14.2.2 F-Test for Significantly Different Variances

The *F-test* tests the hypothesis that two samples have different variances by trying to reject the null hypothesis that their variances are actually consistent. The statistic F is the ratio of one variance to the other, so values either $\gg 1$ or $\ll 1$ will indicate very significant differences. The distribution of F in the null case is given in equation (6.14.49), which is evaluated using the routine `betai`. In the most common case, we are willing to disprove the null hypothesis (of equal variances) by either very large or very small values of F, so the correct p-value is *two-tailed*, the sum of two incomplete beta functions. It turns out, by equation (6.4.3), that the two tails are always equal; we need compute only one, and double it. Occasionally, when the null hypothesis is strongly viable, the identity of the two tails can become confused, giving an indicated probability greater than one. Changing the probability to two minus itself correctly exchanges the tails. These considerations and equation (6.4.3) give the routine

stattests.h

```
void ftest(VecDoub_I &data1, VecDoub_I &data2, Doub &f, Doub &prob) {
Given the arrays data1[0..n1-1] and data2[0..n2-1], this routine returns the value of f,
and its p-value as prob. Small values of prob indicate that the two arrays have significantly
different variances.
    Beta beta;
    Doub var1,var2,ave1,ave2,df1,df2;
    Int n1=data1.size(), n2=data2.size();
    avevar(data1,ave1,var1);
    avevar(data2,ave2,var2);
    if (var1 > var2) {                      Make F the ratio of the larger variance to the smaller
        f=var1/var2;                           one.
        df1=n1-1;
        df2=n2-1;
    } else {
        f=var2/var1;
        df1=n2-1;
        df2=n1-1;
    }
    prob = 2.0*beta.betai(0.5*df2,0.5*df1,df2/(df2+df1*f));
    if (prob > 1.0) prob=2.-prob;
}
```

CITED REFERENCES AND FURTHER READING:

Spiegel, M.R., Schiller, J., and Srinivasan, R.A. 2000, *Schaum's Outline of Theory and Problem of Probability and Statistics*, 2nd ed. (New York: McGraw-Hill).

Lupton, R. 1993, *Statistics in Theory and Practice* (Princeton, NJ: Princeton University Press), Chapter 9.

Devore, J.L. 2003, *Probability and Statistics for Engineering and the Sciences*, 6th ed. (Belmont, CA: Duxbury Press), Chapters 7–8.

Norusis, M.J. 2006, *SPSS 14.0 Guide to Data Analysis* (Englewood Cliffs, NJ: Prentice-Hall).

14.3 Are Two Distributions Different?

Given two sets of data, we can generalize the questions asked in the previous section and ask the single question: Are the two sets drawn from the same distribution function, or from different distribution functions? Equivalently, in proper statistical language, "Can we disprove, to a certain required level of significance, the null hypothesis that two data sets are drawn from the same population distribution function?" Disproving the null hypothesis in effect proves that the data sets are from different distributions. Failing to disprove the null hypothesis, on the other hand, only shows that the data sets can be *consistent* with a single distribution function. One can never *prove* that two data sets come from a single distribution, since, e.g., no practical amount of data can distinguish between two distributions that differ only by one part in 10^{10}.

Proving that two distributions are different, or showing that they are consistent, is a task that comes up all the time in many areas of research: Are the visible stars distributed uniformly in the sky? (That is, is the distribution of stars as a function of declination — position in the sky — the same as the distribution of sky area as a function of declination?) Are educational patterns the same in Brooklyn as in the Bronx? (That is, are the distributions of people as a function of last-grade-attended the same?) Do two brands of fluorescent lights have the same distribution of burn-out times? Is the incidence of chicken pox the same for first-born, second-born, third-born children, etc.?

These four examples illustrate the four combinations arising from two different dichotomies: (1) The data are either continuous or binned. (2) Either we wish to compare one data set to a known distribution, or we wish to compare two equally unknown data sets. The data sets on fluorescent lights and on stars are continuous, since we can be given lists of individual burnout times or of stellar positions. The data sets on chicken pox and educational level are binned, since we are given tables of numbers of events in discrete categories: first-born, second-born, etc.; or 6th grade, 7th grade, etc. Stars and chicken pox, on the other hand, share the property that the null hypothesis is a known distribution (distribution of area in the sky, or incidence of chicken pox in the general population). Fluorescent lights and educational level involve the comparison of two equally unknown data sets (the two brands, or Brooklyn and the Bronx).

One can always turn continuous data into binned data, by grouping the events into specified ranges of the continuous variable(s): declinations between 0 and 10 degrees, 10 and 20, 20 and 30, etc. Binning involves a loss of information, however. Also, there is often considerable arbitrariness as to how the bins should be chosen. Along with many other investigators, we prefer to avoid unnecessary binning of data.

The accepted test for differences between binned distributions is the *chi-square test*. For continuous data as a function of a single variable, the most generally accepted test is the *Kolmogorov-Smirnov test*. We consider each in turn.

14.3.1 Chi-Square Test

Suppose that N_i is the number of events observed in the ith bin, and that n_i is the number expected according to some known distribution. Note that the N_i's are integers, while the n_i's may not be. Then the chi-square statistic is

$$\chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i} \qquad (14.3.1)$$

where the sum is over all bins. A large value of χ^2 indicates that the null hypothesis (that the N_i's are drawn from the population represented by the n_i's) is rather unlikely.

Any term j in (14.3.1) with $0 = n_j = N_j$ should be omitted from the sum. A term with $n_j = 0,\ N_j \neq 0$ gives an infinite χ^2, as it should, since in this case the N_i's cannot possibly be drawn from the n_i's!

The *chi-square probability function* $Q(\chi^2|\nu)$ is an incomplete gamma function, and was already discussed in §6.14 (see equation 6.14.38). Strictly speaking, $Q(\chi^2|\nu)$ is the probability that the sum of the squares of ν random *normal* variables of unit variance (and zero mean) will be greater than χ^2. The terms in the sum (14.3.1) are not exactly the squares of a normal variable. However, if the number of events in each bin is large ($\gg 1$), then the normal distribution is approximately achieved and the chi-square probability function is a good approximation to the distribution of (14.3.1) in the case of the null hypothesis. Its use to estimate the p-value significance of the chi-square test is standard (but see §14.3.2).

The appropriate value of ν, the number of degrees of freedom, bears some additional discussion. If the data are collected with the model n_i's fixed — that is, not later renormalized to fit the total observed number of events ΣN_i — then ν equals the number of bins N_B. (Note that this is *not* the total number of *events*!) Much more commonly, the n_i's are normalized after the fact so that their sum equals the sum of the N_i's. In this case, the correct value for ν is $N_B - 1$, and the model is said to have one constraint (knstrn=1 in the program below). If the model that gives the n_i's has additional free parameters that were adjusted after the fact to agree with the data, then each of these additional "fitted" parameters decreases ν (and increases knstrn) by one additional unit.

We have, then, the following program:

stattests.h

```
void chsone(VecDoub_I &bins, VecDoub_I &ebins, Doub &df,
    Doub &chsq, Doub &prob, const Int knstrn=1) {
```

Given the array bins[0..nbins-1] containing the observed numbers of events, and an array ebins[0..nbins-1] containing the expected numbers of events, and given the number of constraints knstrn (normally one), this routine returns (trivially) the number of degrees of freedom df, and (nontrivially) the chi-square chsq and the p-value prob. A small value of prob indicates a significant difference between the distributions bins and ebins. Note that bins and ebins are both double arrays, although bins will normally contain integer values.

```
    Gamma gam;
    Int j,nbins=bins.size();
    Doub temp;
    df=nbins-knstrn;
    chsq=0.0;
    for (j=0;j<nbins;j++) {
        if (ebins[j]<0.0 || (ebins[j]==0. && bins[j]>0.))
            throw("Bad expected number in chsone");
        if (ebins[j]==0.0 && bins[j]==0.0) {
            --df;                                  No data means one less degree of free-
        } else {                                      dom.
            temp=bins[j]-ebins[j];
            chsq += temp*temp/ebins[j];
        }
    }
    prob=gam.gammq(0.5*df,0.5*chsq);           Chi-square probability function. See §6.2.
}
```

Next we consider the case of comparing *two* binned data sets. Let R_i be the number of events in bin i for the first data set and S_i the number of events in the same bin i for the second data set. Then the chi-square statistic is

$$\chi^2 = \sum_i \frac{(R_i - S_i)^2}{R_i + S_i} \tag{14.3.2}$$

Comparing (14.3.2) to (14.3.1), you should note that the denominator of (14.3.2) is *not* just the average of R_i and S_i (which would be an estimator of n_i in 14.3.1). Rather, it is twice the average, the sum. The reason is that each term in a chi-square sum is supposed to approximate the square of a normally distributed quantity with unit variance. The variance of the difference of two normal quantities is the sum of their individual variances, not the average.

If the data were collected in such a way that the sum of the R_i's is necessarily equal to the sum of S_i's, then the number of degrees of freedom is equal to one less than the number of bins, $N_B - 1$ (that is, knstrn $= 1$), the usual case. If this requirement were absent, then the number of degrees of freedom would be N_B. Example: A birdwatcher wants to know whether the distribution of sighted birds as a function of species is the same this year as last. Each bin corresponds to one species. If the birdwatcher takes his data to be the first 1000 birds that he saw in each year, then the number of degrees of freedom is $N_B - 1$. If he takes his data to be all the birds he saw on a random sample of days, the same days in each year, then the number of degrees of freedom is N_B (knstrn $= 0$). In this latter case, note that he is also testing whether the birds were more numerous overall in one year or the other: That is the extra degree of freedom. Of course, any additional constraints on the data set lower the number of degrees of freedom (i.e., increase knstrn to *more positive* values) in accordance with their number.

The program is

stattests.h

```
void chstwo(VecDoub_I &bins1, VecDoub_I &bins2, Doub &df,
    Doub &chsq, Doub &prob, const Int knstrn=1) {
```

Given the arrays bins1[0..nbins-1] and bins2[0..nbins-1], containing two sets of binned data, and given the number of constraints knstrn (normally 1 or 0), this routine returns the number of degrees of freedom df, the chi-square chsq, and the p-value prob. A small value of prob indicates a significant difference between the distributions bins1 and bins2. Note that bins1 and bins2 are both double arrays, although they will normally contain integer values.

```
    Gamma gam;
    Int j,nbins=bins1.size();
    Doub temp;
    df=nbins-knstrn;
    chsq=0.0;
    for (j=0;j<nbins;j++)
        if (bins1[j] == 0.0 && bins2[j] == 0.0)
            --df;                                   No data means one less degree of free-
        else {                                          dom.
            temp=bins1[j]-bins2[j];
            chsq += temp*temp/(bins1[j]+bins2[j]);
        }
    prob=gam.gammq(0.5*df,0.5*chsq);            Chi-square probability function. See §6.2.
}
```

Equation (14.3.2) and the routine chstwo both apply to the case where the total number of data points is the same in the two binned sets, or to the case where any difference in the totals is part of what is being tested for. For intentionally unequal sample sizes, the formula analogous to (14.3.2) is

$$\chi^2 = \sum_i \frac{(\sqrt{S/R}R_i - \sqrt{R/S}S_i)^2}{R_i + S_i} \tag{14.3.3}$$

where

$$R \equiv \sum_i R_i \qquad S \equiv \sum_i S_i \tag{14.3.4}$$

are the respective numbers of data points. It is straightforward to make the corresponding change in `chstwo`. The fact that R_i and S_i occur in the denominator of equation (14.3.3) with equal weights may seem unintuitive, but the following heuristic derivation shows how this comes about: In the null hypothesis that R_i and S_i are drawn from the same distribution, we can estimate the probability associated with bin i as

$$\widehat{p}_i = \frac{R_i + S_i}{R + S} \tag{14.3.5}$$

The expected number of counts is thus

$$\widehat{R}_i = R\widehat{p}_i \qquad \text{and} \qquad \widehat{S}_i = S\widehat{p}_i \tag{14.3.6}$$

and the chi-square statistic summing over all observations is

$$\chi^2 = \sum_i \frac{(R_i - \widehat{R}_i)^2}{\widehat{R}_i} + \sum_i \frac{(S_i - \widehat{S}_i)^2}{\widehat{S}_i} \tag{14.3.7}$$

Substituting equations (14.3.6) and (14.3.5) into equation (14.3.7) gives, after some algebra, exactly equation (14.3.3). Although there are $2N_B$ terms in equation (14.3.7), the number of degrees of freedom is actually $N_B - 1$ (minus any additional constraints), the same as equation (14.3.2), because we implicitly estimated $N_B + 1$ parameters, the $\widehat{p}_i$'s and the ratio of the two sample sizes. This number of degrees of freedom must thus be subtracted from the original $2N_B$.

For three or more samples, see equation (14.4.3) and related discussion.

14.3.2 Chi-Square with Small Numbers of Counts

When a significant fraction of bins have small numbers of counts ($\lesssim 10$, say), then the χ^2 statistics (14.3.1), (14.3.2), and (14.3.3) are not well approximated by a chi-square probability function. Let us quantify this problem and suggest some remedies.

Consider first equation (14.3.1). In the null hypothesis, the count in an individual bin, N_i, is a Poisson deviate of mean n_i, so it occurs with probability

$$p(N_i|n_i) = \exp(-n_i)\frac{n_i^{N_i}}{N_i!} \tag{14.3.8}$$

(cf. equation 6.14.61). We can calculate the mean μ and variance σ^2 of the term $(N_i - n_i)^2/n_i$ by evaluating the appropriate expectation values. There are various analytical ways to do this. The sums, and the answers, are

$$\begin{aligned}\mu &= \sum_{N_i=0}^{\infty} p(N_i|n_i)\,\frac{(N_i - n_i)^2}{n_i} = 1\\ \sigma^2 &= \left\{\sum_{N_i=0}^{\infty} p(N_i|n_i)\left[\frac{(N_i - n_i)^2}{n_i}\right]^2\right\} - \mu^2 = 2 + \frac{1}{n_i}\end{aligned} \tag{14.3.9}$$

Now we can see what the problem is: Equation (14.3.9) says that each term in (14.3.1) adds, on average, 1 to the value of the χ^2 statistic, and slightly more than 2 to its variance. But the variance of the chi-square probability function is *exactly* twice its mean (equation 6.14.37). If a significant fraction of n_i's are small, then quite probable values of the χ^2 statistic will appear to lie farther out on the tail than they actually are, so that the null hypothesis may be rejected even when it is true.

Several approximate remedies are possible. One is simply to rescale the observed χ^2 statistic so as to "fix" its variance, an idea due to Lucy [1]. If we define

$$Y^2 \equiv \nu + \sqrt{\frac{2\nu}{2\nu + \sum_i n_i^{-1}}}\left(\chi^2 - \nu\right) \tag{14.3.10}$$

where ν is the number of degrees of freedom (see discussion above), then Y^2 is asymptotically approximated by the chi-square probability function even when many n_i's are small. The basic idea in (14.3.10) is to subtract off the mean, rescale the difference from the mean, and then add back the mean. Lucy [1] also defines a similar Z^2 statistic by rescaling not the χ^2 sum of all the terms, but the terms individually, using equation (14.3.9) separately for each.

Another possibility, valid when ν is large, is to use the central limit theorem directly. From its mean and standard deviation, we now know that the χ^2 statistic must be approximately the normal distribution,

$$\chi^2 \sim \mathrm{N}\left(\nu, \left[2\nu + \sum_i n_i^{-1}\right]^{1/2}\right) \tag{14.3.11}$$

We can then obtain p-values from equation (6.14.2), computing a complementary error function. (The p-value is one minus that cdf.)

The same ideas go through in the case of two binned data sets, with counts R_i and S_i, and total numbers of counts R and S (equation 14.3.3, with equation 14.3.2 as the special case with $R = S$). Now, in the null hypothesis, and glossing over some technical issues beyond our scope, we can think of $T_i \equiv R_i + S_i$ as being fixed, while R_i is a random variable drawn from the binomial distribution

$$R_i \sim \text{Binomial}\left(T_i, \frac{R}{R+S}\right) \tag{14.3.12}$$

(see equation 6.14.67). Calculating moments over the binomial distribution, one can obtain as analogs of equations (14.3.9)

$$\begin{aligned} \mu &= 1 \\ \sigma^2 &= 2 + \left[\frac{(R+S)^2}{RS} - 6\right]\frac{1}{R_i + S_i} \end{aligned} \tag{14.3.13}$$

Notice that, now, depending on the values of R and S, the variance can be either greater or less than its nominal value 2, and that it is less than 2 for the case $R = S$. The formulas (14.3.9) and (14.3.13) are originally due to Haldane [2] (see also [3]).

Summing over i, one obtains the analogs of equations (14.3.10) and (14.3.11) simply by the replacement,

$$\sum_i n_i^{-1} \longrightarrow \left[\frac{(R+S)^2}{RS} - 6\right]\sum_i \frac{1}{R_i + S_i} \tag{14.3.14}$$

In fact, equation (14.3.9) is a limiting form of equation (14.3.13) in just the same limit that Poisson is a limiting form of binomial, namely

$$S \to \infty, \qquad \frac{R}{R+S} S_i \to n_i, \qquad R_i \to N_i \tag{14.3.15}$$

There are also other ways of treating small-number counts, including the likelihood ratio test [4], the *modified Neyman* χ^2 [5], and the *chi-square-gamma* statistic [5].

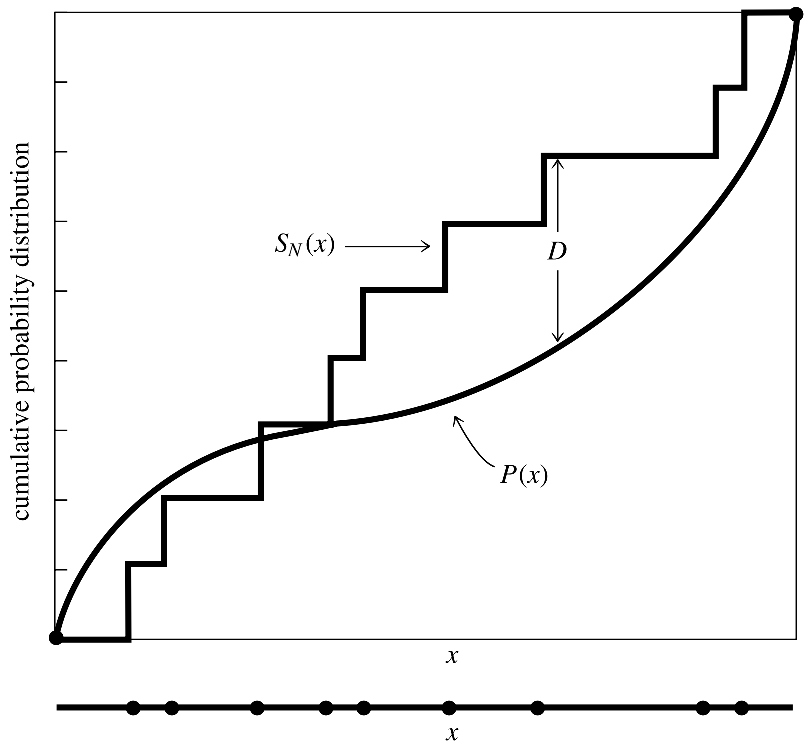

Figure 14.3.1. Kolmogorov-Smirnov statistic D. A measured distribution of values in x (shown as N dots on the lower abscissa) is to be compared with a theoretical distribution whose cumulative probability distribution is plotted as $P(x)$. A step-function cumulative probability distribution $S_N(x)$ is constructed, one that rises an equal amount at each measured point. D is the greatest distance between the two cumulative distributions.

14.3.3 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (or *K–S*) test is applicable to unbinned distributions that are functions of a single independent variable, that is, to data sets where each data point can be associated with a single number (lifetime of each lightbulb when it burns out, or declination of each star). In such cases, the list of data points can be easily converted to an unbiased estimator $S_N(x)$ of the *cumulative* distribution function of the probability distribution from which it was drawn: If the N events are located at values x_i, $i = 0, \ldots, N-1$, then $S_N(x)$ is the function giving the fraction of data points to the left of a given value x. This function is obviously constant between consecutive (i.e., sorted into ascending order) x_i's and jumps by the same constant $1/N$ at each x_i. (See Figure 14.3.1.)

Different distribution functions, or sets of data, give different cumulative distribution function estimates by the above procedure. However, all cumulative distribution functions agree at the smallest allowable value of x (where they are zero) and at the largest allowable value of x (where they are unity). (The smallest and largest values might of course be $\pm\infty$.) So it is the behavior between the largest and smallest values that distinguishes distributions.

One can think of any number of statistics to measure the overall difference between two cumulative distribution functions: the absolute value of the area between them, for example, or their integrated mean square difference. The Kolmogorov-Smirnov D is a particularly simple measure: It is defined as the *maximum value* of the absolute difference between two cumulative distribution functions. Thus, for comparing one data set's $S_N(x)$ to a known cumulative distribution function $P(x)$, the K–S statistic is

$$D = \max_{-\infty<x<\infty} |S_N(x) - P(x)| \tag{14.3.16}$$

while for comparing two different cumulative distribution functions $S_{N_1}(x)$ and $S_{N_2}(x)$, the K–S statistic is

$$D = \max_{-\infty<x<\infty} \left|S_{N_1}(x) - S_{N_2}(x)\right| \tag{14.3.17}$$

What makes the K–S statistic useful is that *its* distribution in the case of the null hypothesis (data sets drawn from the same distribution) can be calculated, at least to a useful approximation, thus giving the p-value significance of any observed nonzero value of D. A central feature of the K–S test is that it is invariant under reparametrization of x; in other words, you can locally slide or stretch the x-axis in Figure 14.3.1, and the maximum distance D remains unchanged. For example, you will get the same significance using x as using $\log x$.

The function that enters into the calculation of the p-value was discussed previously in §6.14, was defined in equations (6.14.56) and (6.14.57), and was implemented in the object `KSdist`. In terms of the function Q_{KS}, the p-value of an observed value of D (as a disproof of the null hypothesis that the distributions are the same) is given approximately [6] by the formula

$$\text{Probability } (D > \text{observed }) = Q_{KS}\left(\left[\sqrt{N_e} + 0.12 + 0.11/\sqrt{N_e}\right] D\right) \tag{14.3.18}$$

where N_e is the effective number of data points, $N_e = N$ for the case (14.3.16) of one distribution, and

$$N_e = \frac{N_1 N_2}{N_1 + N_2} \tag{14.3.19}$$

for the case (14.3.17) of two distributions, where N_1 is the number of data points in the first distribution and N_2 the number in the second.

The nature of the approximation involved in (14.3.18) is that it becomes asymptotically accurate as the N_e becomes large, but is already quite good for $N_e \geq 4$, as small a number as one might ever actually use. (See [6].)

Here is the routine for the case of one distribution:

kstests.h

```
void ksone(VecDoub_IO &data, Doub func(const Doub), Doub &d, Doub &prob)
```

Given an array `data[0..n-1]`, and given a user-supplied function of a single variable `func` that is a cumulative distribution function ranging from 0 (for smallest values of its argument) to 1 (for largest values of its argument), this routine returns the K–S statistic `d` and the p-value `prob`. Small values of `prob` show that the cumulative distribution function of `data` is significantly different from `func`. The array `data` is modified by being sorted into ascending order.

```
{
    Int j,n=data.size();
    Doub dt,en,ff,fn,fo=0.0;
    KSdist ks;
    sort(data);                          If the data are already sorted into as-
                                             cending order, then this call can be
                                             omitted.
    en=n;
    d=0.0;
    for (j=0;j<n;j++) {                  Loop over the sorted data points.
        fn=(j+1)/en;                     Data's c.d.f. after this step.
        ff=func(data[j]);                            Compare to the user-supplied function.
        dt=MAX(abs(fo-ff),abs(fn-ff));               Maximum distance.
        if (dt > d) d=dt;
        fo=fn;
    }
    en=sqrt(en);
    prob=ks.qks((en+0.12+0.11/en)*d);                Compute p-value.
}
```

While the K-S statistic is intended for use with a continuous distribution, it can also be used for a discrete distribution. In this case, it can be shown that the test is conservative, that is, the statistic returned is no larger than in the continuous case. If you allow discrete variables in the case of two distributions, you have to consider how to deal with ties. The standard way to handle ties is to combine all the tied data points and add them to the cdf at once (see, e.g., [7]). This refinement is included in the routine kstwo.

kstests.h

```
void kstwo(VecDoub_IO &data1, VecDoub_IO &data2, Doub &d, Doub &prob)
```

Given an array data1[0..n1-1], and an array data2[0..n2-1], this routine returns the K–S statistic d and the p-value prob for the null hypothesis that the data sets are drawn from the same distribution. Small values of prob show that the cumulative distribution function of data1 is significantly different from that of data2. The arrays data1 and data2 are modified by being sorted into ascending order.

```
{
    Int j1=0,j2=0,n1=data1.size(),n2=data2.size();
    Doub d1,d2,dt,en1,en2,en,fn1=0.0,fn2=0.0;
    KSdist ks;
    sort(data1);
    sort(data2);
    en1=n1;
    en2=n2;
    d=0.0;
    while (j1 < n1 && j2 < n2) {                       If we are not done...
        if ((d1=data1[j1]) <= (d2=data2[j2]))          Next step is in data1.
            do
                fn1=++j1/en1;
            while (j1 < n1 && d1 == data1[j1]);
        if (d2 <= d1)                                  Next step is in data2.
            do
                fn2=++j2/en2;
            while (j2 < n2 && d2 == data2[j2]);
        if ((dt=abs(fn2-fn1)) > d) d=dt;
    }
    en=sqrt(en1*en2/(en1+en2));
    prob=ks.qks((en+0.12+0.11/en)*d);                  Compute p-value.
}
```

14.3.4 Variants on the K–S Test

The sensitivity of the K–S test to deviations from a cumulative distribution function $P(x)$ is not independent of x. In fact, the K–S test tends to be most sensitive around the median value, where $P(x) = 0.5$, and less sensitive at the extreme ends of the distribution, where $P(x)$ is near 0 or 1. The reason is that the difference $|S_N(x) - P(x)|$ does not, in the null hypothesis, have a probability distribution that is independent of x. Rather, its variance is proportional to $P(x)[1 - P(x)]$, which is largest at $P = 0.5$. Since the K–S statistic (14.3.16) is the maximum difference over all x of two cumulative distribution functions, a deviation that might be statistically significant at *its own* value of x gets compared to the expected chance deviation at $P = 0.5$ and is thus discounted. A result is that, while the K–S test is good at finding *shifts* in a probability distribution, especially changes in the median value, it is not always so good at finding *spreads*, which more affect the tails of the probability distribution, and which may leave the median unchanged.

One way of increasing the power of the K–S statistic out on the tails is to replace D (equation 14.3.16) by a so-called *stabilized* or *weighted* statistic [8-10], for example the *Anderson-Darling statistic*,

$$D^* = \max_{-\infty<x<\infty} \frac{|S_N(x) - P(x)|}{\sqrt{P(x)[1-P(x)]}} \tag{14.3.20}$$

Unfortunately, there is no simple formula analogous to equation (14.3.18) for this statistic, although Noé [11] gives a computational method using a recursion relation and provides a graph of numerical results. There are many other possible similar statistics, for example

$$D^{**} = \int_{P=0}^{1} \frac{[S_N(x) - P(x)]^2}{P(x)[1-P(x)]} dP(x) \tag{14.3.21}$$

which is also discussed by Anderson and Darling (see [9]).

Another approach, which we prefer as simpler and more direct, is due to Kuiper [12,13]. We already mentioned that the standard K–S test is invariant under reparametrizations of the variable x. An even more general symmetry, which guarantees equal sensitivities at all values of x, is to wrap the x-axis around into a circle (identifying the points at $\pm\infty$), and to look for a statistic that is now invariant under all shifts and parametrizations on the circle. This allows, for example, a probability distribution to be "cut" at some central value of x and the left and right halves to be interchanged, without altering the statistic or its significance.

Kuiper's statistic, defined as

$$V = D_+ + D_- = \max_{-\infty<x<\infty} [S_N(x) - P(x)] + \max_{-\infty<x<\infty} [P(x) - S_N(x)] \tag{14.3.22}$$

is the sum of the maximum distance of $S_N(x)$ *above and below* $P(x)$. You should be able to convince yourself that this statistic has the desired invariance on the circle: Sketch the indefinite integral of two probability distributions defined on the circle as a function of angle around the circle, as the angle goes through several times 360°. If you change the starting point of the integration, D_+ and D_- change individually, but their sum is constant.

Furthermore, there is a simple formula for the asymptotic distribution of the statistic V, directly analogous to equations (14.3.18) – (14.3.19). Let

$$Q_{KP}(\lambda) = 2\sum_{j=1}^{\infty}(4j^2\lambda^2 - 1)e^{-2j^2\lambda^2} \tag{14.3.23}$$

which is monotonic and satisfies

$$Q_{KP}(0) = 1 \qquad Q_{KP}(\infty) = 0 \tag{14.3.24}$$

In terms of this function the p-value is [6]

$$\text{Probability } (V > \text{observed }) = Q_{KP}\left(\left[\sqrt{N_e} + 0.155 + 0.24/\sqrt{N_e}\right] V\right) \tag{14.3.25}$$

Here N_e is N in the one-sample case or is given by equation (14.3.19) in the case of two samples.

Of course, Kuiper's test is ideal for any problem originally defined on a circle, for example, to test whether the distribution in longitude of something agrees with some theory, or whether two somethings have different distributions in longitude. (See also [14].)

We will leave to you the coding of routines analogous to `ksone`, `kstwo`, and `KSdist::qks`. (For $\lambda < 0.4$, don't try to do the sum 14.3.23. Its value is 1, to 7 figures, but the series can require many terms to converge, and loses accuracy to roundoff.)

Two final cautionary notes: First, we should mention that all varieties of the K–S test lack the ability to discriminate some kinds of distributions. A simple example is a probability distribution with a narrow "notch" within which the probability falls to zero. Such a distribution is of course ruled out by the existence of even one data point within the notch, but, because of its cumulative nature, a K–S test would require many data points in the notch before signaling a discrepancy.

Second, we should note that, if you estimate any parameters from a data set (e.g., a mean and variance), then the distribution of the K–S statistic D for a cumulative distribution function $P(x)$ that *uses the estimated parameters* is no longer given by equation (14.3.18). In general, you will have to determine the new distribution yourself, e.g., by Monte Carlo methods.

CITED REFERENCES AND FURTHER READING:

Devore, J.L. 2003, *Probability and Statistics for Engineering and the Sciences*, 6th ed. (Belmont, CA: Duxbury Press), Chapter 14.

Lupton, R. 1993, *Statistics in Theory and Practice* (Princeton, NJ: Princeton University Press), Chapter 14.

Lucy, L.B. 2000, "Hypothesis Testing for Meagre Data Sets," *Monthly Notices of the Royal Astronomical Society*, vol. 318, pp. 92–100.[1]

Haldane, J.B.S. 1937, "The Exact Value of the Moments of the Distribution of χ^2, Used as a Test of Goodness of Fit, When Expectations Are Small," *Biometrika*, vol. 29, pp. 133–143.[2]

Read, T.R.C., and Cressie, N.A.C. 1988, *Goodness-of-Fit Statistics for Discrete Multivariate Data* (New York: Springer), pp. 140–144.[3]

Baker, S., and Cousins, R.D. 1984, "Clarification of the Use of Chi-Square and Likelihood Functions in Fits to Histograms," *Nuclear Instruments and Methods in Physics Research*, vol. 221, pp. 437–442.[4]

Mighell, K.J. 1999, "Parameter Estimation in Astronomy with Poisson-Distributed Data. I.The χ^2_γ Statistic," *Astrophysical Journal*, vol. 518, pp. 380–393[5]

Stephens, M.A. 1970, "Use of Kolmogorov-Smirnov, Cramer-von Mises and Related Statistics without Extensive Tables," *Journal of the Royal Statistical Society*, ser. B, vol. 32, pp. 115–122.[6]

Hollander, M., and Wolfe, D.A. 1999, *Nonparametric Statistical Methods*, 2nd ed. (New York: Wiley), p. 183.[7]

Anderson, T.W., and Darling, D.A. 1952, "Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes," *Annals of Mathematical Statistics*, vol. 23, pp. 193–212.[8]

Darling, D.A. 1957, "The Kolmogorov-Smirnov, Cramer-von Mises Tests," *Annals of Mathematical Statistics*, vol. 28, pp. 823–838.[9]

Michael, J.R. 1983, "The Stabilized Probability Plot," *Biometrika*, vol. 70, no. 1, pp. 11–17.[10]

Noé, M. 1972, "The Calculation of Distributions of Two-Sided Kolmogorov-Smirnov Type Statistics," *Annals of Mathematical Statistics*, vol. 43, pp. 58–64.[11]

Kuiper, N.H. 1962, "Tests Concerning Random Points on a Circle," *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, ser. A., vol. 63, pp. 38–47.[12]

Stephens, M.A. 1965, "The Goodness-of-Fit Statistic V_n: Distribution and Significance Points," *Biometrika*, vol. 52, pp. 309–321.[13]

Fisher, N.I., Lewis, T., and Embleton, B.J.J. 1987, *Statistical Analysis of Spherical Data* (New York: Cambridge University Press).[14]