14.6 Nonparametric or Rank Correlation

It is precisely the uncertainty in interpreting the significance of the linear correlation coefficient $r$ that leads us to the important concepts of nonparametric or rank correlation. As before, we are given $N$ pairs of measurements $(x_i, y_i)$. Before, difficulties arose because we did not necessarily know the probability distribution function from which the $x_i$'s or $y_i$'s were drawn.

The key concept of nonparametric correlation is this: If we replace the value of each $x_i$ by the value of its rank among all the other $x_i$'s in the sample, that is, $1, 2, 3, \ldots, N$, then the resulting list of numbers will be drawn from a perfectly known distribution function, namely uniformly from the integers between 1 and $N$, inclusive. Better than uniformly, in fact, since if the $x_i$'s are all distinct, then each integer will occur precisely once. If some of the $x_i$'s have identical values, it is conventional to assign to all these “ties” the mean of the ranks that they would have had if their values had been slightly different. This midrank will sometimes be an

\begin{verbatim}
ax=ax+x(j)
ay=ay+y(j)
enddo:
ax=ax/n
ay=ay/n
sxx=0.
syy=0.
sxy=0.
do#: j=1,n
    xt=x(j)-ax
    yt=y(j)-ay
    sxx=sxx+xt**2
    syy=syy+yt**2
    sxy=sxy+xt*yt
endo:
r=sxy/(sqrt(sxx*syy)+TINY)
G
f =n-2
r=t*sqrt(df/((1.+r)+TINY)*(1.-r)+TINY))
prob=betai(0.5*df,0.5,df/(df+t**2))
C prob=erfcc(abs(z*sqrt(n-1.))/1.4142136)
return
END
\end{verbatim}

CITED REFERENCES AND FURTHER READING:
integer, sometimes a half-integer. In all cases the sum of all assigned ranks will be
the same as the sum of the integers from 1 to \( N \), namely \[ \frac{1}{2}N(N + 1). \]

Of course we do exactly the same procedure for the \( y_i \)'s, replacing each value by its
rank among the other \( y_i \)'s in the sample.

Now we are free to invent statistics for detecting correlation between uniform
sets of integers between 1 and \( N \), keeping in mind the possibility of ties in the ranks.
There is, of course, some loss of information in replacing the original numbers by
ranks. We could construct some rather artificial examples where a correlation could be
detected parametrically (e.g., in the linear correlation coefficient \( r \)), but could not be
detected nonparametrically. Such examples are very rare in real life, however,
and the slight loss of information in ranking is a small price to pay for a very major
advantage: When a correlation is demonstrated to be present nonparametrically,
then it is really there! (That is, to a certainty level that depends on the significance
chosen.) Nonparametric correlation is more robust than linear correlation, more
resistant to unplanned defects in the data, in the same sort of sense that the median
is more robust than the mean. For more on the concept of robustness, see §15.7.

As always in statistics, some particular choices of a statistic have already been
invented for us and consecrated, if not beati®ed, by popular use. We will discuss
two, the **Spearman rank-order correlation coefficient** (\( r_s \)), and **Kendall's tau** (\( \tau \)).

**Spearman Rank-Order Correlation Coefficient**

Let \( R_i \) be the rank of \( x_i \) among the other \( x \)'s, \( S_i \) be the rank of \( y_i \) among the
other \( y \)'s, ties being assigned the appropriate midrank as described above. Then the
rank-order correlation coefficient is defined to be the linear correlation coefficient
of the ranks, namely,

\[
r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2} \sqrt{\sum_i (S_i - \bar{S})^2}}
\]

(14.6.1)

The significance of a nonzero value of \( r_s \) is tested by computing

\[
t = r_s \sqrt{\frac{N - 2}{1 - r_s^2}}
\]

(14.6.2)

which is distributed approximately as Student’s distribution with \( N - 2 \) degrees of
freedom. A key point is that this approximation does not depend on the original
distribution of the \( x \)'s and \( y \)'s; it is always the same approximation, and always
pretty good.

It turns out that \( r_s \) is closely related to another conventional measure of
nonparametric correlation, the so-called **sum squared difference of ranks**, defined as

\[
D = \sum_{i=1}^{N} (R_i - S_i)^2
\]

(14.6.3)

(This \( D \) is sometimes denoted \( D^{**} \), where the asterisks are used to indicate that
ties are treated by midranking.)
When there are no ties in the data, then the exact relation between $D$ and $r_s$ is

$$r_s = 1 - \frac{6D}{N^3 - N} \quad (14.6.4)$$

When there are ties, then the exact relation is slightly more complicated: Let $f_k$ be the number of ties in the $k$th group of ties among the $R_i$'s, and let $g_m$ be the number of ties in the $m$th group of ties among the $S_i$'s. Then it turns out that

$$r_s = 1 - \frac{6}{N^3 - N} \left[ D + \frac{1}{12} \sum_k (f_k^3 - f_k) + \frac{1}{12} \sum_m (g_m^3 - g_m) \right]$$

$$\left[ 1 - \frac{\sum_k (f_k^3 - f_k)}{N^3 - N} \right]^{1/2} \left[ 1 - \frac{\sum_m (g_m^3 - g_m)}{N^3 - N} \right]^{1/2} \quad (14.6.5)$$

holds exactly. Notice that if all the $f_k$'s and all the $g_m$'s are equal to one, meaning that there are no ties, then equation (14.6.5) reduces to equation (14.6.4).

In (14.6.2) we gave a $t$-statistic that tests the significance of a nonzero $r_s$. It is also possible to test the significance of $D$ directly. The expectation value of $D$ in the null hypothesis of uncorrelated data sets is

$$D = \frac{1}{6} (N^3 - N) - \frac{1}{12} \sum_k (f_k^3 - f_k) - \frac{1}{12} \sum_m (g_m^3 - g_m) \quad (14.6.6)$$

its variance is

$$\text{Var}(D) = \frac{(N-1)N^2(N+1)^2}{36} \times \left[ 1 - \frac{\sum_k (f_k^3 - f_k)}{N^3 - N} \right] \left[ 1 - \frac{\sum_m (g_m^3 - g_m)}{N^3 - N} \right] \quad (14.6.7)$$

and it is approximately normally distributed, so that the significance level is a complementary error function (cf. equation 14.5.2). Of course, (14.6.2) and (14.6.7) are not independent tests, but simply variants of the same test. In the program that follows, we return both the significance level obtained by using (14.6.2) and the significance level obtained by using (14.6.7); their discrepancy will give you an idea of how good the approximations are. You will also notice that we break off the task of assigning ranks (including tied midranks) into a separate routine, crank.

SUBROUTINE spear(data1,data2,n,wksp1,wksp2,d,zd,probd,rs,probrs)
INTEGER n
REAL d,probd,probrs,rs,zd,data1(n),data2(n),wksp1(n),wksp2(n)
C USES betai,crank,erfcc,sort2

Given two data arrays, data1(1:n) and data2(1:n), each of length n, and given two workspaces of the same length, this routine returns their sum-squared difference of ranks as $D$, the number of standard deviations by which $D$ deviates from its null-hypothesis expected value as $zd$, the two-sided significance level of this deviation as probd, Spearman’s rank correlation $r_s$ as rs, and the two-sided significance level of its deviation from zero as probrs. The workspaces can be identical to the data arrays, but in that case the data arrays are destroyed. The external routines crank (below) and sort2 (§8.2) are used.


small value of either \( \text{probd} \) or \( \text{probrs} \) indicates a significant correlation (\( rs \) positive) or anticorrelation (\( rs \) negative).

INTEGER \( j \)
REAL \( \text{aved}, \text{df}, \text{en}, \text{en3n}, \text{fac}, \text{sf}, \text{sg}, \text{t}, \text{vard}, \text{betai}, \text{erfcc} \)
do :: \( j=1, n \)
\( \text{wksp1}(j) = \text{data1}(j) \)
\( \text{wksp2}(j) = \text{data2}(j) \)
endo ::
call sort2(n, wksp1, wksp2)  
Sort each of the data arrays, and convert the entries to ranks. The values \( sf \) and \( sg \) return the sums \( \sum (f_k^3 - f_k) \)
call crank(n, wksp2, wksp1)  
and \( \sum (g_m^3 - g_m) \), respectively.
call crank(n, wksp2, sf)
d=0.
do :: \( j=1, n \)
\( d = (\text{wksp1}(j) - \text{wksp2}(j))^2 \)
endo ::
\( \text{en} = n \)
\( \text{en3n} = \text{en}^3 - \text{en} \)
\( \text{aved} = \text{en3n}/6. - (\text{sf} + \text{sg})/12. \)
\( \text{fac} = (1. - \text{sf}/\text{en3n})* (1. - \text{sg}/\text{en3n}) \)
\( \text{vard} = ((\text{en} - 1.)*\text{en}^2*(\text{en} + 1.))**2/36.*\text{fac} \)
\( \text{probd} = \text{erfcc}(\text{abs}(zd)/1.4142136) \)
\( \text{rs} = (1. - (6./\text{en3n})* (d + (\text{sf} + \text{sg})/12.))/\text{sqrt}(\text{fac}) \)
if (\( \text{fac} \gt 0. \)) then 
\( t = \text{rs}*\text{sqrt}((\text{en} - 2.)/\text{fac}) \)
\( \text{df} = \text{en} - 2. \)
\( \text{probrs} = \text{betai}(0.5*\text{df}, 0.5, \text{df}/(\text{df} + t**2)) \)
else 
\( \text{probrs} = 0. \)
endif
return
END

SUBROUTINE crank(n, w, s)
INTEGER n
REAL s, w(n)
Given a sorted array \( w(1:n) \), replaces the elements by their rank, including midranking of ties, and returns as \( s \) the sum of \( f^3 - f \), where \( f \) is the number of elements in each tie.
INTEGER j, ji, jt
REAL rank, t
s=0.
do :: \( j=1 \)
The next rank to be assigned.
1 if (\( j \leq n \)) then 
\( \text{if}(w(j+1).\ne.w(j)) \) then 
\( w(j) = j \)
\( j = j + 1 \)
else 
A tie:
do :: \( jt = j + 1 \)
How far does it go?
if (\( w(jt).\ne.w(j) \)) goto 2
endo ::
\(jt = n + 1 \)
\( \text{rank} = 0.5* (j + jt - 1) \)
do :: \( ji = j, jt - 1 \)
\( w(ji) = \text{rank} \)
endo ::
\( t = jt - j \)
\( s = s + t**3 - t \)
\( j = jt \)
endif
2 if here, it goes all the way to the last element.
This is the mean rank of the tie, so enter it into all the tied entries,
and update \( s \).
do :: \( j = jt \)
goto 1
14.6 Nonparametric or Rank Correlation

Kendall’s Tau

Kendall’s \( \tau \) is even more nonparametric than Spearman’s \( r_s \) or \( D \). Instead of using the numerical difference of ranks, it uses only the relative ordering of ranks: higher in rank, lower in rank, or the same in rank. But in that case we don’t even have to rank the data! Ranks will be higher, lower, or the same if and only if the values are larger, smaller, or equal, respectively. On balance, we prefer \( r_s \) as being the more straightforward nonparametric test, but both statistics are in general use. In fact, \( \tau \) and \( r_s \) are very strongly correlated and, in most applications, are effectively the same test.

To define \( \tau \), we start with the \( N \) data points \((x_i, y_i)\). Now consider all \( \frac{1}{2} N(N-1) \) pairs of data points, where a data point cannot be paired with itself, and where the points in either order count as one pair. We call a pair concordant if the relative ordering of the ranks of the two \( x \)'s (or for that matter the two \( x \)'s themselves) is the same as the relative ordering of the ranks of the two \( y \)'s (or for that matter the two \( y \)'s themselves). We call a pair discordant if the relative ordering of the ranks of the two \( x \)'s is opposite from the relative ordering of the ranks of the two \( y \)'s. If there is a tie in either the ranks of the two \( x \)'s or the ranks of the two \( y \)'s, then we don’t call the pair either concordant or discordant. If the tie is in the \( x \)'s, we will call the pair an “extra \( y \) pair.” If the tie is in the \( y \)'s, we will call the pair an “extra \( x \) pair.” If the tie is in both the \( x \)'s and the \( y \)'s, we don’t call the pair anything at all. Are you still with us?

Kendall’s \( \tau \) is now the following simple combination of these various counts:

\[
\tau = \frac{\text{concordant} - \text{discordant}}{\sqrt{\text{concordant} + \text{discordant} + \text{extra-}y} \sqrt{\text{concordant} + \text{discordant} + \text{extra-}x}}
\]

(14.6.8)

You can easily convince yourself that this must lie between 1 and −1, and that it takes on the extreme values only for complete rank agreement or complete rank reversal, respectively.

More important, Kendall has worked out, from the combinatorics, the approximate distribution of \( \tau \) in the null hypothesis of no association between \( x \) and \( y \). In this case \( \tau \) is approximately normally distributed, with zero expectation value and a variance of

\[
\text{Var}(\tau) = \frac{4N + 10}{9N(N-1)}
\]

(14.6.9)

The following program proceeds according to the above description, and therefore loops over all pairs of data points. Beware: This is an \( O(N^2) \) algorithm, unlike the algorithm for \( r_s \), whose dominant sort operations are of order \( N \log N \). If you are routinely computing Kendall’s \( \tau \) for data sets of more than a few thousand points, you may be in for some serious computing. If, however, you are willing to bin your data into a moderate number of bins, then read on.
SUBROUTINE kendl1(data1, data2, n, tau, z, prob)
INTEGER n
REAL prob, tau, z, data1(n), data2(n)
C USES erfcc
Given data arrays data1(1:n) and data2(1:n), this program returns Kendall's $\tau$ as tau, its number of standard deviations from zero as z, and its two-sided significance level as prob.
Small values of prob indicate a significant correlation (tau positive) or anticorrelation (tau negative).
INTEGER is, j, k, n1, n2
REAL a1, a2, aa, var, erfcc
n1=0
This will be the argument of one square root in (14.6.8), and this the other.
n2=0
is=0
This will be the numerator in (14.6.8).
do :: j=1, n-1
Loop over first member of pair, and second member.
do :: k=j+1, n
a1=data1(j)-data1(k)
a2=data2(j)-data2(k)
aa=a1*a2
if(aa.ne.0.)then
Neither array has a tie.
n1=n1+1
n2=n2+1
if(aa.gt.0.)then
is=is+1
else
is=is-1
else
One or both arrays have ties.
if(a1.ne.0.)n1=n1+1
An "extra $x$" event.
if(a2.ne.0.)n2=n2+1
An "extra $y$" event.
endif
endo ::
tau=float(is)/sqrt(float(n1)*float(n2))
Equation (14.6.8).
var=(4.*n+10.)/(9.*n*(n-1.))
Equation (14.6.9).
z=tau/sqrt(var)
prob=erfcc(abs(z)/1.4142136)
Significance.
return
END

Sometimes it happens that there are only a few possible values each for $x$ and $y$. In that case, the data can be recorded as a contingency table (see §14.4) that gives the number of data points for each contingency of $x$ and $y$.

Spearman’s rank-order correlation coefficient is not a very natural statistic under these circumstances, since it assigns to each $x$ and $y$ bin a not-very-meaningful midrank value and then totals up vast numbers of identical rank differences. Kendall’s $\tau$, on the other hand, with its simple counting, remains quite natural. Furthermore, its $O(N^2)$ algorithm is no longer a problem, since we can arrange for it to loop over pairs of contingency table entries (each containing many data points) instead of over pairs of data points. This is implemented in the program that follows.

Note that Kendall’s $\tau$ can be applied only to contingency tables where both variables are ordinal, i.e., well-ordered, and that it looks specifically for monotonic correlations, not for arbitrary associations. These two properties make it less general than the methods of §14.4, which applied to nominal, i.e., unordered, variables and arbitrary associations.

Comparing kendl1 above with kendl2 below, you will see that we have “floated” a number of variables. This is because the number of events in a contingency table might be sufficiently large as to cause overflows in some of the
integer arithmetic, while the number of individual data points in a list could not possibly be that large [for an $O(N^2)$ routine!].

SUBROUTINE kendl2(tab,i,j,ip,jp,tau,z,prob)
INTEGER i,ip,j,jp
REAL prob,tau,z,tab(ip,jp)
C USES erfcc
Given a two-dimensional table tab of physical dimension (ip,jp) and logical dimension (i,j), such that tab(k,l) contains the number of events falling in bin k of one variable and bin l of another, this program returns Kendall’s $\tau$ as tau, its number of standard deviations from zero as z, and its two-sided significance level as prob. Small values of prob indicate a significant correlation (tau positive) or anticorrelation (tau negative) between the two variables. Although tab is a real array, it will normally contain integral values.

INTEGER k,k1,kj,1,l,li,lj,m1,m2,mm,nn
REAL en1,en2,pairs,points,s,var,erfcc
en1=0.
en2=0.
s=0.
nn=1+i*j
points=tab(i,j)
do :: k=0,nn-2
k1=k/j
kj=k-j*k1
points=points+tab(k1+1,kj+1)
do :: l=k+1,nn-1
li=l/j
lj=l-j*li
m1=li-k1
m2=lj-kj
mm=m1*m2
pairs=tab(ki+1,kj+1)*tab(li+1,lj+1)
if(mm.ne.0)then
  en1=en1+pairs
  en2=en2+pairs
  if(mm.gt.0)then
    s=s+pairs
  else
    s=s-pairs
  endif
else
  if(m1.ne.0)en1=en1+pairs
  if(m2.ne.0)en2=en2+pairs
endif
doint ::
tau=s/sqrt(en1*en2)
var=(4.*points+10.)/(9.*points*(points-1.))
z=tau/sqrt(var)
prob=erfcc(abs(z)/1.4142136)
return
END

CITED REFERENCES AND FURTHER READING:
14.7 Do Two-Dimensional Distributions Differ?

We here discuss a useful generalization of the K–S test (§14.3) to two-dimensional distributions. This generalization is due to Fasano and Franceschini [1], a variant on an earlier idea due to Peacock [2].

In a two-dimensional distribution, each data point is characterized by an \((x, y)\) pair of values. An example near to our hearts is that each of the 19 neutrinos that were detected from Supernova 1987A is characterized by a time \(t_i\) and by an energy \(E_i\) (see [3]). We might wish to know whether these measured pairs \((t_i, E_i)\), \(i = 1 \ldots 19\) are consistent with a theoretical model that predicts neutrino flux as a function of both time and energy — that is, a two-dimensional probability distribution in the \((x, y)\) plane. That would be a one-sample test. Or, given two sets of neutrino detections, from two comparable detectors, we might want to know whether they are compatible with each other, a two-sample test.

In the spirit of the tried-and-true, one-dimensional K–S test, we want to range over the \((x, y)\) plane in search of some kind of maximum cumulative difference between two two-dimensional distributions. Unfortunately, cumulative probability distribution is not well-defined in more than one dimension! Peacock’s insight was that a good surrogate is the integrated probability in each of four natural quadrants around a given point \((x_i, y_i)\), namely the total probabilities (or fraction of data) in \((x > x_i, y > y_i)\), \((x < x_i, y > y_i)\), \((x < x_i, y < y_i)\), \((x > x_i, y < y_i)\). The two-dimensional K–S statistic \(D\) is now taken to be the maximum difference (ranging both over data points and over quadrants) of the corresponding integrated probabilities. When comparing two data sets, the value of \(D\) may depend on which data set is ranged over. In that case, define an effective \(D\) as the average of the two values obtained. If you are confused at this point about the exact definition of \(D\), don’t fret; the accompanying computer routines amount to a precise algorithmic definition.

Figure 14.7.1 gives a feeling for what is going on. The 65 triangles and 35 squares seem to have somewhat different distributions in the plane. The dotted lines are centered on the triangle that maximizes the \(D\) statistic; the maximum occurs in the upper-left quadrant. That quadrant contains only 0.12 of all the triangles, but it contains 0.56 of all the squares. The value of \(D\) is thus 0.44. Is this statistically significant?

Even for fixed sample sizes, it is unfortunately not rigorously true that the distribution of \(D\) in the null hypothesis is independent of the shape of the two-dimensional distribution. In this respect the two-dimensional K–S test is not as natural as its one-dimensional parent. However, extensive Monte Carlo integrations have shown that the distribution of the two-dimensional \(D\) is very nearly identical for even quite different distributions, as long as they have the same coefficient of correlation \(r\), defined in the usual way by equation (14.5.1). In their paper, Fasano and Franceschini tabulate Monte Carlo results for (what amounts to) the distribution of \(D\) as a function of (of course) \(D\), sample size \(N\), and coefficient of correlation \(r\). Analyzing their results, one finds that the significance levels for the two-dimensional K–S test can be summarized by the simple, though approximate, formulas,

\[
\text{Probability (} D > \text{observed} \text{) } = Q_{KS} \left( \frac{\sqrt{ND}}{1 + \sqrt{1 - r^2(0.25 - 0.75/\sqrt{N})}} \right)
\]  
(14.7.1)

for the one-sample case, and the same for the two-sample case, but with

\[
N = \frac{N_1 \cdot N_2}{N_1 + N_2}
\]  
(14.7.2)

The above formulas are accurate enough when \(N > 20\), and when the indicated probability (significance level) less than (more significant than) 0.20 or so. When the indicated probability is > 0.20, its value may not be accurate, but the implication that the data and model (or two data sets) are not significantly different is certainly correct. Notice that in the limit of \(r \to 1\) (perfect correlation), equations (14.7.1) and (14.7.2) reduce to equations (14.3.9) and (14.3.10): The two-dimensional data lie on a perfect straight line, and the two-dimensional K–S test becomes a one-dimensional K–S test.