Paper TU04

An Overview of Non-parametric Tests in SAS®: When, Why, and How

Paul A. Pappas and Venita DePuy
Duke Clinical Research Institute
Durham, North Carolina, USA

ABSTRACT

Most commonly used statistical procedures are based on the assumption that distributions are normal. But what if they are not? When should non-parametric statistics be used, and what assumptions do those require? If your data doesn't meet the assumptions for either group, what should you do?

This paper will provide an easy guide to choosing the most appropriate statistical test, whether parametric or non-parametric; how to perform that-test in SAS; and how to interpret the results. Non-parametric SAS procedures covered will include PROCs ANOVA, NPAR1WAY, TTEST and UNIVARIATE. Discussion will include assumptions and assumption violations, robustness, and exact versus approximate tests.

INTRODUCTION

Many statistical tests rely heavily on distributional assumptions, such as normality. When these assumptions are not satisfied, commonly used statistical tests often perform poorly, resulting in a greater chance of committing an error. Non-parametric tests are designed to have desirable statistical properties when few assumptions can be made about the underlying distribution of the data. In other words, when the data are obtained from a non-normal distribution or one containing outliers, a non-parametric test is often a more powerful statistical tool than it's parametric 'normal theory' equivalent. We will explore the use of parametric and non-parametric tests for one- and two-sample location differences, two-sample dispersion differences, and a one way layout analysis. We will also examine testing for general differences between populations.

MEASURES OF LOCATION AND SPREAD

Mean and variance are typically used to describe the center and spread of normally distributed data. If the data are not normally distributed or contain outliers, these measures may not be robust enough to accurately describe the data. The median is a more robust measure of the center of distribution, in that it is not as heavily influenced by outliers and skewed data. As a result, the median is typically used with non-parametric tests. The spread is less easy to quantify but is often represented by the interquartile range, which is simply the difference between the first and third quartiles.

DETERMINING NORMALITY (OR LACK THEREOF)

One of the first steps in test selection should be investigating the distribution of the data. PROC UNIVARIATE can be implemented to determine whether or not your data are normal. This procedure generates a variety of summary statistics, such as the mean and median, as well as numerical representations of properties such as skewness and kurtosis.

If the population from which the data are obtained is normal, the mean and median should be equal or close to equal. The skewness coefficient, which is a measure of symmetry, should be near zero. Positive values for the skewness coefficient indicate that the data are right skewed, and negative values indicate that that data are left skewed. The kurtosis coefficient, which is a measure of spread, should also be near zero. Positive values for the kurtosis coefficient indicate that the distribution of the data is steeper than a normal distribution, and negative values for kurtosis indicate that the distribution of the data is flatter than normal distribution. The NORMAL option in PROC UNIVARIATE produces a table with tests for normality. In general, if the p-values are less than 0.05, then the data should be considered non-normally distributed. However, it is important to remember that these tests are heavily dependent on sample size. Strikingly non-normal data may have a p-value greater than 0.05 due to a small sample size. Therefore, graphical representations of the data should always be examined.

Low resolution plots and high resolution histograms are both available in PROC UNIVARIATE. The PLOTS option in PROC UNIVARIATE creates low-resolution stem-and-leaf, box, and normal probability plots. The stem-and-leaf plot is used to visualize the overall distribution of the data and the box plot is a graphical representation of the 5-number summary. The normal probability plot is designed to investigate whether a variable is normally distributed. If the data are normal, then the plot should display a straight diagonal line. Different departures from the straight diagonal line indicate different types of departures from normality.

The HISTOGRAM statement in PROC UNIVARIATE will produce high resolution histograms. When used in conjunction with the NORMAL option, the histogram will have a line indicating the shape of a normal distribution with the same mean and variance as the sample.

PROC UNIVARIATE is an invaluable tool in visualizing and summarizing data in order to gain an understanding of the underlying populations from which the data are obtained. To produce these results, the following code can be used. Omitting the VAR statement will run the analysis on all the variables in the dataset.

```
PROC UNIVARIATE data=datafile normal plots;
      Histogram;
      Var variable1 variable2 ... variablen;
Run;
```

The determination of the normality of the data should result from evaluation of the graphical output in conjunction with the numerical output. In addition, the user might wish to look at subsets of the data; for example, a CLASS statement might be used to stratify by gender.

DIFFERENCES IN DEPENDENT POPULATIONS

Testing for the difference between two dependent populations, such as before and after measurements on the same subjects, is typically done by testing for a difference in centers of distribution (means or medians). The paired t-test looks for a difference in means, while the non-parametric sign and signed rank tests look for a difference in medians. It is assumed that the spreads and shapes of the two populations are the same. In this situation, the data are paired; two observations are obtained on each of *n* subjects resulting in one sample of *2n* observations.

PAIRED T-TEST

If the data are normal, the one-sample paired t-test is the best statistical test to implement. Assumptions for the one-sample paired t-test are that the observations are independent and identically normally distributed. To perform a one-sample paired t-test in SAS, either PROC UNIVARIATE (as described above) or the following PROC TTEST code can be used.

```
PROC TTEST data=datafile;
      Paired variable1*variable2;
Run;
```

If the p-value for the paired t-test is less than the specified alpha level, then there is evidence to suggest that the population means of the two variables differ. In the case of measurements taken before and after a treatment, this would suggest a treatment effect.

SIGN TEST AND SIGNED RANK TEST

The Signed Rank test and the Sign test are non-parametric equivalents to the one-sample paired t-test. Neither of these tests requires the data to be normally distributed, but both tests require that the observed differences between the paired observations be mutually independent, and that each of the observed paired differences comes from a continuous population symmetric about a common median. However, the observed paired differences do not necessarily have to be obtained from the same underlying distribution.

In general, the signed rank test has more statistical power than the sign test, but if the data contain outliers or are obtained from a heavy-tailed distribution, the sign test will have the most statistical powerful. PROC UNIVARIATE produces both of these tests as well as a paired t-test. Before PROC UNIVARIATE can be used to carry out these tests, the paired differences must be computed, as per the following code:

```
DATA datafile;
      Set datafile;
      Difference=variable1-variable2;
PROC UNIVARIATE data=datafile;
      Var difference;
Run;
```

If the p-values for the signed rank or sign tests are less than the specified alpha level, then there is evidence to suggest that the population medians of the two variables differ.

DIFFERENCES BETWEEN TWO INDEPENDENT POPULATIONS

Investigators may be interested in the difference between centers of distributions, difference in distributional spreads, or simply interested in any differences between two populations. PROC TTEST allows the user to test for differences in means for both equal and unequal variances, as well as providing a test for differences in variance. The non-parametric method provides different-tests, depending on the hypothesis of interest. The Wilcoxon Rank Sum test

investigates differences in medians, with the assumption of identical spreads. The more generalized Kolmogorov-Smirnov test looks for differences in center, spread, or both. It should be noted that "two stage" calculations should not be used. For instance, using the Wilcoxon Rank Sum test for differences in medians followed by the Ansari-Bradley test for differences in spread dos not protect the nominal significance level. When overall differences between populations need to be investigated, consider the Kolmogorov-Smirnov test.

TWO-SAMPLE T-TEST

If the data consists of two samples from independent, normally distributed populations, the two-sample t-test is the best statistical test to implement. Assumptions for the two-sample t-test and Folded F test are that within each sample the observations are independent and identically normally distributed. Also, the two samples must be independent of each other. To perform these tests in SAS, the following code can be used:

```
PROC TTEST data=datafile;
        Class sample;
        Var variable1;
Run;
```

The CLASS statement is necessary in order to distinguish between the two samples. The output from these statements includes the two-sample t-test for both equal and unequal variances, as well as the Folded F test for equality of variances. The equality of variance results should be reviewed first, and used to determine which results from the t-test are applicable. If the p-value for the t-test is less than the specified alpha level, then there is evidence to suggest that the two population means differ.

WILCOXON RANK SUM TEST

The Wilcoxon Rank Sum (which is numerically equivalent to the Mann-Whitney U test) is the non-parametric equivalent to the two-sample t-test. This test can also be performed if only rankings (i.e., ordinal data) are available. It-tests the null hypothesis that the two distributions are identical against the alternative hypothesis that the two distributions differ only with respect to the median. Assumptions for this test are that within each sample the observations are independent and identically distributed, and that the shapes and spreads of the distributions are the same. It does not matter which distribution the data are obtained from as long as the data are randomly selected from the same underlying population. Also, the two samples must be independent of each other. The following code will perform the rank sum test in SAS:

```
PROC NPAR1WAY data=datafile wilcoxon;
       Class sample;
       Var variable1;
       Exact; *OPTIONAL;
Run;
```

The Wilcoxon Rank Sum test will be performed without using the WILCOXON option; however, this option limits the amount of other output. The EXACT statement is optional and requests exact-tests to be performed. Approximate tests perform well when the sample size is sufficiently large. When sample size is small or the data are sparse, skewed, or heavy-tailed, approximate tests may be unreliable. However, exact-tests are computationally intensive which may require computer algorithms to run for hours or days. Thus, if your sample size is sufficiently large, it may not be worth your while to request exact-tests. If the p-value is less than the specified alpha level, then there is evidence to suggest that the two population medians differ.

It should be noted that when the variances of the treatment groups are heterogeneous, the Type I error probabilities in the Wilcoxon Rank Sum test can increase by as much as 300%, with no indication that they asymptotically approach the nominal significance level as the sample size increases. Therefore, care should be taken when assuming that variances are, in fact, equal.

ANSARI-BRADLEY TEST

In some instances, it may be necessary to test for differences in spread while assuming that the centers of two populations are identical. One example is comparing two assay methods to see which is more precise. The Ansari-Bradley test is the non-parametric equivalent to the Folded F test for equality of variances. Assumptions for this test are that within each sample the observations are independent and identically distributed. Also, the two samples must be independent of each other, with equal medians. To perform an Ansari-Bradley test in SAS, the following code can be used:

```
PROC NPAR1WAY data=datafile ab;
       Class sample;
       Var variable;
Run;
```

The EXACT option may also be used if sample sizes are small. If the p-value is less than the specified alpha level, then there is evidence to suggest that the spreads of the two populations are not identical.

KOLMOGOROV-SMIRNOV TEST

In many cases, the Kolmogorov-Smirnov may be the most appropriate of the non-parametric tests for overall differences between two groups. If it can be assumed that spreads and shapes of the two distributions are the same, the Wilcoxon Rank Sum test is more powerful than the Kolmogorov-Smirnov test; if the median and shapes are the same, the Ansari-Bradley test is more powerful. This test is for the general hypothesis that the two populations differ. The assumptions are that the data is independent and identically distributed within the two populations, and that the two samples are independent. To perform this test in SAS, use the following code:

```
PROC NPAR1WAY data=datafile edf;
      Class sample;
      Var variable;
      Exact; *OPTIONAL;
Run;
```

As with the other non-parametric tests, the EXACT option is available in the Kolmogorov-Smirnov test, but is only recommended for small sample sizes or sparse, skewed or heavy tailed data. If the p-value is less than the specified alpha level, then there is evidence to suggest that the two populations are not the same.

DIFFERENCES IN THREE OR MORE INDEPENDENT POPULATIONS

In this situation the data consist of more than two random samples.

ONE-WAY ANALYSIS TEST

If the data are normal, then a one-way ANOVA is the best statistical test to implement. Assumptions for a one-way ANOVA are that within each sample the observations are independent and identically normally distributed. Also, the samples must be independent of each other with equal population variances. To perform a one-way ANOVA in SAS, the following code can be used:

```
PROC ANOVA data=datafile;
      Class sample;
      Model variable=sample;
Run;
```

If the p-value is less than the specified alpha level, then there is evidence to suggest that at least one of the population means differs from the others. Further investigation is required to determine which specific population means can be considered statistically significantly different from each other.

KRUSKAL-WALLIS TEST

The Kruskal-Wallis test is the non-parametric equivalent to the one-way ANOVA. Assumptions for the Kruskal-Wallis test are that within each sample the observations are independent and identically distributed and the samples are independent of each other. To perform a Kruskal-Wallis test in SAS, the following code can be used:

```
PROC NPAR1WAY data=datafile wilcoxon;
      Class sample;
      Var variable;
      Exact; *OPTIONAL;
Run;
```

This block of code is identical to the code used to produce the Wilcoxon Rank Sum test. In fact, the Kruskal-Wallis test reduces to the rank sum test when there are only two samples. The EXACT statement is optional and requests exact-tests to be performed in addition to the large-sample approximation. As before, the exact option is very computationally intensive and should only be used if needed. If the p-value is less than the specified alpha level, then there is evidence to suggest that at least one of the population medians differs from the others. Further investigation is required to determine which specific population medians can be considered statistically significantly different from each other.

CONCLUSION

Statistical analyses may be invalid if the assumptions behind those tests are violated. Prior to conducting analyses, the distribution of the data should be examined for departures from normality, such as skewness or outliers. If the data are normally distributed, and other assumptions are met, parametric tests are the most powerful. If the data are non-normal but other criteria are met, non-parametric statistics provide valid analyses. When neither set of assumptions has been met, both tests should be implemented to see if they agree. Also, further research should be done to discover whether the appropriate parametric or non-parametric test is the most robust to specific data issues.

This paper discusses the most commonly used non-parametric tests, and their parametric equivalents. A variety of other non-parametric tests are available in SAS, both in PROC NPAR1WAY and in other procedures such as PROC FREQ. Further technical information is available in the SAS Online Docs.

REFERENCES

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

SAS Institute (2003). SAS OnlineDoc, Version 8, Cary, NC: SAS Institute, Inc. 1999.

UCLA Academic Technology Services. "SAS Class Notes 2.0 – Analyzing Data" <http://www.ats.ucla.edu/stat/sas/> (July 9, 2004)

Zimmerman, D.W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67,* 55-68.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Paul A. Pappas
Duke Clinical Research Institute
PO Box 17969
Durham, NC 27715
(919) 668-8542
Fax: (919) 668-7053
paul.pappas@duke.edu

Venita DePuy
Duke Clinical Research Institute
PO Box 17969
Durham, NC 27715
(919) 668-8087
Fax: (919) 668-7124
venita.depuy@duke.edu