d's:

$$d_n = \max_{1 \le i \le n} \left[d^+(i),\ d^-(i) \right] \tag{11}$$

as shown by the dashed line in the right-hand panel of Figure 6.3; and then by correcting for the value of n:

$$d_0 = \left[\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}} \right] d_n \tag{12}$$

which, for values of α small enough to be interesting, has the following expression for α:

$$\alpha = \mathcal{P}[d > d_0] = 2\exp(-2d_0^2)$$

This statistic is used in the **Kolmogorov-Smirnov** test for determining whether $\vec{x}$, our sample (supposedly modeled by $\vec{X}$) is in fact compatible with the underlying distribution $\Phi(x)$. If α is very small, then we are justified in rejecting the proposed distribution for X.

If in fact the null hypothesis H_0 is true, and the data can be represented by random variables with pdf ϕ, it turns out that the distribution of the K-S statistic, $\hat{d}$, is independent of the underlying distribution for X; that is, whatever Φ is, $\hat{d}$ will be distributed in the same way. This may seem difficult to believe. What may help to make it less so is to realize that, given any Φ, we could create any other Φ by stretching and shrinking the x-axis appropriately (since all cdf's are monotone functions). But such a transformation of the x axis has no effect on the maximum separation between S_n and Φ, as they transform together.

One disadvantage of the K-S test is that you need to know Φ beforehand. Often, we assume a type of pdf, and use the data to estimate parameters. This is not, strictly speaking, correct. But, it will have the effect that we will fit the data better than we would if we did not have this freedom. Thus our test will be conservative: if we reject the hypothesis, the actual level for rejection will be higher than what we compute.

It is also possible to apply this test to two sample distribution functions derived from different datasets, so as to test whether the two datasets can be modeled by random variables with the same distribution—and we do not need to know what that distribution function is. This is a very useful non-parametric test. The method is to form the same statistic as in (11) and (12), except that we take the difference between the two cdf's S_n and S_m (assuming n and m to be the number of data in the two datasets). For the n in equation (12), we take

$$n_e = \frac{nm}{n+m}$$

Another test for deviation is based on the **Kuiper statistic**, which is found from the d's as

$$v_n = \max_{1\le i\le n}\left[d^+(i))\right] + \max_{1\le i\le n}\left[(d^-(i)\right]$$

followed by a correction for n:

$$v_0 = \left[\sqrt{n} + 0.155 + \frac{0.24}{\sqrt{n}}\right] v_n$$

which, for values of α small enough to be interesting, has the following expression for α:

$$\alpha = \mathcal{P}[v > v_0] = (8v_0^2 - 2)\exp(-2v_0^2)$$

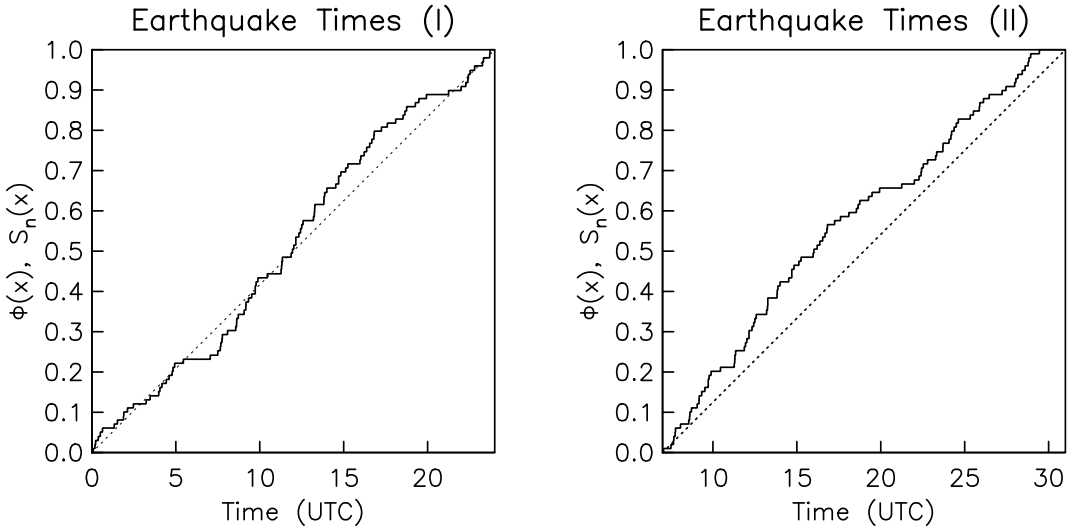

Figure 6.4

That is, we find the maximum deviation up and down separately, and sum them. The statistic $\hat{v}$ is sensitive to departures of S_n from Φ in different ways than $\hat{d}$ is. Most importantly, if we have data defined on a circle, $\hat{v}$ is invariant for different starting values of x, which the K-S statistic would not be. It is therefore suitable for testing, for example, if data are uniformly distributed around a circle or not. Figure 6.4 shows the comparison between Φ and S_n for the California earthquake dataset, assuming ϕ to be uniform, and taking two possible starting times. The K-S statistic d_n is not the same, but the decrease in d^- in going from a start time at 0^h to one at 7^h is exactly compensated for by the increase in d^+, leaving d_n unchanged. The α for this test and this dataset is 0.07—again, tantalizingly close to being "conventionally" small enough to reject the hypothesis of uniformity.

6.4.2. χ^2 Test for Goodness of Fit to $\Phi(x)$

Another widely used quantitative test for goodness of fit to a particular distribution is based on the chi-square statistic (not the same as the chi-square distribution, though of course closely related). This statistic is based on the histogram, and the idea that if we know the underlying distribution we can predict how many observations will be expected on average in each bin or cell of the histogram. This is well suited for problems in which observations naturally fall into discrete groups or cells, but is also widely used for continuous random variables. The general idea is to compare the number of observations that fall within a given cell or interval with the number to be expected for the theoretical probability distribution Φ. If the two numbers are close then Φ is a good model, if they are very different then one might have grounds for rejecting Φ as a model for the data. **Pearson's χ^2 statistic** is given by, for n cells

$$T = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i}$$

where o_i is the number of observations in cell i, and E_i is the number expected for

the theoretical distribution. It can be shown that, when the model is correct, the sampling distribution of $\hat{T}$ is approximately the χ^2 distribution, with m degrees of freedom, where $m = n - p - 1$ p being the number of independent parameters fitted. The approximation by a χ^2 distribution improves as the number of counts in each cell increases; less than 5 counts per cell is usually regarded as inadequate. Note, however, that the grouping of continuous data into cells discards useful information from the sample; unless you start with such "grouped" data, it is probably better to use a test for continuous random variables.

6.5. Probability Plots and Q-Q Plots

There are a couple of methods for examining the distribution of data—somewhat similar to the K-S test in their use of the empirical cdf, but designed more for qualitative assessment of how well a dataset agrees with a particular pdf. This is an example of what is called **exploratory data analysis**: these are graphical methods that are probably the first thing you should apply to a new dataset.

First, there is the **probability plot**, which is derived from the empirical and theoretical cdf's in Figure 6.3. Imagine warping the x-axis so that the function $\Phi(x)$ becomes a straight line, and plotting the points of the ordered data. The required transformation gives the quantiles of the distribution. If the quantile value is q, this gives the mapping

$$x(q) = \Phi^{-1}(q)$$

where Φ is the theoretical cdf. We can define a set of quantiles based on the number of data, n:

$$a_i = \Phi^{-1}\left(\frac{i - \frac{1}{2}}{n}\right) \; i = 1, \ldots, n$$

These quantiles, a_i, have the property that they divide the area under $\Phi(x)$ into $n+1$ areas, each exactly $1/(n+1)$. To make a probability plot, we plot the order statistics of the data, making n pairs a_i, $x_{(i)}$. If the pdf was a good description of the data these points would fall on a straight line. For the particular case in which the theoretical distribution is taken to be normal, a probability plot is useful for identifying long-tailed or short-tailed data distributions.

The two top panels of Figure 6.5 show probability plots for (what else?) our GPS data set. The left-hand plot shows the almost the full dataset; we have not shown all the data because then the plot shows almost only that there are a couple of very large outliers. Even with these omitted, the distribution is still long-tailed on both ends. The right-hand plot, for comparison, shows what the distribution is if we omit points beyond ± 0.25; removing points outside ± 0.2 leaves us with a distribution that appears impressively close to normal. Incidentally, the plot also allows us to estimate the mean (from the zero-intercept) and the standard deviation (from the intercept of a line through the data, evaluated for $\Phi^{-1}(x) = \pm 1$. The lower two panels show our other two often-used datasets. On the left, we have the intervals between reversals, plotted on the assumption that the intervals are exponentially distributed (as they would be for a Poisson process); we have to use log scales on both axes to make

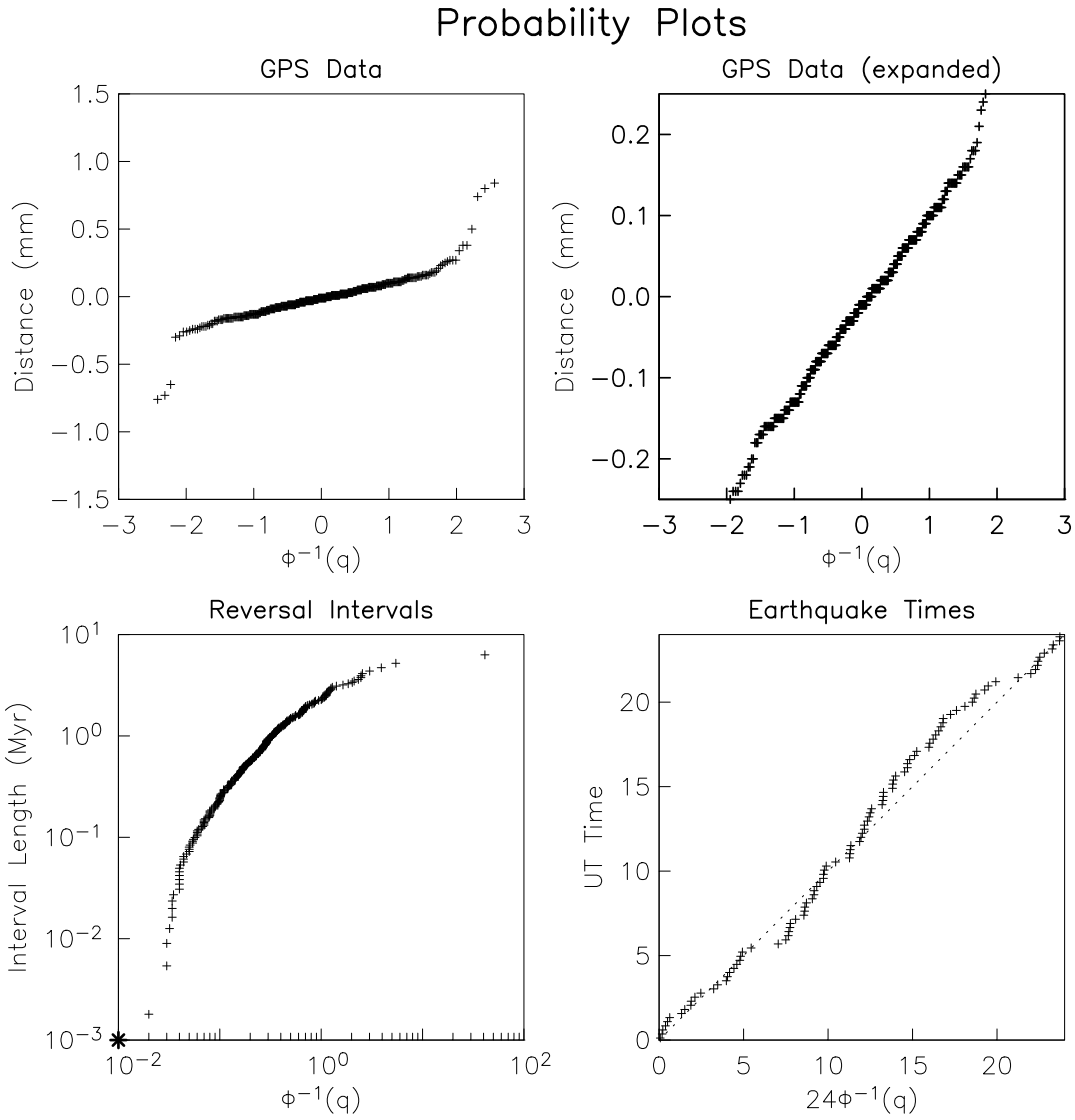

Figure 6.5

the plot readable. The data clearly do not follow a straight line, so we may be sure that this model is not adequate. The lower right-hand plot is for the times of earthquakes from Figure 6.1, assuming a uniform distribution; this is of course very much the same as the Kuiper-statistic plot in Figure 6.4. The event times approximate uniformity, but there are more times around 5 hours UT than would be expected, and fewer from 10 to 20 hours.

This kind of plot can be extended to two datasets, using what is called a **Q-Q plot** (for quantile-quantile). Suppose we have two sets of ordered data, of size n and m:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n-1)}, x_{(n)} \quad \text{and} \quad y_{(1)}, y_{(2)}, \ldots, y_{(m-1)}, y_{(m)}$$

If $n = m$, then we simply plot the ordered pairs $x_{(i)}, y_{(i)}$. If not, we have to interpolate to get the quantiles, which can be done as follows. Take the quantile $0 \le q \le 1$, and

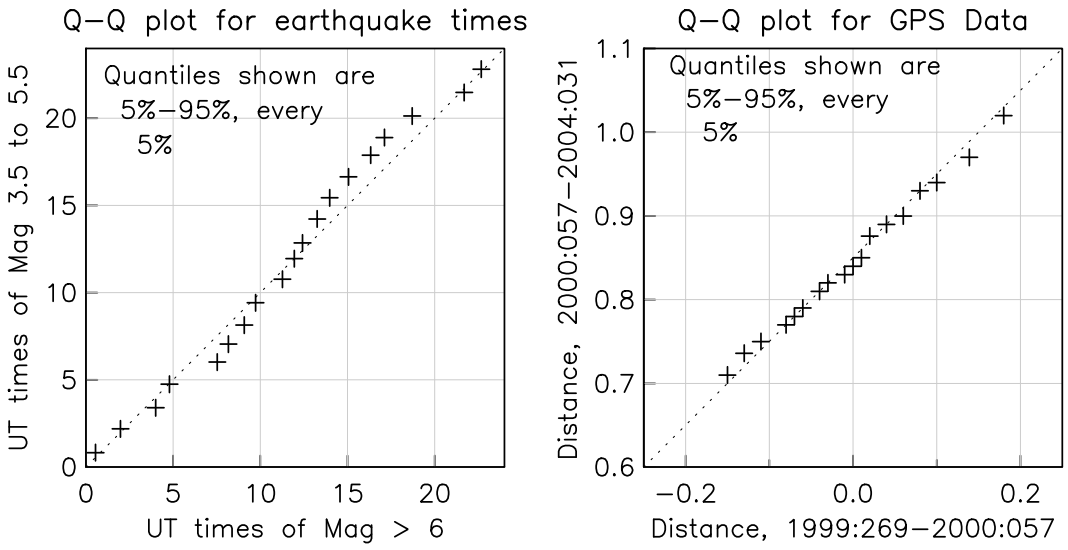

Figure 6.6

from it find the values $r = qn + 0.5$ and $s = qm + 0.5$; truncate these to get the integers k and l, and find the fractional parts $e = r - k$ and $f = s - l$. Then we can create data values that are (approximately) associated with this quantile:

$$x(q) = (1-e)x_{(k)} + ex_{(k+1)} \quad \text{and} \quad y(q) = (1-f)y_{(l)} + fx_{(l+1)}$$

which we can evaluate for a selected set of q's. (It is usually best to use a finer sampling near 0 and 1, and a coarser sampling near 0.5). Note that if n and m are equal, the simplest approach is to take q to be a multiple of n^{-1}, in which case we simply end up with ordered pairs $(x_{(i)}, y_{(i)})$.

Now, plot the quantile values of x and y against each other. If the two data sets have the same distribution, the plotted points will lie along the line $y = x$. Shifts in the mean will move the points right or left from this line; differences in the variance will change the slope away from 1. As with the probability plot, the advantage of the Q-Q plot is that it shows the behavior over the full range of the data, not merely a few summary values. And, it is completely independent of any assumptions at all about what some "underlying" distribution is.

Figure 6.6 shows a couple of examples, comparing a couple of our "standard" datasets with closely related ones. On the left, showing the quantiles for times of large earthquakes against those of smaller ones;[5] the smaller ones have times that are nearly uniformly distributed, so the plot looks very much like the probability plot in Figure 6.5. The GPS data have been compared with data for a later time span; the figure shows a clear shift in location, with the later data (along the y-axis) having just slightly less variance than the earlier data, though a very similar pdf. Since the limiting quantiles are 0.05 and 0.95, this plot cannot show any long-tailed behavior, should this be present.

[5] The smaller earthquakes are those between magnitude 3.5 and 5.4 in the Southern California earthquake catalog between 1981.0 and 2003.5, omitting days with 5.5 and larger shocks.

6.6. Testing for Correlation between Random Variables

The last tests we discuss are to test the hypothesis that there is or is not a correlation between two sets of random variables, $\vec{X}$ and $\vec{Y}$, each with n elements. One diagnostic statistic is the size of ρ, the correlation coefficient. The standard estimate of ρ for n pairs of numbers (x_i, y_i) is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{13}$$

with $\bar{x}$ and $\bar{y}$ being the mean of the x_i's and y_i's respectively. If the variables X and Y are jointly normally distributed, then the standard deviation of r is

$$\sigma_r = \frac{1-r^2}{\sqrt{(n-1)}}$$

and $-1 < r < 1$. We want to decide if r is significantly different from $r = 0$, the case of no correlation. This is done using

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \tag{14}$$

which is t-distributed with $N-2$ degrees of freedom. This is perhaps the most abused test in all of statistics, since it assumes that the data can be modeled by random variables with a bivariate normal distribution, an assumption that is often overlooked by those who use it—sometimes, as we saw in Chapter 1, with deplorable results.

A more general test for correlation that does not rely on this assumption can be gotten by replacing the data with their ranks, and then computing the **Spearman rank-order correlation coefficient**.

This test is almost exactly the same as the previous test, except that we replace the n pairs of values (x_i, y_i) by their ranks, to form pairs $(r_i^x, r_i^{y)}$. Then we find

$$r_s = \frac{\sum_{i=1}^{n}(r_i^x - \bar{r}^x)(r_i^y - \bar{r}^y)}{\sqrt{\sum_{i=1}^{n}(r_i^x - \bar{r}^x)^2}\sqrt{\sum_{i=1}^{n}(r_i^y - \bar{r}^y)^2}} \tag{15}$$

with, for example, $\bar{r}^x$ being the mean of the ranks for the x's. But since the sum of the ranks is just the sum over the first n integers, this is the same for both x and y, as are the sums in the denominator. If we make use of

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \qquad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

we find that the denominator is

$$\frac{n(n^2-1)}{12}$$

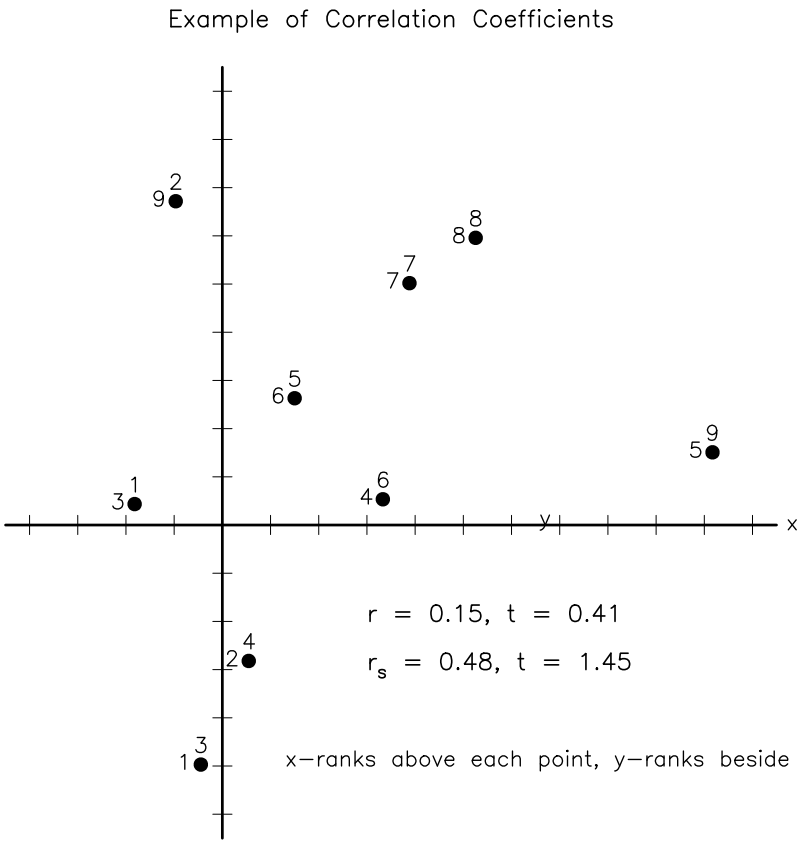

Figure 6.7

We can simplify (15) even further if we sort the pairs so the y values are in increasing order, so that $r_i^y = i$, as illustrated in Figure 6.7. Then the numerator becomes

$$-\left(\frac{(n+1)}{2}\right)^2 + \sum_{k=1}^{n} k r_k^x$$

and the total expression can be written as

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{k=1}^{n} (r_k^x - k)^2$$

which, like r, is between –1 and 1; it will reach a value of 1 only if the ranks for the x's and y's are the same. Continuing in parallel with (14), the statistic

$$t = \frac{r_s \sqrt{(n-2)}}{\sqrt{(1-r_s^2)}}$$

is also approximately distributed as Student's t with $n-2$ degrees of freedom, the approximation being adequate for $n \geq 30$; for smaller n, an exact expression for $\phi(r_s)$ is available. Given the ease with which actual data can violate bivariate normality, it is probably wise to begin with this test if you want to test for correlation.

6.7. The Neyman-Pearson Framework for Hypothesis Testing

We finish with what we might have begun with, which is a sketch of the Neyman-Pearson approach to testing. The formal procedure is not often used in geophysics, but it underlies many discussions of testing, and provides, as the ideas of efficiently and bias did for estimators, a framework for comparing tests.

The Neyman-Pearson approach explicitly frames the test as one between two hypotheses, the null hypothesis H_0 and an **alternative hypothesis**, H_1, that we are said to be testing H_0 against. A simple example would be if we had data modeled by random variables that are normally distributed with known variance and a mean that is either μ_1 or μ_2; the null hypothesis H_0 could be $\mu = \mu_1$, and our test would be against the alternative hypothesis, H_1, that $\mu = \mu_2$.

Although this is a very different approach than the significance testing we have discussed up to now, much of the formal procedure is the same: we decide whether to reject H_0 in favor of H_1 on the basis of a test statistic $t = T(\vec{x})$, using the distribution of the test statistic $\hat{t} = T(\vec{X})$, where the distribution of the random variables $\vec{X}$ is part of the null hypothesis H_0. The set of values for which H_0 is accepted and rejected are, respectively, the **acceptance region** and **rejection regions** of the test. And, exactly what ranges of the parameters these regions cover depends on the value of α, the significance level, which in this framework is always chosen beforehand, at least implicitly.

If we think of testing two hypotheses we can see that we can have two kinds of error:

Type I error: we may reject the null hypothesis H_0 even though it is valid. For our earthquake-time problem, this would be deciding that the distribution of times is nonuniform even though it is actually uniform. Of course, we expect this to happen if we could do the test many times; it should happen a fraction α of the time, where α is the significance level. The probability of a Type I error is therefore just α—and we can in principle choose this to be as small as we like. For the more complicated case in which H_0 is composite, the probability of a Type I error generally depends on which particular member of H_0 (that is, which parameter) we choose, and the significance level is defined to be the maximum of these probabilities.

Type II error. This is where we accept H_0 even though it is false. For our earthquake-time problem, this would be deciding that the times were uniformly distributed even though they were in fact distributed according to H_1. The probability of this occurring is denoted by β. We are probably more interested in the reverse, the probability that H_1 is rejected when it is false; this quantity, $1 - \beta$, is called the **power** of the test. Clearly we want β to be as small, and the power as large, as possible: an ideal test would have a power of one, so we would always reject a false H_0. If H_1 is composite then β depends on the particular parameters of H_1.

Thus to compare tests we can ask which one, for a given α, has the smallest β—that is, is the more powerful. Ideally we could have a power of 1 with $\alpha = 0$; in practice this can never be achieved. Also for any given number of data, n, it is always true decreasing α, will increase β. As indicated above, usually we by fixing

the significance level in advance (at a rather small number, typically 0.05 or .01), and then trying to find a test yielding a small value for β.

Given a fixed α and n, β will depend on the test procedure, and so comparison of powers gives us a means of comparing tests.

You should realize that the power can depend, not only on the nature of the test, but also on the alternative hypothesis H_1, which is usually called what the test is "testing against". For example, the Schuster test is most powerful when testing the hypothesis H_0 (a uniform distribution) against H_1, when H_1 is that the pdf for the times is unimodal (a single peak); it is not difficult to see that this test would do a poorer job of discriminating between a uniform distribution and one with two peaks 12 hours apart. This can be quantified by keeping H_0, α, and n the same, and comparing β for different H_1. For some tests, H_1 can be "anything other than H_0"; this is true of the Kuiper test described above, which tests a uniform distribution on the circle against any alternative. What we lose by employing such a general test is likely to be that, for a given β, α will be larger than it would be for a test against a more specific H_1. This is quite similar to the tradeoff experienced with estimators: a estimator that works well for a wide range of pdf's will be less efficient than a test designed around a specific pdf, if that pdf is appropriate (think of the mean and the median). And the same thinking can be appropriate: we may be willing to sacrifice some power if the test is usable over a wider range of alternative hypotheses.