Chapter 13

Introduction to Nonparametric Analysis

Chapter Table of Contents

Chapter 13
Introduction to Nonparametric Analysis

Overview

In statistical inference, or hypothesis testing, the traditional tests are called *parametric tests* because they depend on the specification of a probability distribution (such as the normal) except for a set of free parameters. Parametric tests are said to depend on distributional assumptions. *Nonparametric tests*, on the other hand, do not require any strict distributional assumptions. Even if the data are distributed normally, nonparametric methods are often almost as powerful as parametric methods.

Many nonparametric methods analyze the ranks of a variable rather than the original values. Procedures such as PROC NPAR1WAY calculate the ranks for you and then perform appropriate nonparametric tests. However, there are some situations in which you use a procedure such as PROC RANK to calculate ranks and then use another procedure to perform the appropriate test. See the section "Obtaining Ranks" on page 184 for details.

Although the NPAR1WAY procedure is specifically targeted for nonparametric analysis, many other procedures also perform nonparametric analyses. Some general references on nonparametrics include Lehman (1975), Conover (1980), Hollander and Wolfe (1973), Hettmansperger (1984), and Gibbons and Chakraborti (1992).

Testing for Normality

Many parametric tests assume an underlying normal distribution for the population. If your data do not meet this assumption, you may prefer to use a nonparametric analysis.

Base SAS software provides several tests for normality in the UNIVARIATE procedure. Depending on your sample size, PROC UNIVARIATE performs the Kolmogorov-Smirnov, Shapiro-Wilk, Anderson-Darling, and Cramér-von Mises tests. For more on PROC UNIVARIATE, see the *SAS Procedures Guide*.

Comparing Distributions

To test the hypothesis that two or more groups of observations have identical distributions, use the NPAR1WAY procedure. The procedure calculates the Kolmogorov-Smirnov statistic, an asymptotic Kolmogorov-Smirnov statistic, and the Cramér-von Mises statistic. In addition, for data with only two groups of observations, the procedure calculates the two-sample Kolmogorov statistic and the Kuiper statistic. To obtain these tests, use the EDF option in the PROC NPAR1WAY statement. For details, see Chapter 47, “The NPAR1WAY Procedure.”

One-Sample Tests

Base SAS software provides two one-sample tests in the UNIVARIATE procedure: a sign test and the Wilcoxon signed rank test. Both tests are designed for situations where you want to make an inference about the location (median) of a population. For example, suppose you want to test if the median resting pulse rate of marathon runners differs from a specified value.

By default, both of these tests examine the hypothesis that the median of the population from which the sample is drawn is equal to a specified value, which is zero by default. The Wilcoxon signed rank test requires that the distribution be symmetric; the sign test does not require this assumption. These tests can also be used for the case of two related samples; see the section “Comparing Two Independent Samples” for more information.

The two tests are automatically provided by the UNIVARIATE procedure. For details, formulas, and examples, see the chapter on the UNIVARIATE procedure in the *SAS Procedures Guide*.

Two-Sample Tests

This section describes tests appropriate for two independent samples (for example, two groups of subjects given different treatments) and for two related samples (for example, before-and-after measurements on a single group of subjects). Related samples are also referred to as paired samples or matched pairs.

Comparing Two Independent Samples

SAS/STAT software provides several nonparametric tests for location and scale differences.

When you perform these tests, your data should consist of a random sample of observations from two different populations. Your goal is either to compare the location parameters (medians) or the scale parameters of the two populations. For example, suppose your data consist of the number of days in the hospital for two groups of patients: those who received a standard surgical procedure and those who received a new, experimental surgical procedure. These patients are a random sample from the population of patients who have received the two types of surgery. Your goal is to decide whether the median hospital stays differ for the two populations.

Tests in the NPAR1WAY Procedure

The NPAR1WAY procedure provides the following location tests: Wilcoxon rank sum test (Mann-Whitney U test), Median test, Savage test, and Van der Waerden test. Also note that the Wilcoxon rank sum test can be obtained from the FREQ procedure. In addition, PROC NPAR1WAY produces the following tests for scale differences: Siegel-Tukey test, Ansari-Bradley test, Klotz test, and Mood test.

When data are sparse, skewed, or heavily tied, the usual asymptotic tests may not be appropriate. In these situations, exact tests may be suitable for analyzing your data. The NPAR1WAY procedure can produce exact p-values for all of the two-sample tests for location and scale differences.

Chapter 47, “The NPAR1WAY Procedure,” provides detailed statistical formulas for these statistics, as well as examples of their use.

Tests in the FREQ Procedure

This procedure provides a test for comparing the location of two groups and for testing for independence between two variables.

The situation in which you want to compare the location of two groups of observations corresponds to a table with two rows. In this case, the asymptotic Wilcoxon rank sum test can be obtained by using SCORES=RANK in the TABLES statement and by looking at either of the following:

- the Mantel-Haenszel statistic in the list of tests for no association. This is labeled as “Mantel Haenszel Chi-square” and PROC FREQ displays the statistic, the degrees of freedom, and the p-value.
- the CMH statistic 2 in the section on Cochran-Mantel-Haenszel statistics. PROC FREQ displays the statistic, the degrees of freedom, and the p-value. To obtain this statistic, specify the CMH2 option in the TABLES statement.

When you test for independence, the question being answered is whether the two variables of interest are related in some way. For example, you might want to know if student scores on a standard test are related to whether students attended a public or private school. One way to think of this situation is to consider the data as a two-way table; the hypothesis of interest is whether the rows and columns are independent. In the preceding example, the groups of students would form the two rows, and the scores would form the columns. The special case of a two-category response (Pass/Fail) leads to a 2×2 table; the case of more than two categories for the response (A/B/C/D/F) leads to a $2 \times c$ table, where c is the number of response categories.

For testing whether two variables are independent, PROC FREQ provides Fisher's exact test. For a 2×2 table, PROC FREQ automatically provides Fisher's exact test when you use the CHISQ option in the TABLES statement. For a $2 \times c$ table, use the EXACT option in the TABLES statement to obtain the test.

Comparing Two Related Samples

SAS/STAT software provides the following nonparametric tests for comparing the locations of two related samples:

- Wilcoxon signed rank test
- sign test
- McNemar's test

The first two tests are available in the UNIVARIATE procedure, and the last test is available in the FREQ procedure. When you perform these tests, your data should consist of pairs of measurements for a random sample from a single population. For example, suppose your data consist of SAT scores for students before and after attending a course on how to prepare for the SAT. The pairs of measurements are the scores before and after the course, and the students should be a random sample of students who attended the course. Your goal in analysis is to decide if the median change in scores is significantly different from zero.

Tests in the UNIVARIATE Procedure

By default, PROC UNIVARIATE performs a Wilcoxon signed rank test and a sign test. To use these tests on two related samples, perform the following steps:

1. In the DATA step, create a new variable that contains the differences between the two related variables.
2. Run PROC UNIVARIATE, using the new variable in the VAR statement.

For discussion of the tests, formulas, and examples, see the chapter on the UNIVARIATE procedure in the *SAS Procedures Guide*.

Tests in the FREQ Procedure

The FREQ procedure can be used to obtain McNemar's test, which is simply another special case of a Cochran-Mantel-Haenszel statistic (and also of the sign test). The AGREE option in the TABLES statement produces this test for 2×2 tables, and exact p-values are available for this test.

Tests for k Samples

Comparing k Independent Samples

One goal in comparing k independent samples is to determine whether the location parameters (medians) of the populations are different. Another goal is to determine whether the scale parameters for the populations are different. For example, suppose new employees are randomly assigned to one of three training programs. At the end of the program, the employees receive a standard test that gives a rating score of their job ability. The goal of analysis is to compare the median scores for the three groups and decide whether the differences are real or due to chance alone.

To compare k independent samples, either the NPAR1WAY or the FREQ procedure provides a Kruskal-Wallis test. PROC NPAR1WAY also provides the Savage, median, and Van der Waerden tests. In addition, PROC NPAR1WAY produces the following tests for scale differences: Siegel-Tukey test, Ansari-Bradley test, Klotz test, and Mood test. Note that you can obtain exact p-values for all of these tests.

In addition, you can specify the SCORES=DATA option to use the input data observations as scores. This enables you to produce a very wide variety of tests. You can construct any scores using the DATA step, and then PROC NPAR1WAY computes the corresponding linear rank and one-way ANOVA tests. You can also analyze the raw data with the SCORES=DATA option; for two-sample data, this permutation test is known as Pitman's test.

See Chapter 47, "The NPAR1WAY Procedure," for details, formulas, and examples.

To produce a Kruskal-Wallis test in the FREQ procedure, use SCORES=RANK and the CMH2 option in the TABLES statement. Then, look at the second Cochran-Mantel-Haenszel statistic (labeled "Row Mean Scores Differ") to obtain the Kruskal-Wallis test. The FREQ procedure also provides the Jonckheere-Terpstra test, which is more powerful than the Kruskal-Wallis test for comparing k samples against ordered alternatives. The exact test is also available. In addition, you can obtain a ridit analysis, developed by Bross (1958), by specifying SCORES=RIDIT or SCORES=MODRIDIT in the TABLES statement in the FREQ procedure.

Comparing k Dependent Samples

Friedman’s test enables you to compare the locations of three or more dependent samples. You can obtain Friedman’s Chi-square with the FREQ procedure by using the CMH2 option and SCORES=RANK and looking at the second CMH statistic in the output. For an example, see Chapter 28, “The FREQ Procedure”; this chapter also contains formulas and other details on the CMH statistics. For a discussion of how to use the RANK and GLM procedures to obtain Friedman’s test, see Ipe (1987).

Measures of Correlation and Associated Tests

The CORR procedure in base SAS software provides several nonparametric measures of association and associated tests. It computes Spearman’s rank-order correlation, Kendall’s tau-b, and Hoeffding’s measure of dependence, and it provides tests for each of these statistics. PROC CORR also computes Spearman’s partial rank-order correlation and Kendall’s partial tau-b. Finally, PROC CORR computes Cronbach’s coefficient alpha for raw and standardized variables. This statistic can be used to estimate the reliability coefficient. For a general discussion of correlations, formulas, interpretation, and examples, see the chapter on the CORR procedure in the *SAS Procedures Guide*.

The FREQ procedure also provides some nonparametric measures of association: gamma, Kendall’s tau-b, Stuart’s tau-c, Somer’s D, and the Spearman rank correlation. The output includes the measure, the asymptotic standard error, confidence limits, and the asymptotic test that the measure equals zero. For the Spearman rank correlation, you can optionally request an exact p-value that the correlation is equal to zero.

Obtaining Ranks

The primary procedure for obtaining ranks is the RANK procedure in base SAS software. Note that the PRINQUAL and TRANSREG procedures also provide rank transformations. With all three of these procedures, you can create an output data set and use it as input to another SAS/STAT procedure or to the IML procedure. See the *SAS Procedures Guide* for information on the RANK procedure, and see the chapters in this book for information on the PRINQUAL and TRANSREG procedures.

In addition, you can specify SCORES=RANK in the TABLES statement in the FREQ procedure. PROC FREQ then uses ranks to perform the analyses requested and generates nonparametric analyses.

For more discussion of using the rank transform, see Iman and Conover (1979), Conover and Iman (1981), Hora and Conover (1984), Iman, Hora, and Conover (1984), Hora and Iman (1988), and Iman (1988).

Kernel Density Estimation

The KDE procedure performs either univariate or bivariate kernel density estimation. Statistical *density estimation* involves approximating a hypothesized probability density function from observed data. *Kernel density estimation* is a nonparametric technique for density estimation in which a known density function (the kernel) is averaged across the observed data points to create a smooth approximation.

PROC KDE uses a Gaussian density as the kernel, and its assumed variance determines the smoothness of the resulting estimate. PROC KDE outputs the kernel density estimate to a SAS data set, which you can then use with other procedures for plotting or analysis. PROC KDE also computes a variety of common statistics, including estimates of the percentiles of the hypothesized probability density function.

References

Bross, I.D.J. (1958), "How to Use Ridit Analysis," *Biometrics*, 14, 18–38.

Conover, W.J. (1980), *Practical Nonparametric Statistics*, Second Edition, New York: John Wiley & Sons, Inc.

Conover, W.J. and Iman, R.L. (1981), "Rank Transformations as a Bridge between Parametric and Nonparametric Statistics," *The American Statistician*, 35, 124–129.

Gibbons, J.D. and Chakraborti, S. (1992), *Nonparametric Statistical Inference*, Third Edition, New York: Marcel Dekker, Inc.

Hettmansperger, T.P. (1984), *Statistical Inference Based on Ranks*, New York: John Wiley & Sons, Inc.

Hollander, M. and Wolfe, D.A. (1973), *Nonparametric Statistical Methods*, New York: John Wiley & Sons, Inc.

Hora, S.C. and Conover, W.J. (1984), "The F Statistic in the Two-Way Layout with Rank-Score Transformed Data," *Journal of the American Statistical Association*, 79, 668–673.

Hora, S.C. and Iman, R.L. (1988), "Asymptotic Relative Efficiencies of the Rank-Transformation Procedure in Randomized Complete Block Designs," *Journal of the American Statistical Association*, 83, 462–470.

Iman, R.L. and Conover, W.J. (1979), "The Use of the Rank Transform in Regression," *Technometrics*, 21, 499–509.

Iman, R.L., Hora, S.C., and Conover, W.J. (1984), "Comparison of Asymptotically Distribution-Free Procedures for the Analysis of Complete Blocks," *Journal of the American Statistical Association*, 79, 674–685.

Iman, R.L. (1988), “The Analysis of Complete Blocks Using Methods Based on Ranks,” *Proceedings of the Thirteenth Annual SAS Users Group International Conference*, 13, 970–978.

Ipe, D. (1987), “Performing the Friedman Test and the Associated Multiple Comparison Test Using PROC GLM,” *Proceedings of the Twelfth Annual SAS Users Group International Conference*, 12, 1146–1148.

Lehmann, E.L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.