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.

6.1 Introduction

We now turn from estimating parameters of probability density functions, to testing **statistical hypotheses**. In general scientific usage, a hypothesis is some assertion we make about the way the world is. A statistical hypothesis is more restricted, being an assertion about how a dataset relates to some kind of probability model. We can test either kind, but the tests for statistical hypotheses are more formalized.

Here are examples of scientific hypotheses and the statistical hypotheses they are related to:

1. As described in Section 1.2, we may hypothesize that there was a change in the core dynamo between the time of the Cretaceous Superchron and the subsequent period of frequent reversals. A statistical

hypothesis that formalizes this is that a point process that fits 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.

2. We may hypothesize that earthquakes are triggered by earth tides. The statistical hypothesis to go with this would be that more often than not earthquakes happen at times related to (say) high and low tides – as opposed to occurring "at random" relative to the tides.

3. 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.

For each example, we start by formulating a probability model for our scientific hypothesis; how to do this is not statistical analysis, but requires informed judgment, both about the particular problem and about the methods available for deciding if a probability model agrees with the data or not. In this chapter we discuss some general principles about testing statistical hypotheses, and also present some of the most frequently used tests.

Statistical tests can keep us from a common error caused by the normal human propensity is to find patterns even when there are none. A test can show that what we have observed does not indicate some regular behavior, but might well have happened by chance.

6.2 Problems and Caveats

We start with some general remarks about this branch of statistics. Thanks to the range of questions we may try to answer there are many different hypothesis tests. But some of their diversity arises from 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 subjects: what is appropriate in an economic context (where costs and benefits are clear) is less so when deciding between scientific theories. We shall select what seems most useful while admitting that it may have less of a logical basis than we would like.

There are two general approaches to hypothesis testing:

- The procedures developed by R. A. Fisher, which use 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.

- The **Neyman-Pearson** approach, which seeks 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 approach combines parts of both the Fisher and Neyman-Pearson procedures to produce a methodology that is easier to describe, though not logically consistent. But it does satisfy the aim of inferring no less and (especially) no more from the data than we should.

6.3 A Framework for Tests

If we have a known pdf ϕ that describes a random variable X, and also know that X is an appropriate model for our data, we would know all that we could, statistically speaking. For example we might know that the data are modeled by Normal random variables with known mean and variance. Chapter 5 was devoted to procedures for finding the "best values" of a pdf's parameters. **Hypothesis testing** is about testing statements about the pdf of the random variables we use to model the data. One kind of statement (fundamental to the Neyman-Pearson approach) is which of two statistical hypotheses we should choose. 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 procedures for hypothesis testing are the same either way, and in fact resemble the procedure used to find how good an estimate is. As mentioned above, the first step is deciding what statistical hypothesis we want to create that will help us evaluate some more general hypothesis – and this is a matter of judgment. Often, what we need to do is set up a statistical hypothesis contrary to what we want to show is true; this is called the **null hypothesis**, conventionally denoted as H_0. Whatever hypothesis we choose includes the stipulation that the data can be modeled by a specified kind of random variable. But in hypothesis testing we can allow much more general pdf's than we can in estimation; for example, in the class of tests called "distribution-free" we assume only that the rv's that model the data come from a pdf, otherwise unspecified: about as general an assumption as we could ask for.

Having set up a statistical model we take the following steps:

- We use some procedure to compute a **test statistic** $T(\vec{x})$ from the data $\vec{x}$; that is, we take the data and produce some number (or numbers), analogous to finding an estimate.

- The null hypothesis H_0 assumes that the data can be modeled by rv's X with a known pdf. Using this assumption, find the pdf of the rv produced by applying the test procedure to the assumed rv's. We call this rv $\hat{T} = T(\vec{X})$ with pdf $\phi(t)$.

- From $\phi(t)$, compute

$$\alpha = \int_{-\infty}^{T=-T(\vec{x})} \phi(t)\,dt + \int_{T=T(\vec{x})}^{\infty} \phi(t)\,dt \tag{6.1}$$

  This integral is the area under the tails of the pdf of $\hat{T}$. (In some cases, we use the area under only one tail) The tail is the part of the pdf for which t, the argument of the pdf, is greater in magnitude than the value $T(\vec{x})$ found from the data. This area, 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**. The value of α is the end result; what we do with it is described in Section 6.3.2 below.

6.3.1 An Example: Testing Earthquake Times

To show how testing works, we perform a test on the earthquake data shown in Section 1.2. That data was introduced in the context of a claim that California earthquakes tend to occur early in the morning, The basis

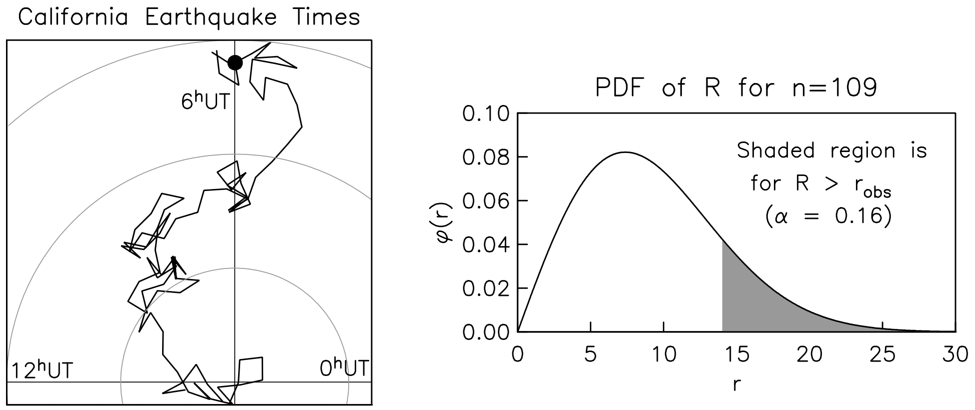

Figure 6.1: The left panel shows the summing of unit vectors for the computation of the Schuster test value, for all earthquakes magnitude 6 and over from 1910 through 2011. The right-hand panel shows the pdf of the distribution of R for the number of data used; the shaded area is the probability of getting the value observed, or more, if the directions of the vectors are uniformly distributed.

for this claim was five events, three of them well-known because damaging: this is just a few anecdotal cases, not a systematic view. What we think we know about how earthquakes work suggests no way that large earthquakes would correlate with local time – so it is therefore reasonable to think that our anecdotal evidence is a mix of coincidence and selection bias. If we look at a much larger set of earthquakes, how do we test the hypothesis that temporal clustering is observed?

In this case our null hypothesis is obviously that there is no clustering, which is to say that the times of earthquakes are uniformly distributed: that is, that the random variable modeling the times of earthquakes has a uniform pdf over [0, 24) (in hours). More formally, if X is the time of earthquakes, we express the null hypothesis by writing $H_0 : X \sim U(x)$.

One possible test statistic[1] is given from the n observed times by first finding

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

and then computing

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

[1] As we will see in Section 6.5.1, this is not the only one.

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 the times are clustered, the bigger R will be. The left panel of Figure 6.1 shows this procedure applied to all large earthquakes in California since 1890, with the individual unit vectors shown head-to-tail; the large dot is the sum, which turns out to have $r_{obs} = 13.90$.

Geophysicists usually call this procedure the **Schuster test**, after the person who introduced it for this very problem; statisticians more often call it the Rayleigh test. For n large, and the X_i's uniformly distributed (our null hypothesis), the pdf of R is

$$\phi(r) = \frac{2r}{ne^{-r^2/n}} \tag{6.2}$$

which is shown on the right panel of Figure 6.1; this is just the Rayleigh distribution of Section 3.5.5. The shaded region shows the part of the pdf for which $r > r_{obs}$; the probability of observing R in this region (supposing the null hypothesis) is $\alpha = 0.16$. The complementary value gives $1-\alpha = 0.84$ for $r < r_{obs}$.

For completeness, if $n < 50$ a better approximation to α is

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

where $z = R^2/n$. For n large this becomes $\alpha = e^{-z}$, the same result as from integrating 6.2 from r to infinity.

6.3.2 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, 9% of the time; we say that we have a significance level of 0.09.

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

1. Report the value of α to summarize what we got: suggestive, perhaps (one chance in six is not that likely), but not conclusive.

2. Take some small value α_0, and say that, since $\alpha > \alpha_0$, we cannot reject the null hypothesis. That is, it is reasonable to say that what we

see could just be chance, and the pattern we started with is just coincidence. If we are going to follow this course, we need to set α_0 before we compute α for the data.

3. Make the stronger statement that because $\alpha > \alpha_0$, the clustering hypothesis is false. The converse, and more common, approach in hypothesis testing comes when we are looking for something more interesting than the null hypothesis; then the stronger statement would be the claim that the alternative hypothesis is true if $\alpha \le \alpha_0$.

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. For example, in an industrial setting we might be using a test to evaluate the quality of our manufacturing, and if $\alpha \le \alpha_0$ we might stop a production line or reject a batch of products. It is not so clear what this approach would entail in doing research: possibilities include deciding what additional data to collect or what other ideas to consider.

We have laid out these options because (in our view) more than one is acceptable, and some are both popular and unwise. All too often, a conventional values for $1-\alpha_0$ is chosen (usually 0.95 or 0.99) and then only option (3) is taken, by 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 interpretation of a hypothesis test is unwise. As a form of words it may be acceptable to say that a hypothesis has been rejected (option 3), but we should realize that

- There is nothing special about a particular value of α_0. In particular, to view $1-\alpha = 0.94$ (say) as being a very different outcome from $1-\alpha = 0.96$ is nonsense.

- Likewise, the value of α for different hypothesese does not bear a clear relationship to how strong the effect assumed by each hypothesis actually is.

- We should not confuse (A) having shown that the data do or do not support a hypothesis at some level, with (B) having proved anything about its truth (option 3) – such a result simply makes a strong case – though 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 absent from geophysics) of declaring that if results did not reach some level of significance, they should not be published. This is much too rigid a way to apply statistics to scientific inference.

For our test on the times of large California earthquakes, the best we can say is that there is not a strong case for temporal clustering. But it is important to note that we cannot say it is false; since the strength of the test depends on the sample size n, it may be that an effect is present, but at too small a level for this test to show it convincingly.

6.3.3 The Perils of Going Fishing

This is an appropriate place to discuss another error often made in testing, especially when (as in geophysics) it is impossible to get more data by doing experiments. It is natural to look for patterns in the data, and, having found one, perform a test of its significance. But then the test is meaningless. Our search for a pattern (a "fishing expedition") amounts to a series of prior tests, and so violates the assumption that the significance level can be treated as though we tested only one set of data one time.

An example[2] may make this clearer. It is possible that earthquakes occur more often at times when the stress from earth tides favors the actual faulting. We collect data on this faulting ("source mechanisms") for many earthquakes, and apply a test (modified from the Schuster test) to see if the data support our hypothesis or not. Since we know so little about earthquake triggering, it seems reasonable to assume that different modes of faulting might react to stress changes differently; also, we do not know which part of the stress tensor might be responsible. So we try the test for different types of stress and different types of faulting, ending up with 12 possible combinations. For one of these our test gives $\alpha = 0.04$; since a standard value for α_0 is 0.05, we 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; hardly unlikely.

[2] This comes from *Heaton* [1982], which updates, corrects, and apologizes for *Heaton* [1975].

There is nothing wrong with using tests to go fishing for a possible result, so long as we do not claim that whatever result we get is in fact significant. The order in which we try different things matters: 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]

What do we do if we cannot collect more data? One solution is to divide the data, in advance and at random, into two sets: one for fishing in, and one for testing what we find. 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 (0.04/12).

6.4 Specific Tests I: Do Means Differ?

We now put aside the philosophical complexities of statistical testing, and discuss, in a more "cookbook" way, some common tests. We start with the simplest kind: tests for differences in means, followed by tests of whether data conform to a particular pdf. Tests involving variances will be discussed later, in the context of least-squares fitting.

6.4.1 Means and Variances Known

We start with a slightly artificial case, in which the null hypothesis H_0 is that all the data come from a normal distribution with specified mean μ_0 and variance σ^2; we test the part of H_0 that is about the mean. This test would not be common in geophysics, but can easily arise in other settings: for example, in asking if some property of a set of manufactured items is within a specified tolerance. The usual shorthand notation for this test is

$$H_0: \qquad \mu = \mu_0$$

Such a statistical hypothesis, in which all the parameters of the pdf are known, is called a **simple hypothesis**. When we used the Schuster test in Section 6.3.1 we were also testing a simple hypothesis, since the pdf for the null hypothesis had all parameters specified – which is to say, none, since

[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.

the pdf was completely specified by the statement that it was uniform over $[0,24)$.

For this H_0, the value we get from the data is the difference between the sample mean and the and the assumed mean: $\bar{x}-\mu_0$. So the test statistic is

$$T(\vec{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu_0 \tag{6.3}$$

We showed in Section 5.3.1 that the pdf of T, for random variables X_i from the assumed normal distribution, would be $T \sim \mathrm{N}(0,\sigma^2/n)$:

$$\phi(t) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} e^{-nt^2/2\sigma^2} \tag{6.4}$$

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 6.1):

$$\alpha_0 = \int_{-\infty}^{-t_0} \phi(t)\,dt + \int_{t_0}^{\infty} \phi(t)\,dt \tag{6.5}$$

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 6.3 – or rather, in terms of its inverse, $\Phi^{-1}(\alpha)$; equation 6.5 will be satisfied for

$$t_0 = \Phi^{-1}(\alpha_0/2) \tag{6.6}$$

where the $\alpha_0/2$ comes from the inclusion of both tails of the pdf in 6.4. H_0 would thus be rejected, with a confidence of $1-\alpha_0$, if

$$|\bar{x}-\mu_0| \;>=\; t_0 \tag{6.7}$$

We naturally think of $\bar{x}$ as being fixed and μ_0 varying; but we can also view equation (6.7) as saying that H_0 should be rejected if μ_0 fell outside the $1-\alpha$ confidence interval for $\bar{x}$. This interpretation follows from this interval being

$$[\bar{x}+\Phi^{-1}(\alpha/2),\ \bar{x}+\Phi^{-1}(1-\alpha/2)] \tag{6.8}$$

This is an example of a more general result, namely that there is a close relationship between confidence intervals on a statistic, and a test applied to that statistic. Since both specify intervals within which a pdf integrates to a specified amount of probability, they have equivalent limits, though these limits are used differently.

We can easily extend our treatment to the case of two data sets, which we call $\vec{x}_A$ and $\vec{x}_B$, with assumed variances σ^2 for both, and assumed means that differ by $\Delta\mu$. To test if the difference in sample means is in fact $\Delta\mu$, we compute

$$t = (\bar{x}_A - \bar{x}_B) - (\Delta\mu) \tag{6.9}$$

which would reduces to $(\bar{x}_A - \bar{x}_B)$ if we are testing to see if the means are equal. The test statistic T, assuming a normal distribution for both data sets, is the convolution of the two distributions for $\hat{\mu}_A$ and $\hat{\mu}_B$. So T is distributed as $\mathrm{N}(0, \sigma^2(n_A^{-1} + n_b^{-1}))$. If we use the cdf $\Phi(t)$ appropriate to this distribution, the critical value for the test, t_0 is again given by equation 6.6, with equations 6.7 and 6.8 following as before. This is sometimes called the **z-test**.

6.4.2 Testing Against a Known Mean, with Unknown Variance

A more interesting case is $H_0: \quad \mu = \mu_0$ with σ^2 unknown. Because this involves additional unknown parameters, it is called a **composite hypothesis**. The test parameter is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{6.10}$$

where s^2 is the sample variance defined in Section 5.3.1; as we showed there, the statistic T, which is

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

is the ratio between an rv with a normal distribution, and an rv distributed as the square root of a $\chi_n - 1^2$ 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 6.6. (By now we hope you appreciate that all these examples have the same basic structure; only the pdf changes).

6.4.3 Means Unknown, Equal but Unknown Variances

Next we consider testing if two datasets have different means, assuming (as usual) normal distributions, and unknown (but equal) variances. Our test parameter combines 6.9 and 6.10:

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

where $\bar{x}_A$ and $\bar{x}_B$ are the sample means for the data sets A and B. The difference (minus any assumed difference) is normalized by the **pooled variance**

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

where n_A and n_B are the number of data in datasets A and B. The corresponding statistic T is then distributed as Student's t with $n_A + n_B - 2$ degrees of freedom.

This is clearly the most general of the tests we have seen so far, since it requires only that σ^2 is the same for both datasets, not that we know it. Unhappily, if we drop this assumption, and allow the variances to be unknown and different, the problem becomes much more complicated – indeed, there is no test for this specific case (called the **Fisher-Behrens problem**). But all is not lost, as we will now proceed to show.

6.4.4 A Nonparametric Test for Differences in Location

All the tests we have described so far assume some form (usually Normal) for the pdf, and in most cases also assume that we know or can estimate the parameters associated with that pdf. But there are tests that make no such assumptions; these are called **non-parametric** or **distribution-free** to indicate that they are independent of a specific pdf.

One such test, which includes testing for differences in means, tests whether or not two data sets came from the same (unspecified) pdf: H_0 : $\phi_A = \phi_B$ for ϕ unknown. This is about as general a test for equality of parameters between two data sets as we could ask for.

How can we do this? Suppose 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. How likely is it that they can be modeled as as random variables from a single pdf? We can imagine a pdf peaked in these two regions – but then we would expect each dataset to include about equal numbers from each region. And if the pdf was nonzero anywhere else, we would expect to get some data outside these regions. So, we can say it is very unlikely.

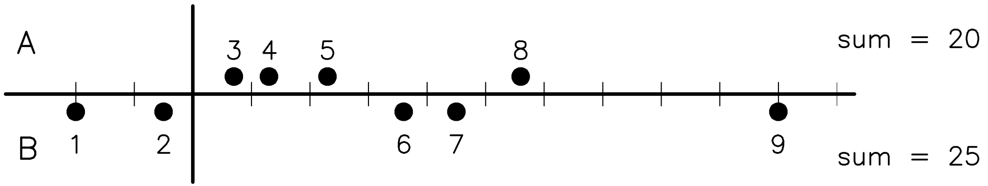

Figure 6.2: Two data sets A and B, indicated by the dots. All dots are numbered by their rank, and the sums for each set are given at the right.

What is important in this reasoning is that all the x_A's are smaller than the x_B's. We can quantify this behavior using a **rank-sum test** (other names are the **Mann-Whitney** or **Wilcoxon** test). How this works is shown in 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, found by sorting. To form the test parameter we sum the ranks for one dataset; this automatically gives the sum for the other because these must add up to the sum of the first n integers, $halfn(n+1)$ where $n = n_A + n_B$ is the total number of data. The figure shows rank sums of 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 pdf of these statistics is complicated, but for n larger than about 20 a good approximation is (what else?) a normal distribution:

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

[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.

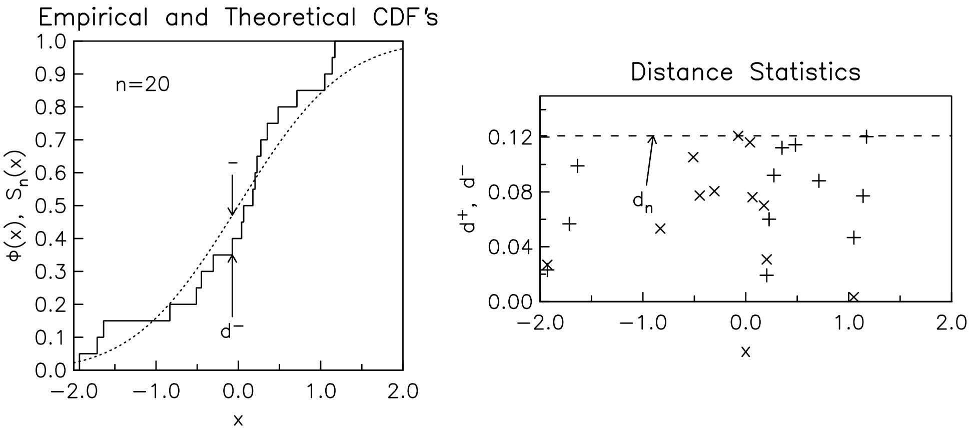

Figure 6.3: Illustration of the computation of the Kolmogorov-Smirnov statistic.

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 pdf 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).

6.5 Specific Tests II: Do We Know the PDF?

The last, nonparametric, test aside, all the tests above assume that the data can be modeled by rv's with a normal distribution. But we have seen that it can be dangerous to assume this, even though the central limit theorem suggests that we can. Can we test if data can be modeled as Normal – or more generally, test the assumption that our data can be modeled by rv's with some specified pdf? We can, and it is this kind of hypothesis test that we now discuss; since much statistical theory depends on getting the model pdf right, this kind of test is very important. The test statistic for this question, somewhat surprisingly, does not depend on the underlying distribution having some particular form – though we do, of course, have to specify the form to make the test.

6.5.1 Kolmogorov-Smirnov Test

Suppose that we have n data values $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ and we want to test whether they can be modeled as independent random variables $\vec{X}$, each element of which has a pdf $\phi(x)$. A number of tests for this use the empirical cdf, $S_n(x)$, that we described in Section 2.6.2: a stairstep function that increases monotonically (though discontinuously) from zero to one. If S_n is derived from n random variables with cdf Φ, the law of large numbers guarantees that as n approaches infinity $S_n(x)$ approaches $\Phi(x)$.

We want a test statistic for deciding how different $S_n(x)$ and $\Phi(x)$ are; the value of this statistic determines if the data are consistent with a model using iid rv's pdf ϕ. At each step (indexed by i), we 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":

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

The left-hand panel of Figure 6.3 shows a possible S_n and Φ, 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_0, is the maximum deviation between S_n and Φ; it is computed in two steps. First, take the maximum value over all the d's

$$d_n = \max_{1 \le i \le n} \left[d^+(i), d^-(i) \right] \tag{6.11}$$

as shown by the dashed line in the right-hand panel of Figure 6.3. Second, correct for the value of n:

$$d_0 = \left[\frac{\sqrt{n} + 0.12 + 0.11}{\sqrt{n}} \right] d_n \tag{6.12}$$

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

$$\alpha = \Pr[d > d_0] = 2\exp(-2d_0^2)$$

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 model distribution $\Phi(x)$; this is the null hypothesis H_0. As usual, we argue that if α is very small, then we are justified in believing that the $\phi(x)$ would be very unlikely to produce rv's with the distribution shown by the data, and hence in rejecting ϕ as a suitable pdf for modeling $\vec{x}$. If α is not too small, then the pdf cannot be ruled out.

If in fact H_0 is true (ϕ is the correct pdf to model the data with), we can find the distribution of the K-S statistic, $\hat{d}$. Perhaps surprisingly, this distribution is independent of the form of Φ (and hence of ϕ). To make this result more reasonable, we note that, since all cdf's are monotone functions, we can create any Φ from any other Φ by stretching and shrinking the x-axis appropriately. But such alterations of the x axis have no effect on the maximum separation between S_n and Φ, as they transform together.

One disadvantage of the K-S test is that we need to know Φ beforehand. More often, we assume that the data can be modeled by some particular pdf (perhaps chosen from the collection in Chapter 3), but estimate the parameters of the pdf from the data. This is not, strictly speaking, consistent with the assumption of the K-S test, and it means that our pdf will be more consistent with the data than it would if we did not estimate the parameters. But this means that 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 in this test we do not even need to know what that distribution function is, so this is a non-parametric test. The method is to form the same statistic as in 6.11 and 6.12, 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 6.12, 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

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

followed by a correction for n:

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

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

$$\alpha = \Pr[v > v_0] = (8v_0^2 - 2)\exp(-2v_0^2)$$

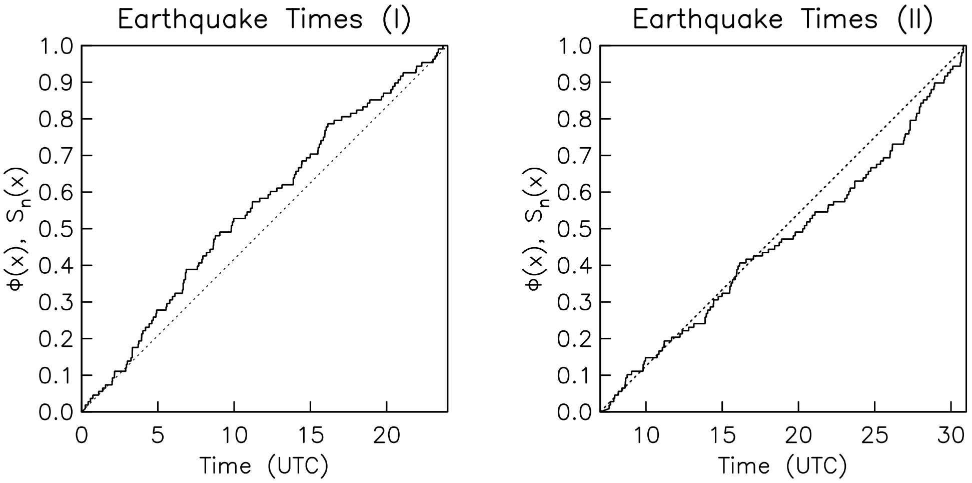

Figure 6.4: Empirical cdf's for the California earthquake dataset used in Figure 6.1, plotted against the expected cdf for a uniform distribution, for two different choices of where to start the S_n – which may be anywhere because this is defined on a circle.

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 v_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.

6.5.2 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 – but should not be, given that the tests we just described do not require binning.

The general idea of this test 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; fewer than five counts per cell is usually regarded as inadequate.

6.6 Specific Tests III: Are Variables Correlated?

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

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6.13}$$

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

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \tag{6.14}$$

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 Section 1.4, 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

$$r_s = \frac{\sum_{i=1}^{n}(r_i^x - \bar{r}^x)(r_i^y - \bar{r}^y)}{\sqrt{\sum_{i=1}^{n}(r_i^x - \bar{r}^x)^2}\sqrt{\sum_{i=1}^{n}(r_i^y - \bar{r}^y)^2}} \tag{6.15}$$

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

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

We can simplify 6.15 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 6.5. 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

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

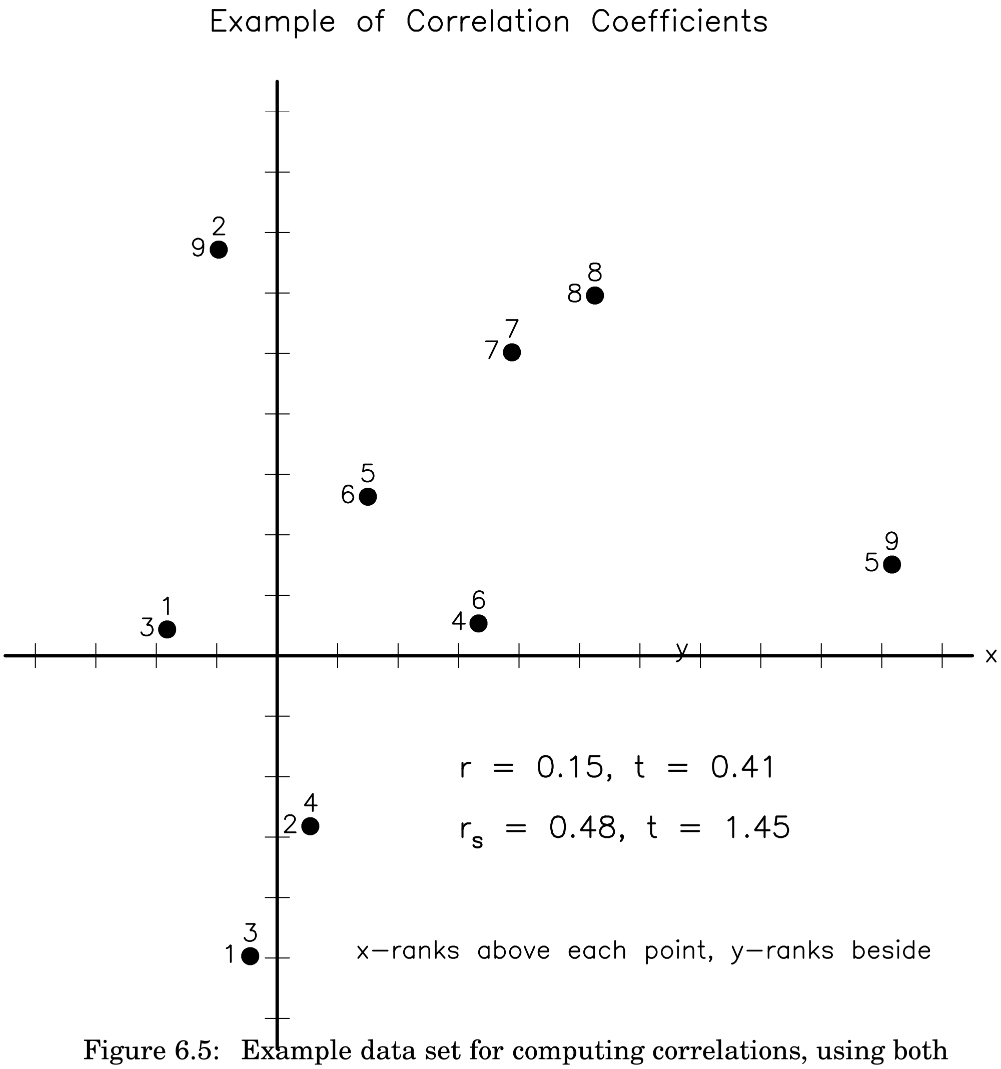

Figure 6.5: Example data set for computing correlations, using both ranks and values.

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 6.14, 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 $nge30$; for smaller n, an exact expression for $\phi(r_s)$ is available. Given how often actual data depart from bivariate normality, you should begin with this test if you want to test for correlation.

6.7 The Neyman-Pearson Approach to 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:

1. **Type I error**: we may reject the null hypothesis H_0 even though it is valid. For our earthquake example, this error would be deciding that the distribution of times is nonuniform even though it is actually uniform. Because of the significance level, we expect exactly this to happen if we do the test many times; it should happen a fraction α of the time. 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.

2. **Type II error**. This is where we accept H_0 even though it is false. For our earthquake problem, this error 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 that decreasing α, will increase β. As indicated above, usually we fix the significance level in advance at a rather small number (typically 0.05 or .01), and then try 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 α will be larger for a given β than it would be for a test against a more specific H_1.

This behavior is quite similar to the tradeoff experienced with estimators: an estimator that works well for a wide range of pdf's will be less efficient than a one designed around a specific pdf, always assuming that this pdf is appropriate: for a Normal, the mean is more efficient than the median, but the latter does better over a wide range of pdf's. Similarly, we may be willing use a test with lower power if this is consistent over a wider range of alternative hypotheses.