

## NAG Library Chapter Introduction

### G08 – Nonparametric Statistics

#### Contents

<b>1</b>	<b>Scope of the Chapter</b> .....	2
<b>2</b>	<b>Background to the Problems</b> .....	2
2.1	Parametric and Nonparametric Hypothesis Testing .....	2
2.2	Types of Nonparametric Test .....	2
2.3	Principles of Nonparametric Tests .....	3
2.3.1	Location tests .....	3
2.3.2	Dispersion tests .....	4
2.3.3	Tests of fit .....	4
2.3.4	Association and correlation tests .....	4
2.3.5	Tests of randomness .....	4
2.4	Regression using ranks .....	5
<b>3</b>	<b>Recommendations on Choice and Use of Available Routines</b> .....	5
3.1	G08A – Location Tests .....	5
3.1.1	One-sample or matched-pairs case .....	5
3.1.2	Two independent samples .....	6
3.1.3	More than two related samples .....	6
3.1.4	More than two independent samples .....	7
3.2	G08B – Dispersion Tests .....	7
3.3	G08C – Tests of Fit .....	7
3.4	G08D – Association and Correlation Tests .....	7
3.5	G08E – Tests of Randomness .....	7
3.6	G08R – Regression Using Ranks .....	8
3.7	Related Routines .....	8
<b>4</b>	<b>Index</b> .....	8
<b>5</b>	<b>Routines Withdrawn or Scheduled for Withdrawal</b> .....	9
<b>6</b>	<b>References</b> .....	9

## 1 Scope of the Chapter

The routines in this chapter perform nonparametric statistical tests which are based on distribution-free methods of analysis. For convenience, the chapter contents are divided into five types of test: tests of location, tests of dispersion, tests of distribution, tests of association and correlation, and tests of randomness. There are also routines to fit linear regression models using the ranks of the observations.

The emphasis in this chapter is on testing; if you wish to compute nonparametric correlations you are referred to Chapter G02, which contains several routines for that purpose.

There are a large number of nonparametric tests available. A selection of some of the more commonly used tests are included in this chapter.

## 2 Background to the Problems

### 2.1 Parametric and Nonparametric Hypothesis Testing

Classical techniques of statistical inference often make numerous or stringent assumptions about the nature of the population or populations from which the observations have been drawn. For instance, a testing procedure might assume that the set of data was obtained from Normally distributed populations. It might be further assumed that the populations involved have equal variances, or that there is a known relationship between the variances. In the Normal case, the test statistic derived would usually be a function of the sample means and variances, since a Normal distribution is completely characterised by its mean and variance. Alternatively, it might be assumed that the set of data was obtained from other distributions of known form, such as the gamma or the exponential. Again, a testing procedure would be devised based upon the parameters characterising such a distribution.

The type of hypothesis testing just described is usually termed **parametric** inference. Distributional assumptions are made which imply that the parameters of the chosen distribution, as estimated from the data, are sufficient to characterise the difference in distribution between the populations.

However, problems arise with parametric methods of inference when these assumptions cannot be made, either because they are contrary to the known nature of the mechanism generating a population, or because the data obviously do not satisfy the assumptions. Some parametric procedures become unreliable under relatively minor departures from the hypothesised distributional form. In the Normal case for example, tests on variances are extremely sensitive to departures from Normality in the underlying distribution.

There are also common situations, particularly in the behavioural sciences, where much more basic assumptions than that of Normality cannot be made. Data values are not always measured on continuous or even numerical scales. They may be simply categorical in nature, relating to such quantities as voting intentions or food preferences.

Techniques of inference are therefore required which do not involve making detailed assumptions about the underlying mechanism generating the observations. The routines in this chapter perform such distribution-free tests, evaluating from a set of data the value of a test statistic, together with an estimate of its significance.

For a comparison of some distribution-based and distribution-free tests, the interested reader is referred to Chapter 31 of Kendall and Stuart (1973). For a briefer and less mathematical account, see Conover (1980) or Siegel (1956).

### 2.2 Types of Nonparametric Test

This introduction is concerned with explaining the basic concepts of hypothesis testing, and some familiarity with the subject is assumed. Chapter 22 of Kendall and Stuart (1973) contains a detailed account, and the outline given in Conover (1980) or Siegel (1956) should be sufficient to understand this section.

Nonparametric tests may be grouped into five categories:

1. Tests of location
2. Tests of dispersion
3. Distribution-free tests of fit

4. Tests of association or correlation
5. Tests of randomness

Tests can also be categorised by the design that they can be applied to:

1. One sample
2. Two related (paired) samples
3. Two independent samples
4.  $k(> 2)$  related (matched) samples
5.  $k(> 2)$  independent samples

A third classification of a test relates to the type of data to which it may be applied. Variables are recorded on four scales of measurement: nominal (categorical), ordinal, interval, and ratio.

The nominal scale is used only to categorise data; for each category a name, perhaps numeric, is assigned so that two different categories will be identified by distinct names. The ordinal scale, as well as categorising the observations, orders the categories. Each is assigned a distinct identifying symbol, in such a way that the order of the symbols corresponds to the order of the categories. (The most common system for ordinal variables is to assign numerical identifiers to the categories, though if they have previously been assigned alphabetic characters, these may be transformed to a numerical system by any convenient method which preserves the ordering of the categories.) The interval scale not only categorises and orders the observations, but also quantifies the comparison between categories; this necessitates a common unit of measurement and an arbitrary zero-point. Finally, the ratio scale is similar to the interval scale, except that it has an absolute (as opposed to arbitrary) zero-point.

It is apparent that there are many possible combinations of these three characteristics of a problem, and many nonparametric tests have been derived to meet the different experimental situations. However, it is not usually a difficult matter to choose an appropriate test given the nature of the data and the type of test which one wishes to perform.

## 2.3 Principles of Nonparametric Tests

In this section, each type of test is considered in turn, and remarks are made on the design principles on which each is based.

### 2.3.1 Location tests

These tests are primarily concerned with inferences about differences in the location of the population distributions. In some cases, however, the tests are only concerned with inferences about the population distributions unless added assumptions are made which allow the hypotheses to be stated in terms of the location parameters.

For most of these tests, data must be measured numerically on at least an ordinal scale, in order that a measure of location may be devised. Ordinal measurement implies that pairs of values may be compared and numerically ordered. A vector of  $n$  values may therefore be **ranked** from smallest to largest using the ordering operation. The resultant **ranks** contain all the information in the original data, but have the advantage that tests may be derived easily based on them, and no testing bias is introduced by the use of ordinal values as though they were measured on an interval scale. Note that the requirement of the measurement scale being ordinal does not imply that all tests of this type involve the actual ranking of the original data.

For the one-sample or matched pairs case, test statistics may be derived based on the number of observations (or differences) lying either side of zero (or some other fixed value), as in the sign test, for example. Under the hypothesis that the median of the single population is zero or the difference in the medians of the paired populations is zero the number of positive and negative values should be similar. The Wilcoxon signed rank test goes further than the sign test by taking into account the magnitude of the single sample values or of the differences.

For the two-sample case, if median equality is hypothesised, the distribution of the ranks of each sample in the total pooled sample should be similar. Test statistics, such as the Mann–Whitney  $U$  statistic, which are based on the ranks of each sample and summarize the differences in rank sums for each sample, may be

computed. These statistics are referred to their expected distributions under the null hypothesis. The above hypothesis can also be tested using the median test. Its test statistic is based on the number of values in each sample which are greater than or less than the pooled median of the two samples, rather than the ranks of each sample.

If median equality is hypothesised for several samples, the distribution ranks of the members of each sample in the total pooled sample should be 'homogeneous'. Test statistics can be derived which summarize the differences in rank sums for the various samples, and again referred to their expected distributions under the null hypothesis.

### 2.3.2 Dispersion tests

These provide a distribution-free alternative to such tests as the  $F$ -(variance-ratio) test for variance equality, which is very sensitive to non-Normality in the generating distribution.

The dispersions of two or more samples may be compared by pooling the samples and observing the distribution of ranks in the ranked pooled sample. Equal dispersions should be recognizable by there being a wide distribution of the extreme ranks between the members of different samples. Statistics are evaluated which quantify the dispersion of ranks between samples, and their significance may be found by evaluating their permutation distributions assuming that no dispersion difference exists.

### 2.3.3 Tests of fit

In the one-sample case, these are tests which investigate whether or not a sample of observations can be considered to follow a specified distribution. In the two-sample case, a test of fit investigates whether the two samples can be considered to have arisen from a common probability distribution.

For the one-sample problem, the null hypothesis may specify only the distributional form, for example  $\text{Normal}(\mu, \sigma^2)$ , or it may incorporate actual parameter values, for example Poisson with mean 10.

Some tests of this type proceed by forming the sample cumulative distribution function of the observations and computing a statistic whose value measures the departure of the sample cumulative distribution function from that of the null distribution. In the two-sample case, a statistic is computed which provides a measure of the difference between the sample cumulative distribution function of each sample. These tests are known as one- or two-sample Kolmogorov–Smirnov tests.

The significance for these test statistics can be computed directly for moderate sample sizes but for larger sample sizes asymptotic results are often used.

Another goodness-of-fit test is the  $\chi^2$  test. For this test, the data is first grouped into intervals and then the difference between the observed number of observations in each interval and the number expected, if the null hypothesis is true, is computed. A statistic based on these differences has asymptotically a  $\chi^2$ -distribution.

### 2.3.4 Association and correlation tests

These are distribution-free analogues of tests based on such statistics as Pearson product-moment correlation coefficients.

Essentially they are based on rankings rather than the observed data values, and involve summing some function of the rank differences between the samples to obtain an overall measure of the concordance of ranks. This measure can be standardized by dividing by its theoretical maximum value for the given sample size and number of samples.

Significance levels may be calculated for quite small sample sizes by using an approximation to a  $\chi^2$ -distribution.

### 2.3.5 Tests of randomness

These tests are designed to investigate sequences of observations and attempt to identify any deviations from randomness. There are clearly many ways in which a sequence may deviate from randomness. The tests provided here primarily detect some form of dependency between the observations in the sequence.

The most common application of this type of tests is in the area of random number generation. The tests are used as empirical tests on a sample of output from a generator to establish local randomness. Theoretical tests are necessary and useful for testing global randomness. Some of the more common empirical tests are discussed below.

A runs-up or runs-down test investigates whether runs of different lengths are occurring with greater or lesser frequency than would be expected under the null hypothesis of randomness. A run up is defined as a sequence of observations in which each observation is larger than the previous observation. The run up ends when an observation is smaller than the previous observation. A test statistic, modified to take into account the dependency between successive run lengths, is computed. The test statistic has an asymptotic  $\chi^2$ -distribution.

The pairs test investigates the condition that, under the null hypothesis of randomness, the non-overlapping 2-tuples (pairs) of a sequence of observations from the interval  $[0, 1]$  should be uniformly distributed over the unit square  $([0, 1]^2)$ . The triplets test follows the same idea but considers 3-tuples and checks for uniformity over the unit cube  $([0, 1]^3)$ . In each test, a test statistic, based on differences between the observed and expected distribution of the 2- or 3-tuples, is computed which has an asymptotic  $\chi^2$ -distribution.

The gaps test considers the ‘gaps’ between successive occurrences of observations in the sequence lying in a specified range. Under the null hypothesis of randomness, the gap length should follow a geometric distribution with a parameter based on the length of the specified range, relative to the overall length of the interval containing all possible observations. The expected number of ‘gaps’, of a certain length, under the null hypothesis may thus be computed together with a test statistic based on the differences between the observed and expected numbers of ‘gaps’ of different length. Again the test statistic has an asymptotic  $\chi^2$ -distribution.

Other empirical tests such as the  $\chi^2$  goodness-of-fit test and the one-sample Kolmogorov–Smirnov test may be used to investigate a sequence for non-uniformity.

## 2.4 Regression using ranks

If you wish to fit a regression model but is unsure about what transformation to take for the observed response to obtain a linear model, then one strategy is to replace response observations by their ranks. Estimates for regression parameters can be found by maximizing a likelihood function based on the ranks and the proposed regression model. The present routines give approximate estimates which are adequate when the signal-to-noise ratio is small, which is often the case with data from the medical and social sciences. Approximate standard errors of estimated regression coefficients are found. Also  $\chi^2$  statistics can be used to test the null hypothesis of no regression effect.

## 3 Recommendations on Choice and Use of Available Routines

The routines are grouped into six categories. The fourth character of the routine name is used to denote this categorisation.

Sub-chapter	Type of test
G08A	Location
G08B	Dispersion
G08C	Fit
G08D	Correlation and Association
G08E	Randomness
G08R	Regression using Ranks

### 3.1 G08A – Location Tests

#### 3.1.1 One-sample or matched-pairs case

Note that a random sample of matched pairs,  $(x_i, y_i)$ , may be reduced to a single sample by considering the differences,  $d_i = x_i - y_i$  say, of each pair. The matched pair may be thought of as a single observation on a bivariate random variable.

G08AAF performs the sign test on two paired samples. Each pair is classified as a + or – depending on the sign of the difference between the two data values within the pair. Under the assumptions that the  $d_i$  are mutually independent and that the observations are measured on at least an ordinal scale, the sign test tests the hypothesis that for any pair sampled from the population distribution,  $\text{Probability}(+) = \text{Probability}(-)$ . The hypothesis may be stated in terms of the equality of the location parameters but the test is no longer regarded as unbiased and consistent unless further assumptions are made. If you wish to test the hypothesis that the location parameters differ by a fixed amount then that amount must be added or subtracted from one of the samples as required before calling G08AAF.

G08AGF performs the one-sample Wilcoxon signed-rank test. The test may be used to test if the median of the population from which the random sample was taken is equal to some specified value (commonly used to test if the median is zero). In this test not only is the sign of the difference between the data values and the hypothesised median value important but also the magnitude of this difference. Thus, where the magnitude of the differences (or the data values themselves if the hypothesised median value is zero) is important this test is preferred to the sign test because it is more powerful. The test may easily be used to test whether the medians of two related populations are equal by taking the differences between the paired sample values and then testing the hypothesis that the median of the differences is zero, using the single sample of differences. The significance of the test statistic may be computed exactly for a moderate sample size but for a larger sample a Normal approximation is used. The exact method allows for ties in the differences.

### 3.1.2 Two independent samples

G08ACF performs the median test and G08AHF performs the Mann–Whitney  $U$  test.

For both tests the two samples are assumed to be random samples from their respective populations and mutually independent. The measurement scale must be at least ordinal.

Note that, although the median test may be generalized to more than two samples, G08ACF only deals with the two-sample case. For the median test, each observation is classified as being above or below the pooled median of the two samples. It may be used to test the hypothesis that the two population medians are equal; under the assumption that if the two population medians are equal then the probability of an observation exceeding the pooled median is the same for both populations.

The Mann–Whitney  $U$  test involves the ranking of the pooled sample. The Mann–Whitney test thus attaches importance to the position of each observation relative to the others and not just its position relative to the median of the pooled sample as in the median test. Thus when the magnitude of the differences between the observations is meaningful the Mann–Whitney  $U$  test is preferred as it is more powerful than the median test. The test tests whether the two population distributions are the same or not. If it is assumed that any difference between the two population distributions is a difference in the location then the test is testing whether the population means are the same or not.

In G08AHF, the significance of the  $U$  test statistic is computed using a Normal approximation. If the exact significance is desired then either G08AJF or G08AKF must be used. G08AJF computes the exact significance of the  $U$  test statistic for the case where there are no ties in the pooled sample. It requires only the value of the statistic and the two sample sizes. G08AKF computes the exact significance of the  $U$  test statistic for the case where there are ties in the pooled sample. It requires the value of the statistic and the two sample sizes and the ranks of the observations of the two samples as provided by G08AHF. G08AHF returns an indicator to inform you whether or not ties were found in the pooled sample.

### 3.1.3 More than two related samples

G08AEF performs the Friedman two-way analysis of variance. This test may in some ways be regarded as an extension of the sign test to the case of  $k$  ( $k > 2$ ) related samples. The data is in the form of a number of multivariate observations which are assumed to be mutually independent. This test also assumes that the measurement within each observation across the  $k$  variates is at least ordinal so that the observation for each variate may be ranked according to some criteria.

For data which may be defined as either a zero or one, that is binary response data, G08ALF performs Cochran's  $Q$ -test to examine differences between the treatments within blocks.

### 3.1.4 More than two independent samples

G08AFF performs the Kruskal–Wallis one-way analysis of variance. The test assumes that each sample is a random sample from its respective distributions and in addition that there is both independence within the samples and mutual independence among the various samples. The test requires that the measurement scale is at least ordinal so that the pooled sample may be ranked.

## 3.2 G08B – Dispersion Tests

G08BAF performs either Mood’s or David’s test for dispersion differences, or both, for two independent samples of possibly unequal size.

For both tests the null hypothesis is that the two samples have equal dispersions, the routine returning a probability value which may be used to perform the test against a one-sided or two-sided alternative, in a way described in the routine document.

## 3.3 G08C – Tests of Fit

G08CBF and G08CCF both perform the one-sample Kolmogorov–Smirnov distribution test. This test is used to test the null hypothesis that the random sample arises from a specified null distribution against one of three possible alternatives.

With G08CBF you may choose a null distribution from one of the following: the uniform, Normal, gamma, beta, binomial, exponential, and Poisson. The parameter values may either be specified by you or estimated from the data by the routine. With G08CCF you must provide a function which will compute the value of the cumulative distribution function at any specified point for the null distribution. The alternative hypotheses available correspond to one- and two-sided tests. The distribution of the test statistic is computed using an exact method for a moderate sample size. For a larger sample size an asymptotic result is used.

G08CDF performs the two-sample Kolmogorov–Smirnov test which tests the null hypothesis that the two samples may be considered to have arisen from the same population distribution against one of three possible alternative hypotheses, again corresponding to one-sided and two-sided tests. The distribution of the test statistic is computed using an exact method for moderate sample sizes, but for larger samples approximations based on asymptotic results are used.

Note that G01EYF and G01EZF are available for computing the distributions of the one-sample and two-sample Kolmogorov–Smirnov statistics respectively.

G08CGF performs the  $\chi^2$  goodness-of-fit test on a single sample which again tests the null hypothesis that the sample arises from a specified null distribution. You may choose a null distribution from one of the following: the Normal, uniform, exponential,  $\chi^2$ , and gamma; or may define the distribution by specifying the probability that an observation lies in a certain interval for a range of intervals covering the support of the null distribution. The significance of this test is computed using the  $\chi^2$ -distribution as an approximation to the distribution of the test statistic.

Tests of Normality may also be carried out using routines in Chapter G01.

## 3.4 G08D – Association and Correlation Tests

G08DAF computes Kendall’s coefficient of concordance on  $k$  independent ranks of  $n$  objects. An example of its application would be to compare for consistency the results of a group of IQ tests performed on the same set of people. Allowance is made for tied rankings, and the approximate significance of the computed coefficient is found.

## 3.5 G08E – Tests of Randomness

G08EAF performs the runs-up test on a sequence of observations. The runs-down test may be performed by multiplying each observation by  $-1$  before calling the routine. All runs whose length is greater than or equal to a certain chosen length will be treated as a single group.

G08EBF performs the pairs (serial) test on a sequence of observations from the interval  $[0, 1]$ . The number of equal sub-intervals into which the interval  $[0, 1]$  is to be divided must be specified.

G08ECF performs the triplets test on a sequence of observations from the interval  $[0, 1]$ . The number of equal sub-intervals into which the interval  $[0, 1]$  is to be divided must be specified.

G08EDF performs the gaps test on a sequence of observations. The total of the interval containing all possible values the observations could take must be specified together with the interval being used to define the ‘gaps’. All ‘gaps’ whose length is greater than or equal to a certain chosen length will be treated as a single group.

### 3.6 G08R – Regression Using Ranks

G08RAF fits a multiple linear regression model in which the observations on the response variable are replaced by their ranks.

G08RBF performs the same function but takes into account observations which may be right-censored.

### 3.7 Related Routines

Tests of location and distribution may be based on scores which are estimates of the expected values of the order statistics. G01DHF may be used to compute Normal scores, an approximation to the Normal scores (Blom, Tukey or van der Waerden scores) or Savage (exponential) scores. For more accurate Normal scores G01DAF may be used. Other routines in sub-chapter G01D may be used to test for Normality.

## 4 Index

Regression using ranks:

right-censored data .....	G08RBF
uncensored data .....	G08RAF

Tests of association and correlation:

Kendall’s coefficient of concordance.....	G08DAF
---	--------

Tests of dispersion:

Mood’s and David’s tests on two samples of unequal size .....	G08BAF
---	--------

Tests of fit:

Kolmogorov–Smirnov one-sample distribution test:	
for a user-supplied distribution.....	G08CCF
for standard distributions.....	G08CBF
Kolmogorov–Smirnov two-sample distribution test.....	G08CDF
$\chi^2$ goodness of fit test .....	G08CGF

Tests of location:

Cochran Q test on cross-classified binary data .....	G08ALF
Exact probabilities for Mann–Whitney $U$ statistic:	
no ties in pooled sample.....	G08AJF
ties in pooled sample .....	G08AKF
Friedman two-way analysis of variance on $k$ matched samples .....	G08AEF
Kruskal–Wallis one-way analysis of variance on $k$ samples of unequal size .....	G08AFF
Mann–Whitney $U$ test on two samples of possibly unequal size.....	G08AHF
Median test on two samples of unequal size .....	G08ACF
Sign test on two paired samples .....	G08AAF
Wilcoxon one sample signed rank test .....	G08AGF

Tests of randomness:

Gaps test .....	G08EDF
Pairs (serial) test.....	G08EBF
Runs up or runs down test.....	G08EAF
Triplets test .....	G08ECF

## 5 Routines Withdrawn or Scheduled for Withdrawal

Withdrawn Routine	Mark of Withdrawal	Replacement Routine(s)
G08ABF	16	G08AGF
G08ADF	16	G08AHF, G08AJF and G08AKF
G08CAF	16	G08CBF

## 6 References

Conover W J (1980) *Practical Nonparametric Statistics* Wiley

Kendall M G and Stuart A (1973) *The Advanced Theory of Statistics (Volume 2)* (3rd Edition) Griffin

Siegel S (1956) *Non-parametric Statistics for the Behavioral Sciences* McGraw-Hill

---