SAS/STAT® 9.3 User's Guide

The NPAR1WAY Procedure

(Chapter)

SAS® Documentation

This document is an individual chapter from *SAS/STAT® 9.3 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2011. *SAS/STAT® 9.3 User's Guide*. Cary, NC: SAS Institute Inc.

Copyright © 2011, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

For a Web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government Restricted Rights Notice: Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987).

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

1st electronic book, July 2011

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at **support.sas.com/publishing** or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Chapter 64

The NPAR1WAY Procedure

Contents

Overview: NPAR1WAY Procedure

The NPAR1WAY procedure performs nonparametric tests for location and scale differences across a one-way classification. PROC NPAR1WAY also provides a standard analysis of variance on the raw data and tests based on the empirical distribution function.

PROC NPAR1WAY performs tests for location and scale differences based on the following scores of a response variable: Wilcoxon, median, Van der Waerden (normal), Savage, Siegel-Tukey, Ansari-Bradley, Klotz, Mood, and Conover. Additionally, PROC NPAR1WAY provides tests that use the raw input data as scores. When the data are classified into two samples, tests are based on simple linear rank statistics. When the data are classified into more than two samples, tests are based on one-way ANOVA statistics. Both asymptotic and exact *p*-values are available for these tests. PROC NPAR1WAY also provides Hodges-Lehmann estimation, including exact confidence limits for the location shift.

PROC NPAR1WAY computes empirical distribution function (EDF) statistics, which test whether the distribution of a variable is the same across different groups. These include the Kolmogorov-Smirnov test, the Cramer-von Mises test, and, when the data are classified into only two samples, the Kuiper test. Exact *p*-values are available for the two-sample Kolmogorov-Smirnov test.

PROC NPAR1WAY uses the Output Delivery System (ODS), a SAS subsystem that provides capabilities for displaying and controlling the output from SAS procedures. ODS enables you to convert any of the output from PROC NPAR1WAY into a SAS data set. See the section "ODS Table Names" on page 5287 for more information.

PROC NPAR1WAY uses ODS Graphics to create graphs as part of its output. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." For specific information about the statistical graphics available with the NPAR1WAY procedure, see the PLOTS= option in the PROC NPAR1WAY statement and the section "ODS Graphics" on page 5289.

Getting Started: NPAR1WAY Procedure

This example illustrates how you can use PROC NPAR1WAY to perform a one-way nonparametric analysis. The data from Halverson and Sherwood (1930) consist of weight gain measurements for five different levels of gossypol additive in animal feed. Gossypol is a substance contained in cottonseed shells, and these data were collected to study the effect of gossypol on animal nutrition.

The following DATA step statements create the SAS data set Gossypol:

```
data Gossypol;
   input Dose n;
   do i=1 to n;
      input Gain @@;
      output;
   end;
   datalines;
0 16
228 229 218 216 224 208 235 229 233 219 224 220 232 200 208 232
.04 11
186 229 220 208 228 198 222 273 216 198 213
.07 12
179 193 183 180 143 204 114 188 178 134 208 196
.10 17
130 87 135 116 118 165 151 59 126 64 78 94 150 160 122 110 178
.13 11
154 130 130 118 118 104 112 134 98 100 104
;
```

The data set Gossypol contains the variable Dose, which represents the amount of gossypol additive, and the variable Gain, which represents the weight gain.

Researchers are interested in whether there is a difference in weight gain among animals receiving the different dose levels of gossypol. The following statements invoke the NPAR1WAY procedure to perform a nonparametric analysis of this problem:

```
proc npar1way data=Gossypol;
   class Dose;
   var Gain;
run;
```

The variable Dose is the CLASS variable, and the VAR statement specifies the variable Gain is the response variable. The CLASS statement is required, and you must name only one CLASS variable. You can name one or more analysis variables in the VAR statement. If you omit the VAR statement, PROC NPAR1WAY analyzes all numeric variables in the data set except for the CLASS variable, the FREQ variable, and the BY variables.

Since no analysis options are specified in the PROC NPAR1WAY statement, the ANOVA, WILCOXON, MEDIAN, VW, SAVAGE, and EDF options are invoked by default. The tables in the following figures show the results of these analyses.

The tables in Figure 64.1 are produced with the ANOVA option. For each level of the CLASS variable Dose, PROC NPAR1WAY displays the number of observations and the mean of the analysis variable Gain. PROC NPAR1WAY displays a standard analysis of variance on the raw data. This gives the same results as the GLM and ANOVA procedures. The *p*-value for the *F* test is <0.0001, which indicates that Dose accounts for a significant portion of the variability of the dependent variable Gain.

Figure 64.1 Analysis of Variance

The NPAR1WAY Procedure

Analysis of Variance for Variable Gain
Classified by Variable Dose

Dose	N	Mean
0	16	222.187500
0.04	11	217.363636
0.07	12	175.000000
0.1	17	120.176471
0.13	11	118.363636

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	4	140082.986077	35020.74652	55.8143	<.0001
Within	62	38901.998997	627.45160		

Average scores were used for ties.

The WILCOXON option produces the output in Figure 64.2. PROC NPAR1WAY first provides a summary of the Wilcoxon scores for the analysis variable Gain by class level. For each level of the CLASS variable Dose, PROC NPAR1WAY displays the following information: number of observations, sum of the Wilcoxon scores, expected sum under the null hypothesis of no difference among class levels, standard deviation under the null hypothesis, and mean score.

Next PROC NPAR1WAY displays the one-way ANOVA statistic, which for Wilcoxon scores is known as the Kruskal-Wallis test. The statistic equals 52.6656, with four degrees of freedom, which is the number of class levels minus one. The *p*-value (probability of a larger statistic under the null hypothesis) is <0.0001. This leads to rejection of the null hypothesis that there is no difference in location for Gain among the levels of Dose. This *p*-value is asymptotic, computed from the asymptotic chi-square distribution of the test statistic. For certain data sets it might also be useful to compute the exact *p*-value—for example, for small data sets or for data sets that are sparse, skewed, or heavily tied. You can use the EXACT statement to request exact *p*-values for any of the location or scale tests available in PROC NPAR1WAY.

Figure 64.2 Wilcoxon Score Analysis

Wilcoxon Scores (Rank Sums) for Variable Gain
Classified by Variable Dose

Dose	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	16	890.50	544.0	67.978966	55.656250
0.04	11	555.00	374.0	59.063588	50.454545
0.07	12	395.50	408.0	61.136622	32.958333
0.1	17	275.50	578.0	69.380741	16.205882
0.13	11	161.50	374.0	59.063588	14.681818

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	52.6656
DF	4
Pr > Chi-Square	<.0001

Figure 64.3 through Figure 64.5 display the analyses produced by the MEDIAN, VW, and SAVAGE options. For each score type, PROC NPAR1WAY provides a summary of scores and the one-way ANOVA statistic, as previously described for Wilcoxon scores. Other score types available in PROC NPAR1WAY are Siegel-Tukey, Ansari-Bradley, Klotz, and Mood, which can be used to test for scale differences. Conover scores can be used to test for differences in both location and scale. Additionally, you can specify the SCORES=DATA option, which uses the input data as scores. This option gives you the flexibility to construct any scores for your data with the DATA step and then analyze these scores with PROC NPAR1WAY.

Figure 64.3 Median Score Analysis

Median Scores (Number of Points Above Median) for Variable Gain
Classified by Variable Dose

Dose	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	16	16.0	7.880597	1.757902	1.00
0.04	11	11.0	5.417910	1.527355	1.00
0.07	12	6.0	5.910448	1.580963	0.50
0.1	17	0.0	8.373134	1.794152	0.00
0.13	11	0.0	5.417910	1.527355	0.00

Average scores were used for ties.

Median One-Way Analysis

Chi-Square	54.1765
DF	4
Pr > Chi-Square	<.0001

Figure 64.4 Van der Waerden (Normal) Score Analysis

Van der Waerden Scores (Normal) for Variable Gain
Classified by Variable Dose

Dose	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	16	16.116474	0.0	3.325957	1.007280
0.04	11	8.340899	0.0	2.889761	0.758264
0.07	12	-0.576674	0.0	2.991186	-0.048056
0.1	17	-14.688921	0.0	3.394540	-0.864054
0.13	11	-9.191777	0.0	2.889761	-0.835616

Average scores were used for ties.

Van der Waerden One-Way Analysis

Chi-Square	47.2972
DF	4
Pr > Chi-Square	<.0001

Figure 64.5 Savage Score Analysis

Savage Scores (Exponential) for Variable Gain
Classified by Variable Dose

Dose	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	16	16.074391	0.0	3.385275	1.004649
0.04	11	7.693099	0.0	2.941300	0.699373
0.07	12	-3.584958	0.0	3.044534	-0.298746
0.1	17	-11.979488	0.0	3.455082	-0.704676
0.13	11	-8.203044	0.0	2.941300	-0.745731

Average scores were used for ties.

Savage One-Way Analysis

Chi-Square	39.4908
DF	4
Pr > Chi-Square	<.0001

The tables in Figure 64.6 display the empirical distribution function statistics, comparing the distribution of Gain for the different levels of Dose. These tables are produced by the EDF option, and they include Kolmogorov-Smirnov statistics and Cramer-von Mises statistics.

Figure 64.6 Empirical Distribution Function Analysis

Kolmogorov-Smirnov Test for Variable Gain
Classified by Variable Dose

Dose	N	EDF at Maximum	Deviation from Mean at Maximum
0	16	0.000000	-1.910448
0.04	11	0.000000	-1.584060
0.07	12	0.333333	-0.499796
0.1	17	1.000000	2.153861
0.13	11	1.000000	1.732565
Total	67	0.477612	

Maximum Deviation Occurred at Observation 36
Value of Gain at Maximum = 178.0

Kolmogorov-Smirnov Statistics (Asymptotic)

KS	0.457928	KSa	3.748300

Cramer-von Mises Test for Variable Gain
Classified by Variable Dose

Dose	N	Summed Deviation from Mean
0	16	2.165210
0.04	11	0.918280
0.07	12	0.348227
0.1	17	1.497542
0.13	11	1.335745

Cramer-von Mises Statistics (Asymptotic)

CM	0.093508	CMa	6.265003

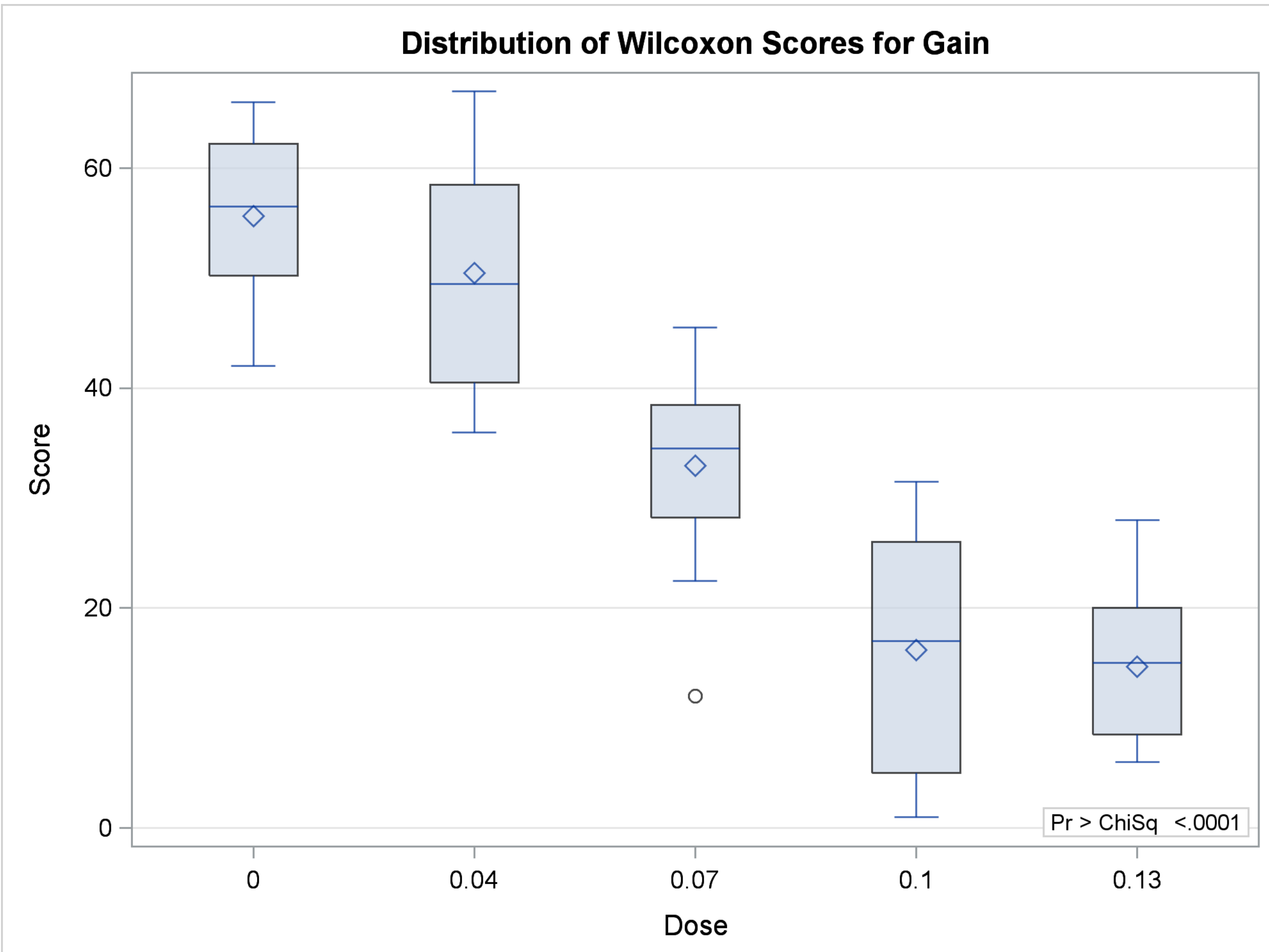
PROC NPAR1WAY uses ODS Graphics to create graphs as part of its output. The following statements produce a box plot of Wilcoxon scores for Gain classified by Dose. ODS Graphics must be enabled before producing graphs.

```
ods graphics on;
proc npar1way data=Gossypol plots(only)=wilcoxonboxplot;
   class Dose;
   var Gain;
run;
ods graphics off;
```

Figure 64.7 displays the box plot of Wilcoxon scores. This graph corresponds to the Wilcoxon scores analysis shown in Figure 64.2. To remove the *p*-value from the box plot display, you can specify the NOSTATS plot option in parentheses following the WILCOXONBOXPLOT option.

Box plots are available for all PROC NPAR1WAY score types except median scores, which are displayed with a stacked bar chart. If ODS Graphics is enabled but you do not specify the PLOTS= option, then PROC NPAR1WAY produces all plots that are associated with the analyses that you request.

Figure 64.7 Box Plot of Wilcoxon Scores



In the preceding example, the CLASS variable Dose has five levels, and the analyses examine possible differences among these five levels (samples). The following statements invoke the NPAR1WAY procedure to perform a nonparametric analysis of the two lowest levels of Dose:

```
proc npar1way data=Gossypol;
   where Dose <= .04;
   class Dose;
   var Gain;
run;
```

The tables in the following figures show the results of this two-sample analysis. The tables in Figure 64.8 are produced by the ANOVA option.

Figure 64.8 Analysis of Variance for Two-Sample Data

```
                         The NPAR1WAY Procedure

                  Analysis of Variance for Variable Gain
                       Classified by Variable Dose

                  Dose              N                Mean
                  --------------------------------------
                     0             16          222.187500
                  0.04             11          217.363636

   Source     DF     Sum of Squares     Mean Square      F Value     Pr > F
   -------------------------------------------------------------------------
   Among       1         151.683712      151.683712       0.5587     0.4617
   Within     25        6786.982955      271.479318

                   Average scores were used for ties.
```

Figure 64.9 displays the output produced by the WILCOXON option. PROC NPAR1WAY provides a summary of the Wilcoxon scores for the analysis variable Gain for each of the two class levels. Since there are only two levels, PROC NPAR1WAY displays the two-sample test, based on the simple linear rank statistic with Wilcoxon scores. The normal approximation includes a continuity correction. To remove the continuity correction, you can specify the CORRECT=NO option. PROC NPAR1WAY also gives a *t* approximation for the Wilcoxon two-sample test. Like the multisample analysis, PROC NPAR1WAY computes a one-way ANOVA statistic, which for Wilcoxon scores is known as the Kruskal-Wallis test. All these *p*-values show no difference in Gain for the two Dose levels at the 0.05 level of significance.

Figure 64.10 through Figure 64.12 display the two-sample analyses produced by the MEDIAN, VW, and SAVAGE options.

Figure 64.9 Wilcoxon Two-Sample Analysis

Wilcoxon Scores (Rank Sums) for Variable Gain
Classified by Variable Dose

Dose	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	16	253.50	224.0	20.221565	15.843750
0.04	11	124.50	154.0	20.221565	11.318182

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic	124.5000
Normal Approximation	
Z	-1.4341
One-Sided Pr < Z	0.0758
Two-Sided Pr > |Z|	0.1515
t Approximation	
One-Sided Pr < Z	0.0817
Two-Sided Pr > |Z|	0.1635

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test

Chi-Square	2.1282
DF	1
Pr > Chi-Square	0.1446

Figure 64.10 Median Two-Sample Analysis

Median Scores (Number of Points Above Median) for Variable Gain
Classified by Variable Dose

Dose	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	16	9.0	7.703704	1.299995	0.562500
0.04	11	4.0	5.296296	1.299995	0.363636

Average scores were used for ties.

Median Two-Sample Test

Statistic	4.0000
Z	-0.9972
One-Sided Pr < Z	0.1593
Two-Sided Pr > |Z|	0.3187

Median One-Way Analysis

Chi-Square	0.9943
DF	1
Pr > Chi-Square	0.3187

Figure 64.11 Van der Waerden (Normal) Two-Sample Analysis

Van der Waerden Scores (Normal) for Variable Gain
Classified by Variable Dose

Dose	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	16	3.346520	0.0	2.320336	0.209157
0.04	11	-3.346520	0.0	2.320336	-0.304229

Average scores were used for ties.

Van der Waerden Two-Sample Test

Statistic	-3.3465
Z	-1.4423
One-Sided Pr < Z	0.0746
Two-Sided Pr > |Z|	0.1492

Van der Waerden One-Way Analysis

Chi-Square	2.0801
DF	1
Pr > Chi-Square	0.1492

Figure 64.12 Savage Two-Sample Analysis

Savage Scores (Exponential) for Variable Gain
Classified by Variable Dose

Dose	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	16	1.834554	0.0	2.401839	0.114660
0.04	11	-1.834554	0.0	2.401839	-0.166778

Average scores were used for ties.

Savage Two-Sample Test

Statistic	-1.8346
Z	-0.7638
One-Sided Pr < Z	0.2225
Two-Sided Pr > |Z|	0.4450

Savage One-Way Analysis

Chi-Square	0.5834
DF	1
Pr > Chi-Square	0.4450

The tables in Figure 64.13 display the empirical distribution function statistics, comparing the distribution of Gain for the two levels of Dose. The *p*-value for the Kolmogorov-Smirnov two-sample test is 0.6199, which indicates no rejection of the null hypothesis that the Gain distributions are identical for the two levels of Dose.

Figure 64.13 Two-Sample EDF Tests

```
         Kolmogorov-Smirnov Test for Variable Gain
                Classified by Variable Dose

                          EDF at     Deviation from Mean
     Dose          N     Maximum          at Maximum
     --------------------------------------------------
        0         16     0.250000          -0.481481
     0.04         11     0.545455           0.580689
     Total        27     0.370370

        Maximum Deviation Occurred at Observation 4
              Value of Gain at Maximum = 216.0

      Kolmogorov-Smirnov Two-Sample Test (Asymptotic)
            KS    0.145172    D          0.295455
            KSa   0.754337    Pr > KSa   0.6199

          Cramer-von Mises Test for Variable Gain
                Classified by Variable Dose

                                     Summed Deviation
          Dose             N             from Mean
          ----------------------------------------
             0            16             0.098638
          0.04            11             0.143474

          Cramer-von Mises Statistics (Asymptotic)
               CM   0.008967    CMa   0.242112

               Kuiper Test for Variable Gain
                Classified by Variable Dose

                                       Deviation
               Dose          N         from Mean
               ------------------------------
                  0         16         0.090909
               0.04         11         0.295455

            Kuiper Two-Sample Test (Asymptotic)
       K   0.386364    Ka   0.986440    Pr > Ka   0.8383
```

Syntax: NPAR1WAY Procedure

The following statements are available in PROC NPAR1WAY:

PROC NPAR1WAY < *options* > **;**
  BY *variables* **;**
  CLASS *variable* **;**
  EXACT *statistic-options* < / *computation-options* > **;**
  FREQ *variable* **;**
  OUTPUT < **OUT=***SAS-data-set* > < *options* > **;**
  VAR *variables* **;**

Both the PROC NPAR1WAY statement and the CLASS statement are required for the NPAR1WAY procedure.

The rest of this section gives detailed syntax information for the BY, CLASS, EXACT, FREQ, OUTPUT, and VAR statements in alphabetical order after the description of the PROC NPAR1WAY statement. Table 64.1 summarizes the basic function of each PROC NPAR1WAY statement.

Table 64.1 Summary of PROC NPAR1WAY Statements

Statement	Description
BY	Provides separate analyses for each BY group
CLASS	Identifies the classification variable
EXACT	Requests exact tests
FREQ	Identifies a frequency variable
OUTPUT	Requests an output data set
VAR	Identifies analysis variables

PROC NPAR1WAY Statement

PROC NPAR1WAY < *options* > **;**

The PROC NPAR1WAY statement invokes the procedure and optionally identifies the input data set or requests particular analyses. By default, the procedure uses the most recently created SAS data set and omits missing values from the analysis. If you do not specify any analysis options, PROC NPAR1WAY performs an analysis of variance (ANOVA option), tests for location differences (WILCOXON, MEDIAN, SAVAGE, and VW options), and performs empirical distribution function tests (EDF option).

Table 64.2 lists the *options* available in the PROC NPAR1WAY statement. Descriptions of the *options* follow in alphabetical order.

Table 64.2 PROC NPAR1WAY Statement Options

Task	Options
Specify the input data set	DATA=
Include missing CLASS values	MISSING
Suppress all displayed output	NOPRINT
Request analyses	AB
	ANOVA
	CONOVER
	D
	EDF
	HL
	KLOTZ
	MEDIAN
	MOOD
	SAVAGE
	SCORES=DATA
	ST
	VW | NORMAL
	WILCOXON
Set confidence level	ALPHA=
Suppress continuity correction	CORRECT=NO
Request plots	PLOTS=

You can specify the following *options* in the PROC NPAR1WAY statement.

AB

requests an analysis of Ansari-Bradley scores. See the section "Ansari-Bradley Scores" on page 5267 for more information.

ALPHA=α

specifies the level of the confidence limits for location shift, which you request with the HL option. The value of α must be between 0 and 1, and the default is 0.05. A confidence level of α produces $100(1-\alpha)\%$ confidence limits. The default of ALPHA=0.05 produces 95% confidence limits for the Hodges-Lehmann estimate.

ANOVA

requests a standard analysis of variance on the raw data.

CONOVER

requests an analysis of Conover scores. See the section "Conover Scores" on page 5268 for more information.

CORRECT=NO

suppresses the continuity correction for the Wilcoxon two-sample test and the Siegel-Tukey two-sample test. See the section "Continuity Correction" on page 5265 for more information.

D

requests the one-sided Kolmogorov-Smirnov $D+$ and $D-$ statistics and their asymptotic p-values, in addition to the two-sided D statistic produced by the EDF option for two-sample data. The D option invokes the EDF option. The statistics $D+$ and $D-$ are provided by default if you request exact Kolmogorov-Smirnov statistics with the KS option in the EXACT statement for two-sample data. See the section "Tests Based on the Empirical Distribution Function" on page 5270 for details about Kolmogorov-Smirnov statistics.

DATA=*SAS-data-set*

names the SAS data set to be analyzed by PROC NPAR1WAY. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

EDF

requests statistics based on the empirical distribution function. These include the Kolmogorov-Smirnov and Cramer-von Mises tests and, if there are only two classification levels, the Kuiper test. See the section "Tests Based on the Empirical Distribution Function" on page 5270 for more information.

The EDF option produces the Kolmogorov-Smirnov D statistic for two-sample data. You can request the one-sided $D+$ and $D-$ statistics for two-sample data with the D option.

HL

requests Hodges-Lehmann estimation of the location shift for two-sample data. The HL option provides asymptotic confidence limits for the location shift. These are sometimes known as Moses confidence limits. See the section "Hodges-Lehmann Estimation of Location Shift" on page 5268 for details. You can specify the level of the confidence limits by using the ALPHA= option. The default of ALPHA=0.5 produces 95% confidence limits for the location shift.

KLOTZ

requests an analysis of Klotz scores. See the section "Klotz Scores" on page 5268 for more information.

MEDIAN

requests an analysis of median scores. When there are two classification levels, this option produces the two-sample median test. When there are more than two samples, this option produces the multi-sample median test, which is also known as the Brown-Mood test. See the section "Median Scores" on page 5266 for more information.

MISSING

treats missing values of the CLASS variable as a valid class level.

MOOD

requests an analysis of Mood scores. See the section "Mood Scores" on page 5268 for more information.

NOPRINT

suppresses the display of all output. You can use the NOPRINT option when you only want to create an output data set. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 20, "Using the Output Delivery System."

PLOTS < **(** *global-plot-options* **)** > < **=** *plot-request* < **(***plot-option***)** > >

PLOTS < **(** *global-plot-options* **)** > < **= (** *plot-request* < **(***plot-option***)** > <**. . .** *plot-request* < **(***plot-option***)** > > **)** >

controls the plots that are produced through ODS Graphics. Available plots include box plots, median plots, and empirical distribution plots. See Figure 64.7, Output 64.1.2, Output 64.1.4, and Output 64.2.2 for examples of plots that PROC NPAR1WAY produces. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS."

Plot-requests specify the plots to produce, *plot-options* apply to individual plots, and *global-plot-options* apply to all plots. When you specify only one *plot-request*, you can omit the parentheses around the request. For example:

```
plots=all
plots=wilcoxonboxplot
plots=(wilcoxonboxplot edfplot)
plots(only)=(medianplot normalboxplot)
```

ODS Graphics must be enabled before requesting plots. For example:

```
ods graphics on;
proc npar1way plots=wilcoxonboxplot;
   variable response;
   class treatment;
run;
ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 609 in Chapter 21, "Statistical Graphics Using ODS."

If ODS Graphics is enabled but you do not specify the PLOTS= option, PROC NPAR1WAY produces all plots that are associated with the analyses that you request. If you request a plot with the PLOTS= option but do not request the corresponding analysis, PROC NPAR1WAY automatically invokes that analysis. For example, if you specify PLOTS=CONOVERBOXPLOT but do not also specify the CONOVER option in the PROC NPAR1WAY statement, PROC NPAR1WAY produces the Conover scores analysis in addition to the box plot.

You can suppress default plots and request specific plots by using the PLOTS(ONLY)= option; PLOTS(ONLY)=(*plot-requests*) produces only the plots that are specified as *plot-requests*. You can suppress all plots with the PLOTS=NONE option. The PLOTS= option has no effect when you specify the NOPRINT option.

Global Plot Options

Global-plot-options apply to all plots produced by PROC NPAR1WAY unless they are altered by specific *plot-options*. You can specify the following *global-plot-options* in parentheses after the PLOTS option. You cannot specify both STATS and NOSTATS as *global-plot-options* in the same statement.

NOSTATS

suppresses the *p*-values that are displayed on the plots by default.

ONLY

suppresses the default plots and requests only the plots that are specified as *plot-requests*.

STATS
displays *p*-values on the plots. This is the default.

Plot Requests

The following *plot-requests* are available with the PLOTS= option.

ABBOXPLOT | AB
requests a box plot of Ansari-Bradley scores. This plot is associated with the Ansari-Bradley analysis, which you request with the AB option.

ALL
requests all plots that are associated with the specified analyses. This is the default if you do not specify the ONLY *global-plot-option*.

ANOVABOXPLOT | ANOVA
requests a box plot of the raw data. This plot is associated with the analysis of variance based on the raw data, which you request with the ANOVA option.

CONOVERBOXPLOT | CONOVER
requests a box plot of Conover scores. This plot is associated with the Conover analysis, which you request with the CONOVER option.

DATASCORESBOXPLOT | DATASCORES
requests a box plot of raw data scores. This plot is associated with the analysis that uses input data as scores, which you request with the SCORES=DATA option.

EDFPLOT | EDF
requests an empirical distribution plot. This plot is associated with the analyses based on the empirical distribution function, which you request with the EDF option.

KLOTZBOXPLOT | KLOTZ
requests a box plot of Klotz scores. This plot is associated with the Klotz analysis, which you request with the KLOTZ option.

MEDIANPLOT | MEDIAN
requests a stacked bar chart showing the frequencies above and below the overall median. This plot is associated with the median score analysis, which you request with the MEDIAN option.

MOODBOXPLOT | MOOD
requests a box plot of Mood scores. This plot is associated with the Mood analysis, which you request with the MOOD option.

NONE
suppresses all plots.

SAVAGEBOXPLOT | SAVAGE
requests a box plot of Savage scores. This plot is associated with the Savage analysis, which you request with the SAVAGE option.

STBOXPLOT | ST
requests a box plot of Siegel-Tukey scores. This plot is associated with the Siegel-Tukey analysis, which you request with the ST option.

VWBOXPLOT | VW

NORMALBOXPLOT | NORMAL

requests a box plot of Van der Waerden (normal) scores. This plot is associated with the Van der Waerden analysis, which you request with the VW or NORMAL option.

WILCOXONBOXPLOT | WILCOXON

requests a box plot of Wilcoxon scores. This plot is associated with the Wilcoxon analysis, which you request with the WILCOXON option.

Plot Options

The following *plot-options* are available for any *plot-request*. You cannot specify both STATS and NOSTATS as *plot-options* for the same plot. If you specify NOSTATS as a *global-plot-option*, specifying STATS as an individual *plot-option* overrides the *global-plot-option* for the individual plot and displays statistics on the plot.

NOSTATS

suppresses the p-values that are displayed on the plot by default.

STATS

displays p-values on the plot. This is the default.

SAVAGE

requests an analysis of Savage scores. See the section "Savage Scores" on page 5267 for more information.

SCORES=DATA

requests an analysis that uses input data as scores. This option gives you the flexibility to construct any scores for your data with the DATA step and then analyze these scores with PROC NPAR1WAY. See the section "Scores for Linear Rank and One-Way ANOVA Tests" on page 5266 for more information.

To produce the two-sample permutation test that is known as Pitman's test, provide raw (unscored) data in the input data set and specify the SCORES=DATA option in the EXACT statement. See the section "Exact Tests" on page 5273 for more information.

ST

requests an analysis of Siegel-Tukey scores. See the section "Siegel-Tukey Scores" on page 5267 for more information.

VW | NORMAL

requests an analysis of Van der Waerden (normal) scores. See the section "Van der Waerden (Normal) Scores" on page 5267 for more information.

WILCOXON

requests an analysis of Wilcoxon scores. When there are two classification levels (samples), this option produces the Wilcoxon rank-sum test. For any number of classification levels, this option produces the Kruskal-Wallis test. See the section "Wilcoxon Scores" on page 5266 for more information.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC NPAR1WAY to obtain separate analyses of observations in groups that are defined by the BY variables. If you specify more than one BY statement, the procedure uses only the last BY statement and ignores any previous BY statements.

When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the NPAR1WAY procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

CLASS Statement

CLASS *variable* **;**

The CLASS statement, which is required, names one and only one classification variable. The variable can be character or numeric. The CLASS variable identifies groups (samples) in the data, and PROC NPAR1WAY provides analyses to examine differences among these groups. There can be two or more groups in the data.

EXACT Statement

EXACT *statistic-options* < / *computation-options* > **;**

The EXACT statement requests exact tests for the specified statistics. Optionally, PROC NPAR1WAY computes Monte Carlo estimates of the exact *p*-values. The *statistic-options* specify the exact tests to compute. The *computation-options* specify options for the computation of exact statistics. See the section "Exact Tests" on page 5273 for details.

NOTE: PROC NPAR1WAY computes exact tests with fast and efficient algorithms that are superior to direct enumeration. Exact tests are appropriate when a data set is small, sparse, skewed, or heavily tied.

For some large problems, computation of exact tests might require a large amount of time and memory. Consider using asymptotic tests for such problems. Alternatively, when asymptotic methods might not be sufficient for such large problems, consider using Monte Carlo estimation of exact *p*-values. See the section "Computational Resources" on page 5274 for more information.

Statistic Options

Statistic-options specify the exact tests to compute.

Exact *p*-values are available for all nonparametric tests of location and scale differences that are produced by PROC NPAR1WAY. These include tests based on the following scores: Wilcoxon, median, Van der Waerden (normal), Savage, Siegel-Tukey, Ansari-Bradley, Klotz, Mood, and Conover. Additionally, exact *p*-values are available for tests that use the raw input data as scores. The procedure computes exact *p*-values when the data are classified into two levels (two-sample tests) and when the data are classified into more than two levels (multisample tests). Two-sample tests are based on simple linear rank statistics. Multisample tests are based on one-way ANOVA statistics. See the section "Exact Tests" on page 5273 for details.

Exact *p*-values are also available for the two-sample Kolmogorov-Smirnov test. Additionally, exact confidence limits are available for the Hodges-Lehmann estimate of location shift. See the section "Hodges-Lehmann Estimation of Location Shift" on page 5268 for details.

Table 64.3 lists the available *statistic-options* and the exact tests computed. The option names are identical to the corresponding options in the PROC NPAR1WAY statement and the OUTPUT statement.

If you list no *statistic-options* in the EXACT statement, then PROC NPAR1WAY computes all available exact *p*-values for those tests that you request in the PROC NPAR1WAY statement.

Table 64.3 EXACT Statement Statistic Options

Statistic Option	Exact Test
AB	Ansari-Bradley test
CONOVER	Conover test
HL	Hodges-Lehmann confidence limits
KLOTZ	Klotz test
KS | EDF	Two-sample Kolmogorov-Smirnov test
MEDIAN	Median test
MOOD	Mood test
SAVAGE	Savage test
SCORES=DATA	Test with input data as scores
ST	Siegel-Tukey test
VW | NORMAL	Van der Waerden (normal scores) test
WILCOXON	Wilcoxon test for two-sample data or Kruskal-Wallis test for multisample data

Computation Options

Computation-options specify options for computation of exact statistics. You can specify the following *computation-options* in the EXACT statement after a slash (/).

ALPHA=α

specifies the level of the confidence limits for Monte Carlo p-value estimates. The value of α must be between 0 and 1, and the default is 0.01. A confidence level of α produces $100(1-\alpha)\%$ confidence limits. The default of ALPHA=0.01 produces 99% confidence limits for the Monte Carlo estimates.

The ALPHA= option invokes the MC option.

MAXTIME=*value*

specifies the maximum clock time (in seconds) that PROC NPAR1WAY can use to compute an exact p-value. If the procedure does not complete the computation within the specified time, the computation terminates. The value of MAXTIME= must be a positive number. The MAXTIME= option is valid for both Monte Carlo estimation of exact p-values and direct exact p-value computation. See the section "Computational Resources" on page 5274 for more information.

MC

requests Monte Carlo estimation of exact p-values, instead of direct exact p-value computation. Monte Carlo estimation can be useful for large problems that require a great amount of time and memory for exact computations but for which asymptotic approximations might not be sufficient. See the section "Monte Carlo Estimation" on page 5275 for more information.

The MC option is available for all EXACT statement *statistic-options* except the HL option, which produces exact Hodges-Lehmann confidence limits. The ALPHA=, N=, and SEED= options also invoke the MC option.

N=*n*

specifies the number of samples for Monte Carlo estimation. The value of *n* must be a positive integer, and the default is 10,000 samples. Larger values of *n* produce more precise estimates of exact p-values. Because larger values of *n* generate more samples, the computation time increases.

The N= option invokes the MC option.

POINT

requests exact point probabilities for the test statistics.

The POINT option is available for all EXACT statement *statistic-options* except the HL option, which produces exact Hodges-Lehmann confidence limits. The POINT option is not available with the MC option.

SEED=*number*

specifies the initial seed for random number generation for Monte Carlo estimation. The value of the SEED= option must be an integer. If you do not specify the SEED= option or if the SEED= value is negative or zero, PROC NPAR1WAY uses the time of day from the computer's clock to obtain the initial seed.

The SEED= option invokes the MC option.

FREQ Statement

FREQ *variable* **;**

The FREQ statement names a numeric variable that provides a frequency for each observation in the input data set. If you use a FREQ statement, PROC NPAR1WAY assumes that an observation occurs *n* times, where *n* is the value of the FREQ variable for the observation. The sum of the FREQ variable values represents the total number of observations, and the analysis is based on this expanded number of observations.

If the value of the FREQ variable is missing or is less than one, PROC NPAR1WAY does not use that observation in the analysis. If the value of the FREQ variable is not an integer, PROC NPAR1WAY uses only the integer portion as the frequency of the observation.

OUTPUT Statement

OUTPUT < **OUT=***SAS-data-set*> < *options*> **;**

The OUTPUT statement creates a SAS data set that contains statistics computed by PROC NPAR1WAY. You specify which statistics to store in the output data set by using options that are identical to those that are available in the PROC NPAR1WAY statement. The output data set contains one observation for each analysis variable named in the VAR statement. For more information about the contents of the output data set, see the section "Output Data Set" on page 5277.

Note that you can use the Output Delivery System (ODS) to create a SAS data set from any piece of PROC NPAR1WAY output. For more information, see the section "ODS Table Names" on page 5287 and Chapter 20, "Using the Output Delivery System."

You can specify the following *options* in the OUTPUT statement:

OUT=*SAS-data-set*

names the output data set. If you omit the OUT= option, the data set is named DATA*n*, where *n* is the smallest integer that makes the name unique.

options

specifies the statistics you want in the output data set. The *options* are identical to those that you can use in the PROC NPAR1WAY statement to request analyses. Table 64.4 shows the available *options*. When you specify one of these options in the OUTPUT statement, the output data set contains all statistics from that analysis. See the section "Output Data Set" on page 5277 for a list of the output data set variables corresponding to each option.

If you do not specify any statistics options in the OUTPUT statement, then the output data set includes statistics from all analyses that you request in the PROC NPAR1WAY statement.

Table 64.4 OUTPUT Statement Options

Option	Output Data Set Statistics
AB	Ansari-Bradley test
ANOVA	Analysis of variance
CONOVER	Conover test
EDF	Kolmogorov-Smirnov test, Cramer-von Mises test, and Kuiper test for two-sample data
HL	Hodges-Lehmann estimates
KLOTZ	Klotz test
MEDIAN	Median test
MOOD	Mood test
SAVAGE	Savage test
SCORES=DATA	Test with input data as scores
ST	Siegel-Tukey test
VW | NORMAL	Van der Waerden (normal scores) test
WILCOXON	Wilcoxon test for two-sample data and Kruskal-Wallis test

VAR Statement

VAR *variables* **;**

The VAR statement names the response (dependent) variables to be included in the analysis. These variables must be numeric. If you omit the VAR statement, the procedure includes all numeric variables in the data set except for the CLASS variable, the FREQ variable, and the BY variables.

Details: NPAR1WAY Procedure

Missing Values

If an observation has a missing value for a response (VAR) variable, PROC NPAR1WAY excludes that observation from the analysis. Any observation with a missing or nonpositive value for the FREQ variable is also excluded from the analysis.

By default, PROC NPAR1WAY also excludes observations with missing values of the CLASS variable. If you specify the MISSING option, PROC NPAR1WAY treats missing values of the CLASS variable as a valid class level and includes these observations in the analysis.

PROC NPAR1WAY treats missing BY variable values like any other BY variable value. The missing values form a separate, valid BY group.

Tied Values

Tied values occur when two or more observations are equal, whether the observations occur in the same sample or in different samples. In theory, nonparametric tests were developed for continuous distributions where the probability of a tie is zero. In practice, however, ties often occur. PROC NPAR1WAY uses the same method to handle ties for all score types. The procedure computes the scores as if there were no ties, averages the scores for tied observations, and assigns this average score to each observation with the same value.

When there are tied values, PROC NPAR1WAY first sorts the observations in ascending order and assigns ranks as if there were no ties. Then the procedure computes the scores based on these ranks by using the formula for the specified score type. The procedure averages the scores for tied observations and assigns this average score to each of the tied observations. Thus, all equal data values have the same score value. PROC NPAR1WAY then computes the test statistic from these scores.

Note that the asymptotic tests might be less accurate when the distribution of the data is heavily tied. For such data, it might be appropriate to use the exact tests provided by PROC NPAR1WAY as described in the section “Exact Tests” on page 5273.

When computing empirical distribution function statistics for data with ties, PROC NPAR1WAY uses the formulas given in the section “Tests Based on the Empirical Distribution Function” on page 5270. No special handling of ties is necessary.

Note that PROC NPAR1WAY bases its computations on the internal numeric values of the analysis variables; the procedure does not format or round these values before analysis. When values differ in their internal representation, even slightly, PROC NPAR1WAY does not treat them as tied values. If this is a concern for your data, then round the analysis variables by an appropriate amount before invoking PROC NPAR1WAY. For information about the ROUND function, see the discussion in *SAS Language Reference: Dictionary*.

Statistical Computations

Simple Linear Rank Tests for Two-Sample Data

Statistics of the form

$$S = \sum_{j=1}^{n} c_j \, a(R_j)$$

are called *simple linear rank statistics*, where

R_j	is the rank of observation j
$a(R_j)$	is the score based on the rank of observation j
c_j	is an indicator variable denoting the class to which the jth observation belongs
n	is the total number of observations

For two-sample data (where the observations are classified into two levels), PROC NPAR1WAY calculates simple linear rank statistics for the scores that you specify. The section "Scores for Linear Rank and One-Way ANOVA Tests" on page 5266 describes the available scores, which you can use to test for differences in location and differences in scale.

To compute the linear rank statistic S, PROC NPAR1WAY sums the scores of the observations in the smaller of the two samples. If both samples have the same number of observations, PROC NPAR1WAY sums those scores for the sample that appears first in the input data set.

For each score that you specify, PROC NPAR1WAY computes an asymptotic test of the null hypothesis of no difference between the two classification levels. Exact tests are also available for these two-sample linear rank statistics. PROC NPAR1WAY computes exact tests for each score type that you specify in the EXACT statement. See the section "Exact Tests" on page 5273 for details.

To compute an asymptotic test for a linear rank sum statistic, PROC NPAR1WAY uses a standardized test statistic z, which has an asymptotic standard normal distribution under the null hypothesis. The standardized test statistic is computed as

$$z = (S - \mathrm{E}_0(S)) \,/\, \sqrt{\mathrm{Var}_0(S)}$$

where $\mathrm{E}_0(S)$ is the expected value of S under the null hypothesis, and $\mathrm{Var}_0(S)$ is the variance under the null hypothesis. As shown in Randles and Wolfe (1979),

$$\mathrm{E}_0(S) = \frac{n_1}{n} \sum_{j=1}^{n} a(R_j)$$

where n_1 is the number of observations in the first (smaller) class level (sample), n_2 is the number of observations in the other class level, and

$$\mathrm{Var}_0(S) = \frac{n_1 n_2}{n(n-1)} \sum_{j=1}^{n} \left(a(R_j) - \bar{a} \right)^2$$

where $\bar{a}$ is the average score,

$$\bar{a} = \frac{1}{n} \sum_{j=1}^{n} a(R_j)$$

Definition of p*-Values*

PROC NPAR1WAY computes one-sided and two-sided asymptotic p-values for each two-sample linear rank test. When the test statistic z is greater than its null hypothesis expected value of zero, PROC NPAR1WAY computes the right-sided p-value, which is the probability of a larger value of the statistic occurring under the null hypothesis. When the test statistic is less than or equal to zero, PROC NPAR1WAY computes the left-sided p-value, which is the probability of a smaller value of the statistic occurring under the null hypothesis. The one-sided p-value $P_1(z)$ can be expressed as

$$P_1(z) = \begin{cases} \mathrm{Prob}(Z > z) & \text{if } z > 0 \\ \mathrm{Prob}(Z < z) & \text{if } z \le 0 \end{cases}$$

where Z has a standard normal distribution. The two-sided p-value $P_2(z)$ is computed as

$$P_2(z) = \mathrm{Prob}(|Z| > |z|)$$

Continuity Correction

PROC NPAR1WAY uses a continuity correction for the asymptotic two-sample Wilcoxon and Siegel-Tukey tests by default. You can remove the continuity correction by specifying the CORRECT=NO option. PROC NPAR1WAY incorporates the continuity correction when computing the standardized test statistic z by subtracting 0.5 from the numerator $(S - \mathrm{E}_0(S))$ if it is greater than zero. If the numerator is less than zero, PROC NPAR1WAY adds 0.5. Some sources recommend a continuity correction for nonparametric tests that use a continuous distribution to approximate a discrete distribution. (See Sheskin 1997.)

If you specify CORRECT=NO, PROC NPAR1WAY does not use a continuity correction for any test.

One-Way ANOVA Tests

PROC NPAR1WAY computes a one-way ANOVA test for each score type that you specify. Under the null hypothesis of no difference among class levels (samples), this test statistic has an asymptotic chi-square distribution with $r - 1$ degrees of freedom, where r is the number of class levels. For Wilcoxon scores, this test is known as the Kruskal-Wallis test.

Exact one-way ANOVA tests are also available for multisample data (where the data are classified into more than two levels). For two-sample data, exact simple linear rank tests are available. PROC NPAR1WAY computes exact tests for each score type that you specify in the EXACT statement. See the section "Exact Tests" on page 5273 for details.

PROC NPAR1WAY computes the one-way ANOVA test statistic as

$$C = \left(\sum_{i=1}^{r} (T_i - \mathrm{E}_0(T_i))^2 \; / \; n_i \right) / \; S^2$$

where T_i is the total of scores for class level i, $\mathrm{E}_0(T_i)$ is the expected total for level i under the null hypothesis of no difference among levels, n_i is the number of observations in level i, and S^2 is the sample variance of the scores. The total of scores for class level i is given by

$$T_i = \sum_{j=1}^{n} c_{ij} \; a(R_j)$$

where $a(R_j)$ is the score for observation j, and c_{ij} indicates whether observation j is in level i. The expected total of scores for class level i under the null hypothesis is equal to

$$\mathrm{E}_0(T_i) = \frac{n_i}{n} \sum_{j=1}^{n} a(R_j)$$

The sample variance of the scores is computed as

$$S^2 = \frac{1}{(n-1)} \sum_{j=1}^{n} \left(a(R_j) - \bar{a} \right)^2$$

where $\bar{a}$ is the average score,

$$\bar{a} = \frac{1}{n} \sum_{j=1}^{n} a(R_j)$$

Scores for Linear Rank and One-Way ANOVA Tests

For each score type that you specify, PROC NPAR1WAY computes a one-way ANOVA statistic and also a linear rank statistic for two-sample data. The following score types are used primarily to test for differences in location: Wilcoxon, median, Van der Waerden (normal), and Savage. The following scores types are used to test for scale differences: Siegel-Tukey, Ansari-Bradley, Klotz, and Mood. Conover scores can be used to test for differences in both location and scale. This section gives formulas for the score types available in PROC NPAR1WAY. For further information about the formulas and the applicability of each score, see Randles and Wolfe (1979), Gibbons and Chakraborti (1992), Conover (1999), and Hollander and Wolfe (1999).

In addition to the score types described in this section, you can specify the SCORES=DATA option to use the input data observations as scores. This enables you to produce a wide variety of tests. You can construct any scores by using the DATA step, and then you can use PROC NPAR1WAY to compute the corresponding linear rank and one-way ANOVA tests for these scores. You can also analyze raw (unscored) data with the SCORES=DATA option; for two-sample data, the corresponding exact test is a permutation test that is known as Pitman's test.

Wilcoxon Scores

Wilcoxon scores are the ranks of the observations, which can be written as

$$a(R_j) = R_j$$

where R_j is the rank of observation j, and $a(R_j)$ is the score of observation j.

Using Wilcoxon scores in the linear rank statistic for two-sample data produces the rank sum statistic of the Mann-Whitney-Wilcoxon test. Using Wilcoxon scores in the one-way ANOVA statistic produces the Kruskal-Wallis test. Wilcoxon scores are locally most powerful for location shifts of a logistic distribution.

When computing the asymptotic Wilcoxon two-sample test, PROC NPAR1WAY uses a continuity correction by default, as described in the section "Continuity Correction" on page 5265. If you specify the CORRECT=NO option in the PROC NPAR1WAY statement, the procedure does not use a continuity correction.

Median Scores

Median scores equal 1 for observations greater than the median, and 0 otherwise. In terms of the observation ranks, median scores are defined as

$$a(R_j) = \begin{cases} 1 & \text{if} \quad R_j > (n+1)/2 \\ 0 & \text{if} \quad R_j \le (n+1)/2 \end{cases}$$

Using median scores in the linear rank statistic for two-sample data produces the two-sample median test. The one-way ANOVA statistic with median scores is equivalent to the Brown-Mood test. Median scores are particularly powerful for distributions that are symmetric and heavy-tailed.

Van der Waerden (Normal) Scores

Van der Waerden scores are the quantiles of a standard normal distribution and are also known as *quantile normal scores*. Van der Waerden scores are computed as

$$a(R_j) = \Phi^{-1}\left(\frac{R_j}{n+1}\right)$$

where Φ is the cumulative distribution function of a standard normal distribution. These scores are powerful for normal distributions.

Savage Scores

Savage scores are expected values of order statistics from the exponential distribution, with 1 subtracted to center the scores around 0. Savage scores are computed as

$$a(R_j) = \sum_{i=1}^{R_j}\left(\frac{1}{n-i+1}\right) - 1$$

Savage scores are powerful for comparing scale differences in exponential distributions or location shifts in extreme value distributions (Hajek 1969, p. 83).

Siegel-Tukey Scores

Siegel-Tukey scores are defined as

$$\begin{array}{llll} a(1) = 1, & a(n) = 2, & a(n-1) = 3, & a(2) = 4, \\ a(3) = 5, & a(n-2) = 6, & a(n-3) = 7, & a(4) = 8, \ \ldots \end{array}$$

where the score values continue to increase in this pattern toward the middle ranks until all observations have been assigned a score.

When computing the asymptotic Siegel-Tukey two-sample test, PROC NPAR1WAY uses a continuity correction by default, as described in the section “Continuity Correction” on page 5265. If you specify the CORRECT=NO option in the PROC NPAR1WAY statement, the procedure does not use a continuity correction.

Ansari-Bradley Scores

Ansari-Bradley scores are similar to Siegel-Tukey scores, but Ansari-Bradley scoring assigns the same score value to corresponding extreme ranks. (Siegel-Tukey scores are a permutation of the ranks $1, 2, \ldots, n$.) Ansari-Bradley scores are defined as

$$\begin{array}{ll} a(1) = 1, & a(n) = 1, \\ a(2) = 2, & a(n-1) = 2, \ \ldots \end{array}$$

Equivalently, Ansari-Bradley scores are equal to

$$a(R_j) = \frac{n+1}{2} - \left| R_j - \frac{n+1}{2} \right|$$

Klotz Scores

Klotz scores are the squares of the Van der Waerden (normal) scores. Klotz scores are computed as

$$a(R_j) = \left(\Phi^{-1} \left(\frac{R_j}{n+1} \right) \right)^2$$

where Φ is the cumulative distribution function of a standard normal distribution.

Mood Scores

Mood scores are computed as the square of the difference between the observation rank and the average rank. Mood scores can be written as

$$a(R_j) = \left(R_j - \frac{n+1}{2} \right)^2$$

Conover Scores

Conover scores are based on the squared ranks of the absolute deviations from the sample means. For observation j the absolute deviation from the mean is computed as

$$U_j = |X_{j(i)} - \bar{X}_i|$$

where $X_{j(i)}$ is the value of observation j, observation j belongs to sample i, and $\bar{X}_i$ is the mean of sample i. The values of U_j are ranked, and the Conover score for observation j is computed as

$$\text{Score}_j = \left(\text{Rank}(U_j) \right)^2$$

Following Conover (1999), when there are ties among the values of U_j, PROC NPAR1WAY assigns the average rank to each of the tied observations. To compute the average rank, PROC NPAR1WAY ranks the U_j as if there were no ties, and then averages the ranks of the tied observations.

The Conover score test is also known as the squared ranks test for variances. See Conover (1999) for more information.

Hodges-Lehmann Estimation of Location Shift

If you specify the HL option, PROC NPAR1WAY computes the Hodges-Lehmann estimate of location shift for two-sample data. PROC NPAR1WAY also provides confidence limits for the location shift. These confidence limits are sometimes called Moses confidence limits. You can set the level of the confidence limits with the ALPHA= option. The default is ALPHA=0.05, which produces 95% confidence limits. Additionally, you can request exact confidence limits for the location shift by specifying the HL option in the EXACT statement.

The Hodges-Lehmann estimator of location shift is associated with the Wilcoxon linear rank statistic. See Hollander and Wolfe (1999) and Hodges and Lehmann (1983) for details.

PROC NPAR1WAY computes the Hodges-Lehmann estimate $\hat{\Delta}$ as the median of all paired differences between observations in the two samples, which can be written as

$$\hat{\Delta} = \text{median}\left((Y_j - X_i) \quad \text{where} \quad j = 1, 2, \ldots, n_1;\ i = 1, 2, \ldots, n_2 \right)$$

The Y_j are observations in sample 1, the X_i are observations in sample 2, and n_1 and n_2 denote the number of observations in sample 1 and sample 2, respectively. PROC NPAR1WAY uses the smaller of the two samples as sample 1. If both samples have the same number of observations, PROC NPAR1WAY uses the sample that appears first in the input data set as sample 1. Sample 1 is the same sample that PROC NPAR1WAY uses to compute the two-sample linear rank statistic.

Let m denote the total number of differences ($n_1 \times n_2$), and let $U^{(k)}$ denote the kth value of $(Y_j - X_i)$ among the ordered differences. When m is an odd number, then the median difference is the value with rank $(m + 1)/2$,

$$\hat{\Delta} = U^{(k)} \quad \text{where} \quad k = (m+1)/2$$

When m is an even number, the median difference is the average of the values with ranks $(m/2)$ and $((m/2) + 1)$,

$$\hat{\Delta} = \left(U^{(k)} + U^{(k+1)} \right) / 2 \quad \text{where} \quad k = m/2$$

Following Hollander and Wolfe (1999), the asymptotic lower and upper confidence limits for the location shift are

$$\left(\Delta_L = U^{(C_\alpha)}, \quad \Delta_U = U^{(m+1-C_\alpha)} \right)$$

where C_α is the largest integer less than or equal to $C_\alpha{}^*$, which is computed as

$$C_\alpha{}^* = \text{E}_0(S) - z_{\alpha/2} \sqrt{\text{Var}_0(S)}$$

where $\text{E}_0(S)$ and $\text{Var}_0(S)$ are the expected value and variance, respectively, of the Wilcoxon statistic S under the null hypothesis (as described in the section "Simple Linear Rank Tests for Two-Sample Data" on page 5263), and $z_{\alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the standard normal distribution. For Wilcoxon rank scores,

$$\text{E}_0(S) = n_1 n_2 / 2$$

When there are no tied values, $\text{Var}_0(S)$ for Wilcoxon scores equals

$$\text{Var}_0(S) = n_1 n_2 (n_1 + n_2 + 1) / 12$$

PROC NPAR1WAY displays the midpoint of the confidence interval (Δ_L, Δ_U), which can also be used as an estimate of location shift. See Lehmann (1963) for details. Additionally, PROC NPAR1WAY provides an estimate of the asymptotic standard error of $\hat{\Delta}$ based on the length of the confidence interval, which is computed as

$$\text{se}(\hat{\Delta}) = (\Delta_U - \Delta_L) \ / \ (2\ z_{\alpha/2})$$

Exact Confidence Limits

If you specify the HL option in the EXACT statement, PROC NPAR1WAY computes exact confidence limits for the location shift between the two samples. As for the asymptotic confidence limits, you can set the confidence level with the ALPHA= option. The default is ALPHA=0.05, which produces 95% confidence limits.

PROC NPAR1WAY computes exact confidence limits for the location shift as described in Randles and Wolfe (1979, p. 180). PROC NPAR1WAY first generates the exact conditional distribution of the Mann-Whitney U statistic, which equals the number of pairwise differences $(Y_i - X_j)$ that are positive, plus half the number of pairwise differences that are zero. The Mann-Whitney statistic is defined as

$$MW = \sum_{i=1}^{n_1} \sum_{j=i}^{n_2} \phi \left(X_i - Y_j \right)$$

where

$$\phi(Y_i, X_j) = \begin{cases} 1 & \text{if } \; Y_i > X_j \\ 1/2 & \text{if } \; Y_i = X_j \\ 0 & \text{otherwise} \end{cases}$$

From the exact conditional distribution of the Mann-Whitney statistic MW, PROC NPAR1WAY chooses $C^*_{L,\alpha}$ as the largest value such that $\text{Prob}(MW \geq C^*_{L,\alpha}) \geq \alpha/2$. Rounding $C^*_{L,\alpha}$ up to the nearest integer $C_{L,\alpha}$, the lower confidence limit equals the difference $(Y_i - X_j)$ that has a rank of $(n_1 n_2 - C_{L,\alpha} + 1)$.

To find the upper confidence limit, PROC NPAR1WAY chooses $C^*_{U,\alpha}$ as the smallest value such that $\text{Prob}(MW \leq C^*_{U,\alpha}) \geq \alpha/2$. Rounding $C^*_{U,\alpha}$ down to the nearest integer $C_{L,\alpha}$, the upper confidence limit equals the difference $(Y_i - X_j)$ that has a rank of $(n_1 n_2 - C_{U,\alpha})$.

Because this is a discrete problem, the confidence coefficient for these exact confidence limits is not exactly $(1 - \alpha)$ but is at least $(1 - \alpha)$. Thus, these confidence limits are conservative.

Tests Based on the Empirical Distribution Function

If you specify the EDF option, PROC NPAR1WAY computes tests based on the empirical distribution function. These include the Kolmogorov-Smirnov and Cramer-von Mises tests, and also the Kuiper test for two-sample data. This section gives formulas for these test statistics. For further information about the formulas and the interpretation of EDF statistics, see Hollander and Wolfe (1999) and Gibbons and Chakraborti (1992). For details about the k-sample analogs of the Kolmogorov-Smirnov and Cramer-von Mises statistics, see Kiefer (1959).

The *empirical distribution function* (EDF) of a sample $\{x_j\}$, $j = 1, 2, \ldots, n$, is defined as

$$F(x) = \frac{1}{n} (\text{number of } x_j \leq x) = \frac{1}{n} \sum_{j=1}^{n} I(x_j \leq x)$$

where $I(\cdot)$ is an indicator function. PROC NPAR1WAY uses the subsample of values within the ith class level to generate an EDF for the class, F_i. The EDF for the overall sample, pooled over classes, can also be expressed as

$$F(x) = \frac{1}{n} \sum_i (n_i F_i(x))$$

where n_i is the number of observations in the ith class level, and n is the total number of observations.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov statistic measures the maximum deviation of the EDF within the classes from the pooled EDF. PROC NPAR1WAY computes the Kolmogorov-Smirnov statistic as

$$KS = \max_j \sqrt{\frac{1}{n} \sum_i n_i \left(F_i(x_j) - F(x_j) \right)^2} \quad \text{where} \quad j = 1, 2, \ldots, n$$

The asymptotic Kolmogorov-Smirnov statistic is computed as

$$KS_a = KS \times \sqrt{n}$$

For each class level i and overall, PROC NPAR1WAY displays the value of F_i at the maximum deviation from F and the value $\sqrt{n_i}\ (F_i - F)$ at the maximum deviation from F. PROC NPAR1WAY also gives the observation where the maximum deviation occurs.

If there are only two class levels, PROC NPAR1WAY computes the two-sample Kolmogorov-Smirnov test statistic D as

$$D = \max_j \left| F_1(x_j) - F_2(x_j) \right| \quad \text{where} \quad j = 1, 2, \ldots, n$$

The p-value for this test is the probability that D is greater than the observed value d under the null hypothesis of no difference between class levels (samples). PROC NPAR1WAY computes the asymptotic p-value for D with the approximation

$$\text{Prob}(D > d) = 2 \sum_{i=1}^{\infty} (-1)^{(i-1)}\ e^{(-2i^2\ z^2)}$$

where

$$z = d \sqrt{n_1\ n_2\ /\ n}$$

The quality of this approximation has been studied by Hodges (1957).

If you specify the D option, or if you request exact Kolmogorov-Smirnov p-values with the KS option in the EXACT statement, PROC NPAR1WAY also computes the one-sided Kolmogorov-Smirnov statistics $D+$ and $D-$ for two-sample data as

$$D+ = \max_j \left(F_1(x_j) - F_2(x_j) \right) \quad \text{where} \quad j = 1, 2, \ldots, n$$

$$D- = \max_j \left(F_2(x_j) - F_1(x_j) \right) \quad \text{where} \quad j = 1, 2, \ldots, n$$

The asymptotic probability that $D+$ is greater than the observed value d^+, under the null hypothesis of no difference between the two class levels, is computed as

$$\text{Prob}(D+ > d^+) = e^{-2z^2} \quad \text{where} \quad z = d^+ \sqrt{n_1\, n_2 \,/\, n}$$

Similarly, the asymptotic probability that $D-$ is greater than the observed value d^- is computed as

$$\text{Prob}(D- > d^-) = e^{-2z^2} \quad \text{where} \quad z = d^- \sqrt{n_1\, n_2 \,/\, n}$$

To request exact p-values for the Kolmogorov-Smirnov statistics, you can specify the KS option in the EXACT statement. See the section "Exact Tests" on page 5273 for more information.

Cramer-von Mises Test

The Cramer-von Mises statistic is defined as

$$CM = \frac{1}{n^2} \sum_i \left(n_i \sum_{j=1}^{p} t_j \left(F_i(x_j) - F(x_j) \right)^2 \right)$$

where t_j is the number of ties at the jth distinct value and p is the number of distinct values. The asymptotic value is computed as

$$CM_a = CM \times n$$

PROC NPAR1WAY displays the contribution of each class level to the sum CM_a.

Kuiper Test

For data with two class levels, PROC NPAR1WAY computes the Kuiper statistic, its scaled value for the asymptotic distribution, and the asymptotic p-value. The Kuiper statistic is computed as

$$K = \max_j \left(F_1(x_j) - F_2(x_j) \right) - \min_j \left(F_1(x_j) - F_2(x_j) \right) \quad \text{where} \quad j = 1, 2, \ldots, n$$

The asymptotic value is

$$K_a = K \sqrt{n_1\, n_2 \,/\, n}$$

PROC NPAR1WAY displays the value of $(\max_j \left| F_1(x_j) - F_2(x_j) \right|)$ for each class level.

The p-value for the Kuiper test is the probability of observing a larger value of K_a under the null hypothesis of no difference between the two classes. PROC NPAR1WAY computes this p-value according to Owen (1962, p. 441).

Exact Tests

PROC NPAR1WAY provides exact p-values for tests for location and scale differences based on the following scores: Wilcoxon, median, van der Waerden (normal), Savage, Siegel-Tukey, Ansari-Bradley, Klotz, Mood, and Conover. Additionally, PROC NPAR1WAY provides exact p-values for tests that use the raw data as scores. Exact tests are available for two-sample and multisample data. When the data are classified into two samples, tests are based on simple linear rank statistics. When the data are classified into more than two samples, tests are based on one-way ANOVA statistics.

Exact tests can be useful in situations where the asymptotic assumptions are not met and the asymptotic p-values are not close approximations for the true p-values. Standard asymptotic methods involve the assumption that the test statistic follows a particular distribution when the sample size is sufficiently large. When the sample size is not large, asymptotic results might not be valid, with the asymptotic p-values differing perhaps substantially from the exact p-values. Asymptotic results might also be unreliable when the distribution of the data is sparse, skewed, or heavily tied. See Agresti (2007) and Bishop, Fienberg, and Holland (1975). Exact computations are based on the statistical theory of exact conditional inference for contingency tables, reviewed by Agresti (1992).

In addition to computation of exact p-values, PROC NPAR1WAY provides the option of estimating exact p-values by Monte Carlo simulation. This can be useful for problems that are so large that exact computations require a great amount of time and memory, but for which asymptotic approximations might not be sufficient.

The following sections summarize the exact computational algorithms, define the exact p-values that PROC NPAR1WAY computes, discuss the computational resource requirements, and describe the Monte Carlo estimation option.

Computational Algorithms

PROC NPAR1WAY computes exact p-values by using the network algorithm developed by Mehta and Patel (1983). This algorithm provides a substantial advantage over direct enumeration, which can be very time-consuming and feasible only for small problems. See Agresti (1992) for a review of algorithms for computation of exact p-values, and see Mehta, Patel, and Tsiatis (1984) and Mehta, Patel, and Senchaudhuri (1991) for information about the performance of the network algorithm.

PROC NPAR1WAY constructs a contingency table from the input data, with rows formed by the levels of the classification variable and columns formed by the response variable values. The reference set for a given contingency table is the set of all contingency tables with the observed marginal row and column sums. Corresponding to this reference set, the network algorithm forms a directed acyclic network consisting of nodes in a number of stages. A path through the network corresponds to a distinct table in the reference set. The distances between nodes are defined so that the total distance of a path through the network is the corresponding value of the test statistic. At each node, the algorithm computes the shortest and longest path distances for all the paths that pass through that node. For the two-sample linear rank statistics, which can be expressed as linear combinations of cell frequencies multiplied by increasing row and column scores, PROC NPAR1WAY computes shortest and longest path distances by using the algorithm given by Agresti, Mehta, and Patel (1990). For the multisample one-way test statistics, PROC NPAR1WAY computes an upper bound for the longest path and a lower bound for the shortest path by following the approach of Valz and Thompson (1994).

The longest and shortest path distances (bounds) for a node are compared to the value of the test statistic to determine whether all paths through the node contribute to the *p*-value, none of the paths through the node contribute to the *p*-value, or neither of these situations occurs. If all paths through the node contribute, the *p*-value is incremented accordingly, and these paths are eliminated from further analysis. If no paths contribute, these paths are eliminated from the analysis. Otherwise, the algorithm continues, still processing this node and the associated paths. The algorithm finishes when all nodes have been accounted for.

In applying the network algorithm, PROC NPAR1WAY uses full numerical precision to represent all statistics, row and column scores, and other quantities involved in the computations. Although it is possible to use rounding to improve the speed and memory requirements of the algorithm, PROC NPAR1WAY does not do this because it can result in reduced accuracy of the *p*-values.

Definition of p-Values

For two-sample linear rank tests, PROC NPAR1WAY computes exact one-sided and two-sided *p*-values for each test that is specified in the EXACT statement. For the one-sided test, PROC NPAR1WAY displays the right-sided *p*-value when the observed value of the test statistic is greater than its expected value. The right-sided *p*-value is the sum of probabilities for those tables having a test statistic greater than or equal to the observed test statistic. Otherwise, when the test statistic is less than or equal to its expected value, PROC NPAR1WAY displays the left-sided *p*-value. The left-sided *p*-value is the sum of probabilities for those tables having a test statistic less than or equal to the one observed. The one-sided *p*-value P_1 can be expressed as

$$P_1(t) = \begin{cases} \text{Prob(Test Statistic} \geq t \text{)} & \text{if } \ t > \text{E}_0(T) \\ \text{Prob(Test Statistic} \leq t \text{)} & \text{if } \ t \leq \text{E}_0(T) \end{cases}$$

where t is the observed value of the test statistic and $\text{E}_0(T)$ is the expected value of the test statistic under the null hypothesis. PROC NPAR1WAY computes the two-sided *p*-value as the sum of the one-sided *p*-value and the corresponding area in the opposite tail of the distribution of the statistic, equidistant from the expected value. The two-sided *p*-value P_2 can be expressed as

$$P_2(t) = \text{Prob}\left(|\text{Test Statistic} - \text{E}_0(T)| \geq |t - \text{E}_0(T)|\right)$$

For multisample data, the tests are based on one-way ANOVA statistics. For a test of this form, large values of the test statistic indicate a departure from the null hypothesis; the test is inherently two-sided. The exact *p*-value is the sum of probabilities for those tables having a test statistic greater than or equal to the value of the observed test statistic.

If you specify the POINT option in the EXACT statement, PROC NPAR1WAY also displays exact point probabilities for the test statistics. The exact point probability is the exact probability that the test statistic equals the observed value.

Computational Resources

PROC NPAR1WAY uses relatively fast and efficient algorithms for exact computations. These algorithms, together with improvements in computer power, now make it feasible to perform exact computations for data sets where previously only asymptotic methods could be applied. Nevertheless, there are still large problems that might require a prohibitive amount of time and memory for exact computations, depending on the speed and memory available on your computer. For large problems, consider whether exact methods are really needed or whether asymptotic methods might give results quite close to the exact results while requiring much less computer time and memory. When asymptotic methods might not be sufficient for such large problems, consider using Monte Carlo estimation of exact *p*-values, as described in the section "Monte Carlo Estimation" on page 5275.

A formula does not exist that can predict in advance how much time and memory are needed to compute an exact *p*-value for a certain problem. The time and memory required depend on several factors, including which test is being performed, the total sample size, the number of rows and columns, and the specific arrangement of the observations into table cells. Generally, larger problems (in terms of total sample size, number of rows, and number of columns) tend to require more time and memory. Additionally, for a fixed total sample size, time and memory requirements tend to increase as the number of rows and columns increase, since this corresponds to an increase in the number of tables in the reference set. Also for a fixed sample size, time and memory requirements increase as the marginal row and column totals become more homogeneous. See Agresti, Mehta, and Patel (1990) and Gail and Mantel (1977) for details.

At any time while PROC NPAR1WAY is computing exact *p*-values, you can terminate the computations by pressing the system interrupt key sequence (see the *SAS Companion* for your system) and choosing to stop computations. After you terminate exact computations, PROC NPAR1WAY completes all other remaining tasks. The procedure produces the requested output and reports missing values for any exact *p*-values not computed by the time of termination.

You can also use the MAXTIME= option in the EXACT statement to limit the amount of time PROC NPAR1WAY uses for exact computations. You specify a MAXTIME= value that is the maximum amount of time (in seconds) that PROC NPAR1WAY can use to compute an exact *p*-value. If PROC NPAR1WAY does not finish computing the exact *p*-value within that time, it terminates the computation and completes all other remaining tasks.

Monte Carlo Estimation

If you specify the MC option in the EXACT statement, PROC NPAR1WAY computes Monte Carlo estimates of the exact *p*-values instead of directly computing the exact *p*-values. Monte Carlo estimation can be useful for large problems that require a great amount of time and memory for exact computations but for which asymptotic approximations might not be sufficient. To describe the precision of each Monte Carlo estimate, PROC NPAR1WAY provides the asymptotic standard error and $100(1-\alpha)\%$ confidence limits. The confidence level α is determined by the ALPHA= option in the EXACT statement, which, by default, equals 0.01 and produces 99% confidence limits. The N= option in the EXACT statement specifies the number of samples PROC NPAR1WAY uses for Monte Carlo estimation; the default is 10,000 samples. You can specify a larger value for n to improve the precision of the Monte Carlo estimates. Because larger values of n generate more samples, the computation time increases. Or you can specify a smaller value of n to reduce the computation time.

To compute a Monte Carlo estimate of an exact *p*-value, PROC NPAR1WAY generates a random sample of tables with the same total sample size, row totals, and column totals as the observed table. PROC NPAR1WAY uses the algorithm of Agresti, Wackerly, and Boyett (1979), which generates tables in proportion to their hypergeometric probabilities conditional on the marginal frequencies. For each sample table, PROC NPAR1WAY computes the value of the test statistic and compares it to the value for the observed table. When estimating a right-sided p-value, PROC NPAR1WAY counts all sample tables for which the test statistic is greater than or equal to the observed test statistic. Then the p-value estimate equals the number of these tables divided by the total number of tables sampled, which can be written as

$$\begin{aligned} \hat{P}_{MC} &= M \;/\; N \\ M &= \text{number of samples with (Test Statistic} \geq t) \\ N &= \text{total number of samples} \\ t &= \text{observed Test Statistic} \end{aligned}$$

PROC NPAR1WAY computes left-sided and two-sided p-value estimates in a similar manner. For left-sided p-values, PROC NPAR1WAY evaluates whether the test statistic for each sampled table is less than or equal to the observed test statistic. For two-sided p-values, PROC NPAR1WAY examines the sample test statistics according to the expression for $P_2(t)$ given in the section "Definition of p-Values" on page 5274.

The variable M is a binomial variable with N trials and success probability p. It follows that the asymptotic standard error of the Monte Carlo estimate is

$$\text{se}(\hat{P}_{MC}) = \sqrt{\hat{P}_{MC}\,(1 - \hat{P}_{MC}) \;/\; (N - 1)}$$

PROC NPAR1WAY constructs asymptotic confidence limits for the p-values according to

$$\hat{P}_{MC} \;\pm\; \left(z_{\alpha/2} \;\times\; \text{se}(\hat{P}_{MC}) \right)$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution, and the confidence level α is determined by the ALPHA= option in the EXACT statement.

When the Monte Carlo estimate $\hat{P}_{MC}$ equals 0, PROC NPAR1WAY computes confidence limits for the p-value as

$$(\,0,\;\; 1 - \alpha^{(1/N)}\,)$$

When the Monte Carlo estimate $\hat{P}_{MC}$ equals 1, PROC NPAR1WAY computes the confidence limits as

$$(\,\alpha^{(1/N)},\;\; 1\,)$$

Output Data Set

The OUTPUT statement creates a SAS data set that contains statistics computed by PROC NPAR1WAY. You specify which statistics to store in the output data set by using options identical to those in the PROC NPAR1WAY statement. When you specify one of these options in the OUTPUT statement, PROC NPAR1WAY includes all available statistics from that analysis in the output data set.

The output data set contains one observation for each analysis variable within a BY group. (You can name the analysis variables in the VAR statement.) The OUTPUT data set includes the following variables:

- BY variables, if you use a BY statement
- _VAR_, which identifies the analysis variable
- Variables containing the specified statistics

Table 64.5 lists the variable names and descriptions for all available statistics. Note that some statistics are available only for the two-sample case (where the CLASS variable groups the data into two classes); other statistics are available only for the multisample case.

If you request exact tests by using the EXACT statement, the output data set also includes exact *p*-values for those tests when you specify the corresponding options in the OUTPUT statement. If you do not request exact tests with the EXACT statement, the output data set does not include exact *p*-values.

Monte Carlo estimates of exact *p*-values are not available in this output data set, but you can use the Output Delivery System (ODS) to store Monte Carlo estimates in a SAS data set. You can use the Output Delivery System to create a SAS data set from any piece of PROC NPAR1WAY output. For more information, see the section "ODS Table Names" on page 5287 and Chapter 20, "Using the Output Delivery System."

Table 64.5 Output Data Set Variable Names and Descriptions

Option	Output Variables		Variable Descriptions
AB	_AB_	*	Two-sample Ansari-Bradley statistic
	Z_AB	*	Ansari-Bradley statistic, standardized
	PL_AB	*	*p*-value (left-sided), Ansari-Bradley test
	PR_AB	*	*p*-value (right-sided), Ansari-Bradley test
	P2_AB	*	*p*-value (two-sided), Ansari-Bradley test
	XPL_AB	*	Exact *p*-value (left-sided), Ansari-Bradley test
	XPR_AB	*	Exact *p*-value (right-sided), Ansari-Bradley test
	XPT_AB	*	Exact point probability, Ansari-Bradley test
	XP2_AB	*	Exact *p*-value (two-sided), Ansari-Bradley test
	CHAB		Ansari-Bradley chi-square
	DF_CHAB		Degrees of freedom, Ansari-Bradley chi-square
	P_CHAB		*p*-value, Ansari-Bradley chi-square test
	XP_CHAB	**	Exact *p*-value, Ansari-Bradley chi-square test
	XPT_CHAB	**	Exact point probability, Ansari-Bradley chi-square

Table 64.5 *continued*

Option	Output Variables		Variable Descriptions
ANOVA	_MSA_		ANOVA effect mean square, among MS
	MSE		ANOVA error mean square, within MS
	_F_		*F* statistic for ANOVA
	P_F		*p*-value, *F* statistic for ANOVA
CONOVER	_CON_	*	Two-sample Conover statistic
	Z_CON	*	Conover statistic, standardized
	PL_CON	*	*p*-value (left-sided), Conover test
	PR_CON	*	*p*-value (right-sided), Conover test
	P2_CON	*	*p*-value (two-sided), Conover test
	XPL_CON	*	Exact *p*-value (left-sided), Conover test
	XPR_CON	*	Exact *p*-value (right-sided), Conover test
	XPT_CON	*	Exact point probability, Conover test
	XP2_CON	*	Exact *p*-value (two-sided), Conover test
	CHCON		Conover chi-square
	DF_CHCON		Degrees of freedom, Conover chi-square
	P_CHCON		*p*-value, Conover chi-square test
	XP_CHCON	**	Exact *p*-value, Conover chi-square test
	XPT_CHCO	**	Exact point probability, Conover chi-square
EDF	_KS_		Kolmogorov-Smirnov statistic
	KSA		Kolmogorov-Smirnov statistic (asymptotic)
	_Dp_	*	Two-sample Kolmogorov-Smirnov D+
	P_Dp	*	*p*-value, D+
	_Dm_	*	Two-sample Kolmogorov-Smirnov D-
	P_Dm	*	*p*-value, D-
	_D_	*	Two-sample Kolmogorov-Smirnov statistic D
	P_KSA	*	*p*-value, D
	XP_Dp	*	Exact *p*-value, D+
	XPT_Dp	*	Exact point probability, D+
	XP_Dm	*	Exact *p*-value, D-
	XPT_Dm	*	Exact point probability, D-
	XP_D	*	Exact *p*-value, D
	XPT_D	*	Exact point probability, D
	_CM_		Cramer-von Mises statistic
	CMA		Cramer-von Mises statistic (asymptotic)
	_K_	*	Kuiper two-sample statistic
	_KA_	*	Kuiper two-sample statistic (asymptotic)
	P_KA	*	*p*-value, two-sample Kuiper test

Table 64.5 *continued*

Option	Output Variables		Variable Descriptions
HL	_HL_	*	Hodges-Lehmann estimate, location shift
	L_HL	*	Lower confidence limit, Hodges-Lehmann
	U_HL	*	Upper confidence limit, Hodges-Lehmann
	M_HL	*	Confidence limit midpoint, Hodges-Lehmann
	E_HL	*	ASE of Hodges-Lehmann estimate
	XL_HL	*	Exact lower confidence limit, Hodges-Lehmann
	XU_HL	*	Exact upper confidence limit, Hodges-Lehmann
	XM_HL	*	Exact confidence limit midpoint
KLOTZ	_KLOTZ_	*	Two-sample Klotz statistic
	Z_K	*	Klotz statistic, standardized
	PL_K	*	*p*-value (left-sided), Klotz test
	PR_K	*	*p*-value (right-sided), Klotz test
	P2_K	*	*p*-value (two-sided), Klotz test
	XPL_K	*	Exact *p*-value (left-sided), Klotz test
	XPR_K	*	Exact *p*-value (right-sided), Klotz test
	XPT_K	*	Exact point probability, Klotz test
	XP2_K	*	Exact *p*-value (two-sided), Klotz test
	CHK		Klotz chi-square
	DF_CHK		Degrees of freedom, Klotz chi-square
	P_CHK		*p*-value, Klotz chi-square test
	XP_CHK	**	Exact *p*-value, Klotz chi-square test
	XPT_CHK	**	Exact point probability, Klotz chi-square
MEDIAN	_MED_	*	Two-sample median statistic
	Z_MED	*	Median statistic, standardized
	PL_MED	*	*p*-value (left-sided), median test
	PR_MED	*	*p*-value (right-sided), median test
	P2_MED	*	*p*-value (two-sided), median test
	XPL_MED	*	Exact *p*-value (left-sided), median test
	XPR_MED	*	Exact *p*-value (right-sided), median test
	XPT_MED	*	Exact point probability, median test
	XP2_MED	*	Exact *p*-value (two-sided), median test
	CHMED		Median chi-square (Brown-Mood test)
	DF_CHMED		Degrees of freedom, median chi-square
	P_CHMED		*p*-value, median chi-square test
	XP_CHMED	**	Exact *p*-value, median chi-square test
	XPT_CHME	**	Exact point probability, median chi-square

Table 64.5 *continued*

Option	Output Variables		Variable Descriptions
MOOD	_MOOD_	*	Two-sample Mood statistic
	Z_MOOD	*	Mood statistic, standardized
	PL_MOOD	*	*p*-value (left-sided), Mood test
	PR_MOOD	*	*p*-value (right-sided), Mood test
	P2_MOOD	*	*p*-value (two-sided), Mood test
	XPL_MOOD	*	Exact *p*-value (left-sided), Mood test
	XPR_MOOD	*	Exact *p*-value (right-sided), Mood test
	XPT_MOOD	*	Exact point probability, Mood test
	XP2_MOOD	*	Exact *p*-value (two-sided), Mood test
	CHMOOD		Mood chi-square
	DF_CHMOO		Degrees of Freedom, Mood chi-square
	P_CHMOOD		*p*-value, Mood chi-square test
	XP_CHMOO	**	Exact *p*-value, Mood chi-square test
	XPT_CHMO	**	Exact point probability, Mood chi-square
SAVAGE	_SAV_	*	Two-sample Savage statistic
	Z_SAV	*	Savage statistic, standardized
	PL_SAV	*	*p*-value (left-sided), Savage test
	PR_SAV	*	*p*-value (right-sided), Savage test
	P2_SAV	*	*p*-value (two-sided), Savage test
	XPL_SAV	*	Exact *p*-value (left-sided), Savage test
	XPR_SAV	*	Exact *p*-value (right-sided), Savage test
	XPT_SAV	*	Exact point probability, Savage test
	XP2_SAV	*	Exact *p*-value (two-sided), Savage test
	CHSAV		Savage chi-square
	DF_CHSAV		Degrees of freedom, Savage chi-square
	P_CHSAV		*p*-value, Savage chi-square test
	XP_CHSAV	**	Exact *p*-value, Savage chi-square test
	XPT_CHSA	**	Exact point probability, Savage chi-square
SCORES=DATA	_DATA_	*	Two-sample data scores statistic
	Z_DATA	*	Data scores statistic, standardized
	PL_DATA	*	*p*-value (left-sided), data scores test
	PR_DATA	*	*p*-value (right-sided), data scores test
	P2_DATA	*	*p*-value (two-sided), data scores test
	XPL_DATA	*	Exact *p*-value (left-sided), data scores test
	XPR_DATA	*	Exact *p*-value (right-sided), data scores test
	XPT_DATA	*	Exact point probability, data scores test
	XP2_DATA	*	Exact *p*-value (two-sided), data scores test
	CHDATA		Data scores chi-square
	DF_CHDAT		Degrees of freedom, data scores chi-square
	P_CHDATA		*p*-value, data scores chi-square test
	XP_CHDAT	**	Exact *p*-value, data scores chi-square test
	XPT_CHDA	**	Exact point probability, data scores chi-square

Table 64.5 *continued*

Option	Output Variables		Variable Descriptions
ST	_ST_	*	Two-sample Siegel-Tukey statistic
	Z_ST	*	Siegel-Tukey statistic, standardized
	PL_ST	*	*p*-value (left-sided), Siegel-Tukey test
	PR_ST	*	*p*-value (right-sided), Siegel-Tukey test
	P2_ST	*	*p*-value (two-sided), Siegel-Tukey test
	XPL_ST	*	Exact *p*-value (left-sided), Siegel-Tukey test
	XPR_ST	*	Exact *p*-value (right-sided), Siegel-Tukey test
	XPT_ST	*	Exact point probability, Siegel-Tukey test
	XP2_ST	*	Exact *p*-value (two-sided), Siegel-Tukey test
	CHST		Siegel-Tukey chi-square
	DF_CHST		Degrees of freedom, Siegel-Tukey chi-square
	P_CHST		*p*-value, Siegel-Tukey chi-square test
	XP_CHST	**	Exact *p*-value, Siegel-Tukey chi-square test
	XPT_CHST	**	Exact point probability, Siegel-Tukey chi-square
VW | NORMAL	_VW_	*	Two-sample Van der Waerden statistic
	Z_VW	*	Van der Waerden statistic, standardized
	PL_VW	*	*p*-value (left-sided), Van der Waerden test
	PR_VW	*	*p*-value (right-sided), Van der Waerden test
	P2_VW	*	*p*-value (two-sided), Van der Waerden test
	XPL_VW	*	Exact *p*-value (left-sided), Van der Waerden test
	XPR_VW	*	Exact *p*-value (right-sided), Van der Waerden test
	XPT_VW	*	Exact point probability, Van der Waerden test
	XP2_VW	*	Exact *p*-value (two-sided), Van der Waerden test
	CHVW		Van der Waerden chi-square
	DF_CHVW		Degrees of freedom, Van der Waerden chi-square
	P_CHVW		*p*-value, Van der Waerden chi-square test
	XP_CHVW	**	Exact *p*-value, Van der Waerden chi-square test
	XPT_CHVW	**	Exact point probability, Van der Waerden chi-square
WILCOXON	_WIL_	*	Two-sample Wilcoxon statistic
	Z_WIL	*	Wilcoxon statistic, standardized
	PL_WIL	*	*p*-value (left-sided), Wilcoxon test
	PR_WIL	*	*p*-value (right-sided), Wilcoxon test
	P2_WIL	*	*p*-value (two-sided), Wilcoxon test
	PTL_WIL	*	*p*-value (left-sided), Wilcoxon *t* approximation
	PTR_WIL	*	*p*-value (right-sided), Wilcoxon *t* approximation
	PT2_WIL	*	*p*-value (two-sided), Wilcoxon *t* approximation
	XPL_WIL	*	Exact *p*-value (left-sided), Wilcoxon test
	XPR_WIL	*	Exact *p*-value (right-sided), Wilcoxon test
	XPT_WIL	*	Exact point probability, Wilcoxon test
	XP2_WIL	*	Exact *p*-value (two-sided), Wilcoxon test

Table 64.5 *continued*

Option	Output Variables		Variable Descriptions
WILCOXON	_KW_		Kruskal-Wallis statistic
	DF_KW		Degrees of freedom, Kruskal-Wallis test
	P_KW		*p*-value, Kruskal-Wallis test
	XP_KW	**	Exact *p*-value, Kruskal-Wallis test
	XPT_KW	**	Exact point probability, Kruskal-Wallis test

* Statistic included only for two-sample cases.

** Statistic included only for multisample cases.

Displayed Output

If you specify the ANOVA option, PROC NPAR1WAY displays a "Class Means" table and an "Analysis of Variance" table for each response variable. The "Class Means" table includes the following information for each CLASS variable value (level):

- N, which is the number of observations
- Mean of the response variable

The "Analysis of Variance" table includes the following information for each Source of variation (Among classes and Within classes):

- DF, which is the degrees of freedom associated with the source
- Sum of Squares
- Mean Square, which is the sum of squares divided by the degrees of freedom

The "Analysis of Variance" table also includes the following:

- F Value for testing the hypothesis that the class means are equal, which is computed by dividing the Mean Square (Among) by the Mean Square (Within)
- Pr > F, which is the significance probability corresponding to the F Value

For each score type that you specify, PROC NPAR1WAY displays a “Class Scores” table. The available score types include Wilcoxon, median, Van der Waerden (normal), Savage, Siegel-Tukey, Ansari-Bradley, Klotz, Mood, Conover, and raw data scores. PROC NPAR1WAY computes the scores for the response variable values and classifies the scored observations according to the CLASS variable values. The “Class Scores” table includes the following information for each CLASS variable level:

- N, which is the number of observations
- Sum of Scores
- Expected Under H0, which is the expected sum of scores under the null hypothesis of no difference among classes
- Std Dev Under H0, which is the standard deviation under the null hypothesis
- Mean Score

When there are two levels of the CLASS variable, PROC NPAR1WAY displays a “Two-Sample Test” table for each analysis of scores. The “Two-Sample Test” table includes the following information:

- Statistic, which is the sum of scores for the class with the smaller sample size
- Z, which is the standardized test statistic and has an asymptotic standard normal distribution under the null hypothesis
- One-Sided Pr < Z or One-Sided Pr > Z, which is the asymptotic one-sided *p*-value. This is displayed as Pr < Z or Pr > Z depending on whether Z is ≤ 0 or > 0.
- Two-Sided Pr > |Z|, which is the asymptotic two-sided *p*-value

For Wilcoxon scores, the “Two-Sample Test” table also includes a *t* Approximation for the Wilcoxon two-sample test.

If you request an exact test by specifying the score type in the EXACT statement, the “Two-Sample Test” table also includes the following exact *p*-values:

- One-Sided Pr $\leq$ S or One-Sided Pr $\geq$ S, which is the exact one-sided *p*-value. This is displayed as Pr $\leq$ S or Pr $\geq$ S depending on whether S $\leq$ Mean or S > Mean, where S is the test statistic and Mean is its expected value under the null hypothesis.
- Point Pr = S, which is the point probability. This is displayed if you specify the POINT option in the EXACT statement.
- Two-Sided Pr $\geq$ |S - Mean|, which is the exact two-sided *p*-value

If you request Monte Carlo estimates for a two-sample exact test by specifying the MC option in the EXACT statement, PROC NPAR1WAY displays the “Monte Carlo Estimates for the Exact Test” table, which includes the following information:

- Estimate of One-Sided Pr $\leq$ S or One-Sided Pr $\geq$ S, which is the exact one-sided p-value, together with its Lower and Upper Confidence Limits
- Estimate of Two-Sided Pr $\geq$ |S - Mean|, which is the exact two-sided p-value, together with its Lower and Upper Confidence Limits
- Number of Samples used to compute the Monte Carlo estimates
- Initial Seed used to compute the Monte Carlo estimates

For both two-sample and multisample data, PROC NPAR1WAY displays a "One-Way Analysis" table, which includes the following information:

- Chi-Square, which is the one-way ANOVA statistic for testing the null hypothesis of no difference among classes
- DF, which is the degrees of freedom
- Pr > Chi-Square, which is the asymptotic p-value

For multisample data, if you request an exact test by specifying the score type in the EXACT statement, the "One-Way Analysis" table also displays the exact p-value as follows:

- Exact Pr $\geq$ Chi-Square
- Exact Pr = Chi-Square, which is the point probability. This is displayed if you specify the POINT option in the EXACT statement.

For multisample data, if you specify the MC option in the EXACT statement, PROC NPAR1WAY displays the following information in the "Monte Carlo Estimate for the Exact Test" table:

- Estimate of Exact Pr $\geq$ Chi-Square, together with its Lower and Upper Confidence Limits
- Number of Samples used to compute the Monte Carlo estimate
- Initial Seed used to compute the Monte Carlo estimate

If you specify the HL option for two-sample data, PROC NPAR1WAY produces a "Hodges-Lehmann Estimation" table, which includes the following information:

- Location Shift estimate
- Confidence Limits for the Location Shift
- Confidence Interval Midpoint
- Asymptotic Standard Error estimate, which is based on the confidence interval

If you request exact Hodges-Lehmann confidence limits by specifying the HL option in the EXACT statement, the “Hodges-Lehmann Estimation” table also includes Exact Confidence Limits and the exact Interval Midpoint.

If you specify the EDF option, PROC NPAR1WAY produces tables for the Kolmogorov-Smirnov test, the Cramer-von Mises test, and for two-sample data only, the Kuiper test.

The “Kolmogorov-Smirnov Test” table includes the following information for each CLASS variable level:

- N, which is the number of observations
- EDF at Maximum, which is the value of the class EDF (empirical distribution function) at its maximum deviation from the pooled EDF
- Deviation from Mean at Maximum, which is the value of $\sqrt{n_i}\sqrt{F_i - F}$ at its maximum, where n_i is the class sample size, F_i is the class EDF, and F is the pooled EDF

The “Kolmogorov-Smirnov Test” table displays the following statistics:

- KS, which is the Kolmogorov-Smirnov statistic
- KSa, which is the asymptotic Kolmogorov-Smirnov statistic, KSa = $\sqrt{n}$ KS

For two-sample data, the “Kolmogorov-Smirnov Test” table also displays the following statistics:

- Pr > KSa, which is the asymptotic p-value for KSa and equals Pr > D
- D, which is the two-sample Kolmogorov-Smirnov statistic, $\max_j |F_1(x_j) - F_2(x_j)|$

If you specify the D option for two-sample data, PROC NPAR1WAY displays the following one-sided Kolmogorov-Smirnov statistics and their asymptotic p-values in the “Kolmogorov-Smirnov Two-Sample Test” table:

- D+, which is $\max_j (F_1(x_j) - F_2(x_j))$
- Pr > D+
- D-, which is $\max_j (F_2(x_j) - F_1(x_j))$
- Pr > D-

For two-sample data, if you request an exact Kolmogorov-Smirnov test by specifying the KS option in the EXACT statement, PROC NPAR1WAY displays the following exact p-values in the “Kolmogorov-Smirnov Two-Sample Test” table:

- Exact Pr $\geq$ D
- Exact Pr $\geq$ D+

- Exact Pr $\geq$ D-
- Exact Point Pr = D, Exact Point Pr = D+, and Exact Point Pr = D-, if you specify the POINT option in the EXACT statement

If you request Monte Carlo estimates for the two-sample exact Kolmogorov-Smirnov test, PROC NPAR1WAY displays the following information in the “Kolmogorov-Smirnov Two-Sample Test” table:

- Estimate of Exact Pr $\geq$ D, together with its Lower and Upper Confidence Limits
- Estimate of Exact Pr $\geq$ D+, together with its Lower and Upper Confidence Limits
- Estimate of Exact Pr $\geq$ D-, together with its Lower and Upper Confidence Limits
- Number of Samples used to compute the Monte Carlo estimates
- Initial Seed used to compute the Monte Carlo estimates

The “Cramer-von Mises Test” table includes the following information for each CLASS variable level:

- N, which is the number of observations
- Summed Deviation from Mean, which is $(n_i/n) \sum_{j=1}^{p} t_j \left(F_i(x_j) - F(x_j) \right)^2$

The “Cramer-von Mises Statistics” table displays the following statistics:

- CM, which is the Cramer-von Mises statistic
- CMa, which is the asymptotic Cramer-von Mises statistic, CMa = n CM

For two-sample data, PROC NPAR1WAY displays the “Kuiper Test” table, which includes the following information for each CLASS variable level:

- N, which is the number of observations
- Deviation from Mean, which is $\max_j |F_1(x_j) - F_2(x_j)|$

The “Kuiper Two-Sample Statistics” table displays the following statistics:

- K, which is the Kuiper two-sample test statistic
- Ka, which is the asymptotic Kuiper two-sample test statistic, Ka = K $\sqrt{n_1 n_2 / n}$
- Pr > Ka

ODS Table Names

PROC NPAR1WAY assigns a name to each table that it creates. You can use these names to refer to tables when you use the Output Delivery System (ODS) to select tables and create output data sets. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

Table 64.6 lists the ODS table names together with their descriptions and the options required to produce the tables. If you do not specify any analysis options in the PROC NPAR1WAY statement, the procedure provides the ANOVA, WILCOXON, MEDIAN, VW (NORMAL), SAVAGE, and EDF analyses by default.

Table 64.6 ODS Tables Produced by PROC NPAR1WAY

ODS Table Name	Description	Statement	Option
ANOVA	Analysis of variance	PROC	ANOVA
ABAnalysis	Ansari-Bradley one-way analysis	PROC	AB
ABMC	Monte Carlo estimates for the Ansari-Bradley exact test	EXACT	AB / MC
ABScores	Ansari-Bradley scores	PROC	AB
ABTest *	Ansari-Bradley two-sample test	PROC	AB
ClassMeans	Class means	PROC	ANOVA
ConoverAnalysis	Conover one-way analysis	PROC	CONOVER
ConoverMC	Monte Carlo estimates for the Conover exact test	EXACT	CONOVER / MC
ConoverScores	Conover scores	PROC	CONOVER
ConoverTest *	Conover two-sample test	PROC	CONOVER
CVMStats	Cramer-von Mises statistics	PROC	EDF
CVMTest	Cramer-von Mises test	PROC	EDF
DataScores	Data scores	PROC	SCORES=DATA
DataScoresAnalysis	Data scores one-way analysis	PROC	SCORES=DATA
DataScoresMC	Monte Carlo estimates for the data scores exact test	EXACT	SCORES=DATA / MC
DataScoresTest *	Data scores two-sample test	PROC	SCORES=DATA
HodgesLehmann *	Hodges-Lehmann estimation	PROC	HL
KlotzAnalysis	Klotz one-way analysis	PROC	KLOTZ
KlotzMC	Monte Carlo estimates for the Klotz exact test	EXACT	KLOTZ / MC
KlotzScores	Klotz scores	PROC	KLOTZ
KlotzTest *	Klotz two-sample test	PROC	KLOTZ
KolSmir2Stats *	Kolmogorov-Smirnov two-sample statistics	PROC	EDF
KolSmirExactTest *	Kolmogorov-Smirnov exact test	EXACT	KS | EDF
KolSmirStats **	Kolmogorov-Smirnov statistics	PROC	EDF
KolSmirTest	Kolmogorov-Smirnov test	PROC	EDF

Table 64.6 *continued*

ODS Table Name	Description	Statement	Option
KruskalWallisMC **	Monte Carlo estimates for the Kruskal-Wallis exact test	EXACT	WILCOXON / MC
KruskalWallisTest	Kruskal-Wallis test	PROC	WILCOXON
KSMC *	Monte Carlo estimates for the Kolmogorov-Smirnov exact test	EXACT	KS | EDF / MC
KuiperStats *	Kuiper two-sample statistics	PROC	EDF
KuiperTest *	Kuiper test	PROC	EDF
MedianAnalysis	Median one-way analysis	PROC	MEDIAN
MedianMC	Monte Carlo estimates for the median exact test	EXACT	MEDIAN / MC
MedianScores	Median scores	PROC	MEDIAN
MedianTest *	Median two-sample test	PROC	MEDIAN
MoodAnalysis	Mood one-way analysis	PROC	MOOD
MoodMC	Monte Carlo estimates for the Mood exact test	EXACT	MOOD / MC
MoodScores	Mood scores	PROC	MOOD
MoodTest *	Mood two-sample test	PROC	MOOD
SavageAnalysis	Savage one-way analysis	PROC	SAVAGE
SavageMC	Monte Carlo estimates for the Savage exact test	EXACT	SAVAGE / MC
SavageScores	Savage scores	PROC	SAVAGE
SavageTest *	Savage two-sample test	PROC	SAVAGE
STAnalysis	Siegel-Tukey one-way analysis	PROC	ST
STMC	Monte Carlo estimates for the Siegel-Tukey exact test	EXACT	ST / MC
STScores	Siegel-Tukey scores	PROC	ST
STTest *	Siegel-Tukey two-sample test	PROC	ST
VWAnalysis	Van der Waerden one-way analysis	PROC	VW | NORMAL
VWMC	Monte Carlo estimates for the Van der Waerden exact test	EXACT	VW | NORMAL / MC
VWScores	Van der Waerden scores	PROC	VW | NORMAL
VWTest *	Van der Waerden two-sample test	PROC	VW | NORMAL
WilcoxonMC *	Monte Carlo estimates for the Wilcoxon two-sample exact test	EXACT	WILCOXON / MC
WilcoxonScores	Wilcoxon scores	PROC	WILCOXON
WilcoxonTest *	Wilcoxon two-sample test	PROC	WILCOXON

* PROC NPAR1WAY produces this table only for two-sample data.

** PROC NPAR1WAY produces this table only for multisample data.

ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, "Statistical Graphics Using ODS."

Before you create graphs, ODS Graphics must be enabled (for example, with the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 609 in Chapter 21, "Statistical Graphics Using ODS."

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section "A Primer on ODS Statistical Graphics" on page 608 in Chapter 21, "Statistical Graphics Using ODS."

When ODS Graphics is enabled, you can request specific plots with the PLOTS= option in the PROC NPAR1WAY statement. If you do not specify the PLOTS= option but have enabled ODS Graphics, PROC NPAR1WAY produces all plots that are associated with the analyses that you request.

PROC NPAR1WAY assigns a name to each graph that it creates with ODS Graphics. You can use these names to refer to the graphs. Table 64.7 lists the names of the graphs that PROC NPAR1WAY generates together with their descriptions and the options that are required to produce the graphs.

Table 64.7 Graphs Produced by PROC NPAR1WAY

ODS Graph Name	Description	Option
ABBoxPlot	Box plot of Ansari-Bradley scores	AB
ANOVABoxPlot	Box plot of raw data	ANOVA
ConoverBoxPlot	Box plot of Conover scores	CONOVER
DataScoresBoxPlot	Box plot of data scores	SCORES=DATA
EDFPlot	Empirical distribution function plot	EDF
KlotzBoxPlot	Box plot of Klotz scores	KLOTZ
MedianPlot	Median plot	MEDIAN
MoodBoxPlot	Box plot of Mood scores	MOOD
SavageBoxPlot	Box plot of Savage scores	SAVAGE
STBoxPlot	Box plot of Siegel-Tukey scores	ST
VWBoxPlot	Box plot of Van der Waerden scores	VW | NORMAL
WilcoxonBoxPlot	Box plot of Wilcoxon scores	WILCOXON

Examples: NPAR1WAY Procedure

Example 64.1: Two-Sample Location Tests and Plots

Fifty-nine female patients with rheumatoid arthritis who participated in a clinical trial were assigned to two groups, active and placebo. The response status (excellent=5, good=4, moderate=3, fair=2, poor=1) of each patient was recorded.

The following SAS statements create the data set Arthritis, which contains the observed status values for all the patients. The variable Treatment denotes the treatment received by a patient, and the variable Response contains the response status of the patient. The variable Freq contains the frequency of the observation, which is the number of patients with the Treatment and Response combination.

```
data Arthritis;
   input Treatment $ Response Freq @@;
   datalines;
Active 5 5 Active 4 11 Active 3 5 Active 2 1 Active 1 5
Placebo 5 2 Placebo 4 4 Placebo 3 7 Placebo 2 7 Placebo 1 12
;
```

The following PROC NPAR1WAY statements test the null hypothesis that there is no difference in the patient response status against the alternative hypothesis that the patient response status differs in the two treatment groups. The WILCOXON option requests the Wilcoxon test for difference in location, and the MEDIAN option requests the median test for difference in location. The variable Treatment is the CLASS variable, and the VAR statement specifies that the variable Response is the analysis variable.

The PLOTS= option requests a box plot of the Wilcoxon scores and a median plot for Response classified by Treatment. ODS Graphics must be enabled before producing plots.

```
ods graphics on;
proc npar1way data=Arthritis wilcoxon median
              plots=(wilcoxonboxplot medianplot);
   class Treatment;
   var Response;
   freq Freq;
run;
ods graphics off;
```

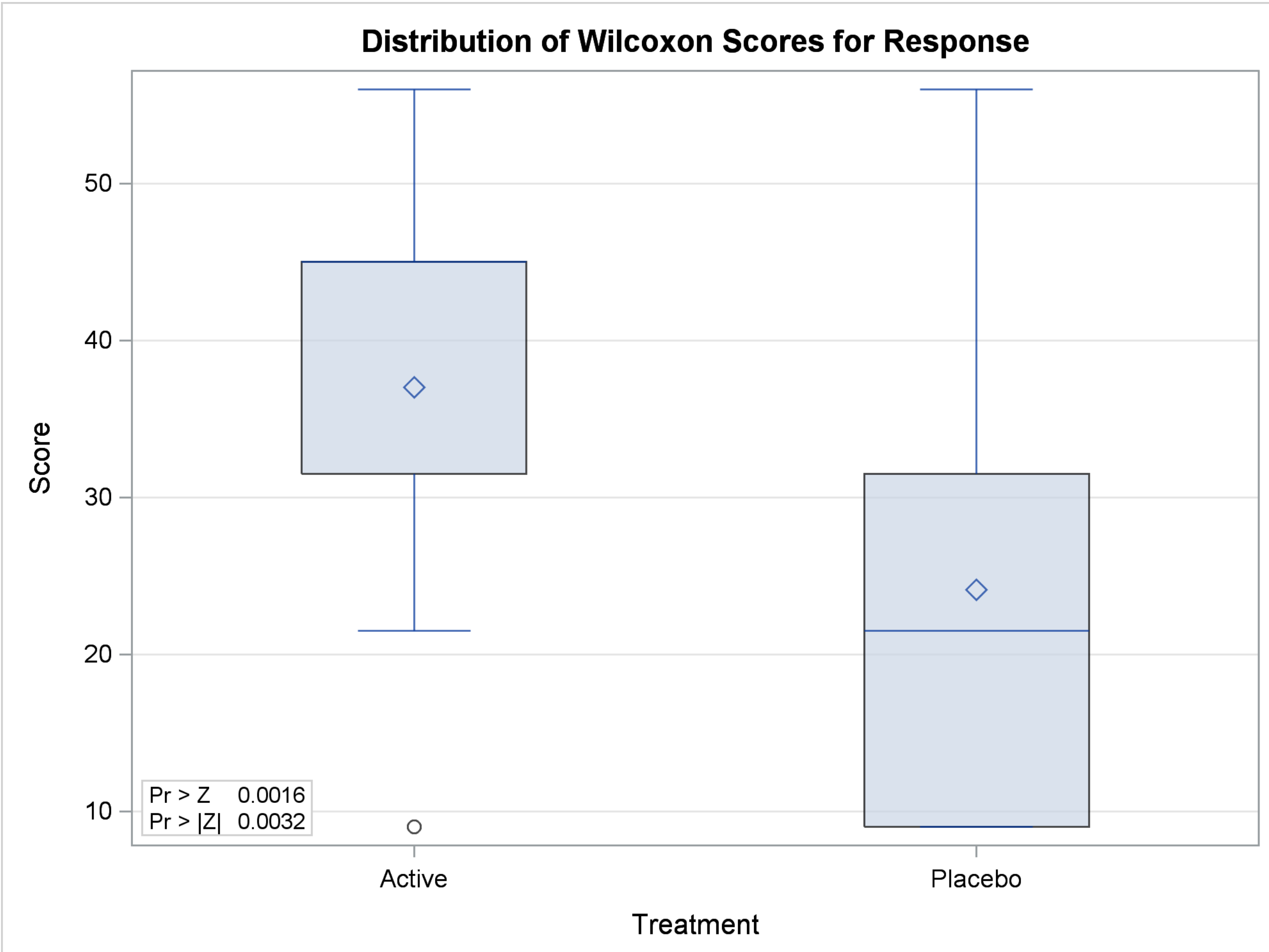
Output 64.1.1 shows the results of the Wilcoxon analysis. The Wilcoxon two-sample test statistic equals 999.0, which is the sum of the Wilcoxon scores for the smaller sample (Active). This sum is greater than 810.0, which is the expected value under the null hypothesis of no difference between the two samples, Active and Placebo. The one-sided p-value is 0.0016, which indicates that the patient response for the Active treatment is significantly more than for the Placebo group.

Output 64.1.1 Wilcoxon Two-Sample Test

```
                         The NPAR1WAY Procedure

            Wilcoxon Scores (Rank Sums) for Variable Response
                    Classified by Variable Treatment

                             Sum of      Expected       Std Dev          Mean
Treatment          N         Scores      Under H0      Under H0         Score
-------------------------------------------------------------------------------
Active            27          999.0         810.0     63.972744     37.000000
Placebo           32          771.0         960.0     63.972744     24.093750

                      Average scores were used for ties.

                            Wilcoxon Two-Sample Test

                        Statistic               999.0000

                        Normal Approximation
                        Z                         2.9466
                        One-Sided Pr >  Z         0.0016
                        Two-Sided Pr > |Z|        0.0032

                        t Approximation
                        One-Sided Pr >  Z         0.0023
                        Two-Sided Pr > |Z|        0.0046

                  Z includes a continuity correction of 0.5.

                              Kruskal-Wallis Test

                        Chi-Square                8.7284
                        DF                             1
                        Pr > Chi-Square           0.0031
```

Output 64.1.2 displays the box plot of Wilcoxon scores classified by Treatment, which corresponds to the Wilcoxon analysis in Output 64.1.1. To remove the *p*-values from the box plot display, you can specify the NOSTATS plot option in parentheses following the WILCOXONBOXPLOT option.

Output 64.1.2 Box Plot of Wilcoxon Scores



Output 64.1.3 shows the results of the median two-sample test. The test statistic equals 18.9167, and its standardized Z value is 3.1667. The one-sided p-value Pr > Z equals 0.0005. This supports the alternative hypothesis that the effect of the Active treatment is greater than that of the Placebo.

Output 64.1.4 displays the median plot for the analysis of Response classified by Treatment. The median plot is a stacked bar chart showing the frequencies above and below the overall median. This plot corresponds to the median scores analysis in Output 64.1.3.

Output 64.1.3 Median Two-Sample Test

Median Scores (Number of Points Above Median) for Variable Response
Classified by Variable Treatment

Treatment	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Active	27	18.916667	13.271186	1.728195	0.700617
Placebo	32	10.083333	15.728814	1.728195	0.315104

Average scores were used for ties.

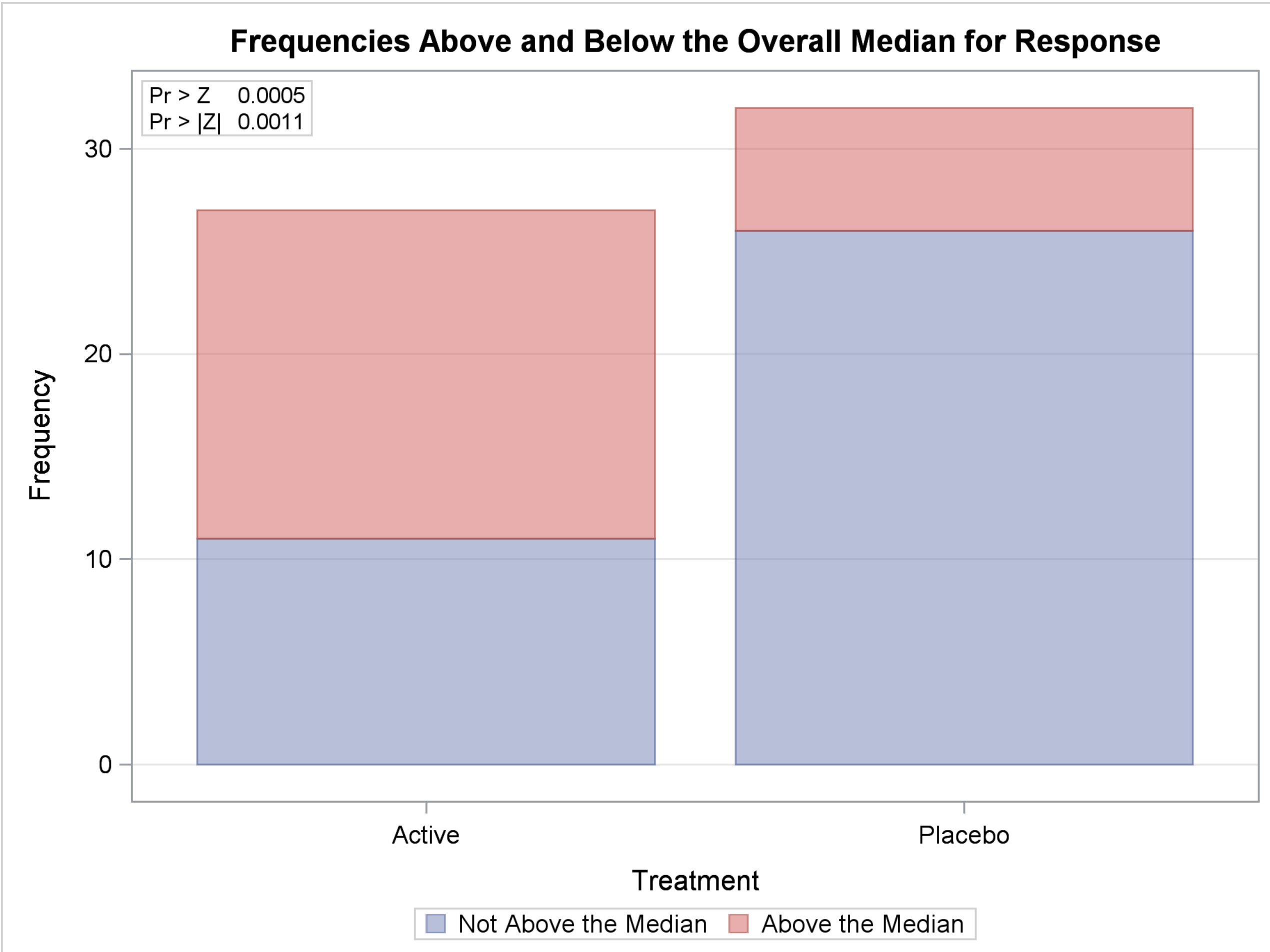
Median Two-Sample Test

Statistic	18.9167
Z	3.2667
One-Sided Pr > Z	0.0005
Two-Sided Pr > |Z|	0.0011

Median One-Way Analysis

Chi-Square	10.6713
DF	1
Pr > Chi-Square	0.0011

Output 64.1.4 Median Plot



Example 64.2: EDF Statistics and EDF Plot

This example uses the SAS data set Arthritis created in Example 64.1. The data set contains the variable Treatment, which denotes the treatment received by a patient, and the variable Response, which contains the response status of the patient. The variable Freq contains the frequency of the observation, which is the number of patients with the Treatment and Response combination.

The following statements request empirical distribution function (EDF) statistics, which test whether the distribution of a variable is the same across different groups. The EDF option requests the EDF analysis. The variable Treatment is the CLASS variable, and the variable Response specified in the VAR statement is the analysis variable. The FREQ statement names Freq as the frequency variable.

The PLOTS= option requests an EDF plot for Response classified by Treatment. ODS Graphics must be enabled before producing plots.

```
ods graphics on;
proc npar1way edf plots=edfplot data=Arthritis;
   class Treatment;
   var Response;
   freq Freq;
run;
ods graphics off;
```

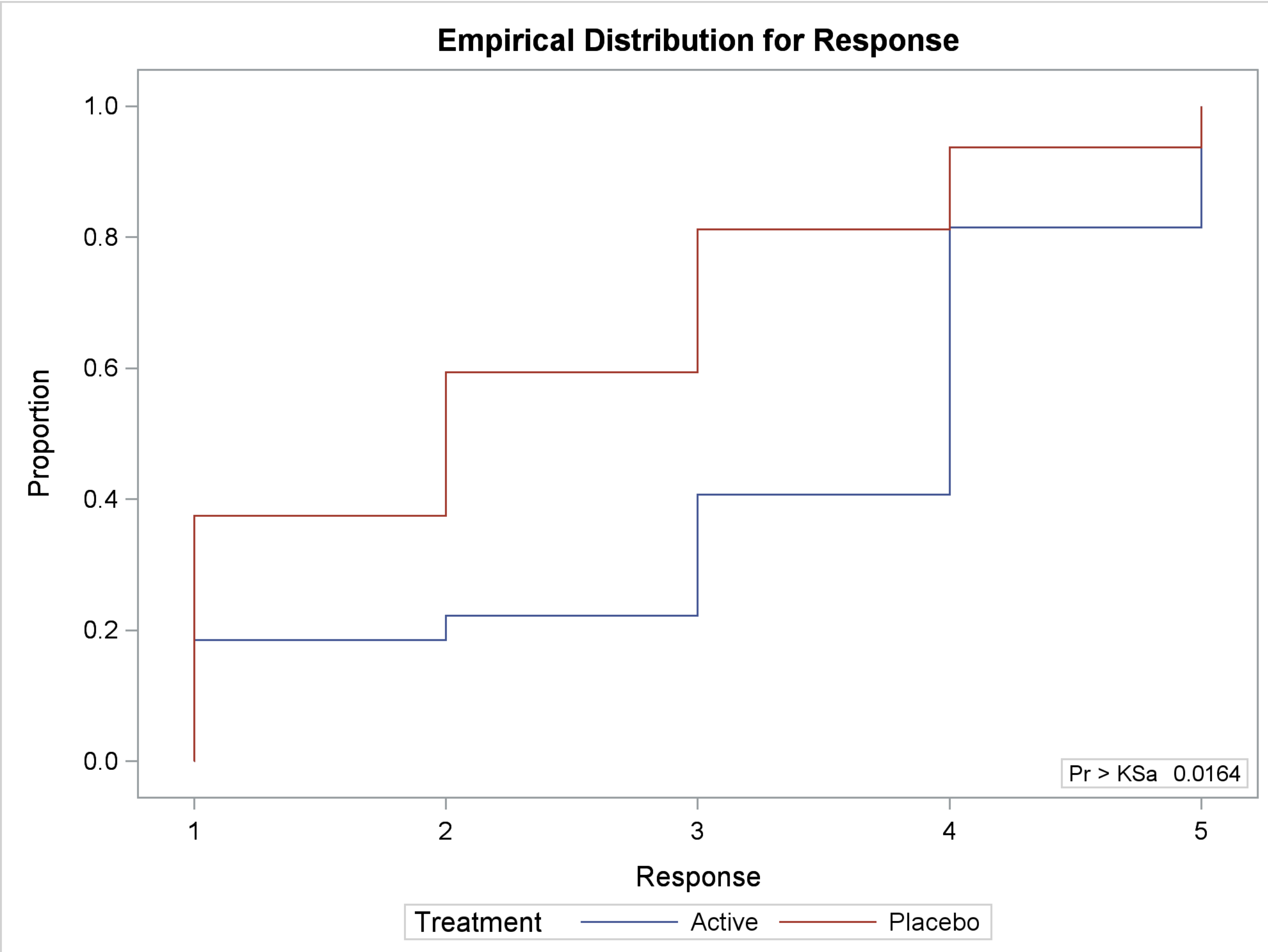
Output 64.2.1 shows EDF statistics that compare the two levels of Treatment, Active and Placebo. The asymptotic p-value for the Kolmogorov-Smirnov test is 0.0164. This supports rejection of the null hypothesis that the distributions are the same for the two samples.

Output 64.2.2 shows the EDF plot for Response classified by Treatment.

Output 64.2.1 Empirical Distribution Function Statistics

```
                    The NPAR1WAY Procedure

        Kolmogorov-Smirnov Test for Variable Response
               Classified by Variable Treatment

                          EDF at    Deviation from Mean
Treatment        N       Maximum        at Maximum
-------------------------------------------------------
Active          27      0.407407         -1.141653
Placebo         32      0.812500          1.048675
Total           59      0.627119

         Maximum Deviation Occurred at Observation 3
              Value of Response at Maximum = 3.0

       Kolmogorov-Smirnov Two-Sample Test (Asymptotic)
             KS    0.201818    D          0.405093
             KSa   1.550191    Pr > KSa   0.0164
```

Output 64.2.2 Empirical Distribution Function Plot



Example 64.3: Exact Wilcoxon Two-Sample Test

Researchers conducted an experiment to compare the effects of two stimulants. Thirteen randomly selected subjects received the first stimulant, and six randomly selected subjects received the second stimulant. The reaction times (in minutes) were measured while the subjects were under the influence of the stimulants.

The following SAS statements create the data set React, which contains the observed reaction times for each stimulant. The variable Stim represents Stimulant 1 or 2. The variable Time contains the reaction times observed for subjects under the stimulant.

```
data React;
   input Stim Time @@;
   datalines;
1 1.94   1 1.94   1 2.92   1 2.92   1 2.92   1 2.92   1 3.27
1 3.27   1 3.27   1 3.27   1 3.70   1 3.70   1 3.74
2 3.27   2 3.27   2 3.27   2 3.70   2 3.70   2 3.74
;
```

The following statements request a Wilcoxon test of the null hypothesis that there is no difference between the effects of the two stimulants. Stim is the CLASS variable, and Time is the analysis variable. The

WILCOXON option requests an analysis of Wilcoxon scores. The CORRECT=NO option removes the continuity correction from the computation of the standardized z test statistic. The WILCOXON option in the EXACT statement requests exact p-values for the Wilcoxon test. Because the sample size is small, the large-sample normal approximation might not be adequate, and it is appropriate to compute the exact test. These statements produce the results shown in Output 64.3.1.

```
proc npar1way wilcoxon correct=no data=React;
   class Stim;
   var Time;
   exact wilcoxon;
run;
```

Output 64.3.1 displays the results of the Wilcoxon two-sample test. The Wilcoxon statistic equals 79.50. Since this value is greater than 60.0, the expected value under the null hypothesis, PROC NPAR1WAY displays the right-sided p-values. The normal approximation for the Wilcoxon two-sample test yields a one-sided p-value of 0.0421 and a two-sided p-value of 0.0843. For the exact Wilcoxon test, the one-sided p-value is 0.0527, and the two-sided p-value is 0.1054.

Output 64.3.1 Wilcoxon Two-Sample Test

```
                        The NPAR1WAY Procedure

            Wilcoxon Scores (Rank Sums) for Variable Time
                     Classified by Variable Stim

                         Sum of      Expected       Std Dev          Mean
Stim         N           Scores      Under H0      Under H0         Score
--------------------------------------------------------------------------
1           13           110.50         130.0     11.004784         8.500
2            6            79.50          60.0     11.004784        13.250

                  Average scores were used for ties.

                      Wilcoxon Two-Sample Test

                 Statistic (S)                79.5000

                 Normal Approximation
                 Z                             1.7720
                 One-Sided Pr >  Z             0.0382
                 Two-Sided Pr > |Z|            0.0764

                 t Approximation
                 One-Sided Pr >  Z             0.0467
                 Two-Sided Pr > |Z|            0.0933

                 Exact Test
                 One-Sided Pr >=  S            0.0527
                 Two-Sided Pr >= |S - Mean|    0.1054

                        Kruskal-Wallis Test

                      Chi-Square         3.1398
                      DF                      1
                      Pr > Chi-Square    0.0764
```

Example 64.4: Hodges-Lehmann Estimation

This example uses the SAS data set React created in Example 64.3. The data set contains the variable Stim, which represents Stimulant 1 or 2, and the variable Time, which contains the reaction times observed for subjects under the stimulant.

The following statements request Hodges-Lehmann estimation of the location shift between the two groups. Stim is the CLASS variable, and Time is the analysis variable. The HL option requests Hodges-Lehmann estimation. The ALPHA= option sets the confidence level for the Hodges-Lehmann confidence limits. The HL option in the EXACT statement requests exact confidence limits for the estimate of location shift. The ODS SELECT statement selects which tables to display. Output 64.4.1 shows the Hodges-Lehmann results.

```
proc npar1way hl alpha=.02 data=React;
   class Stim;
   var Time;
   exact hl;
   ods select WilcoxonScores HodgesLehmann;
run;
```

The HL option automatically invokes the WILCOXON option, producing a table of Wilcoxon scores (Output 64.4.1). The Hodges-Lehmann estimate of location shift is 0.35, and the asymptotic confidence limits are 0.00 and 0.82. The confidence interval midpoint equals 0.41, which can also be used as an estimate of the location shift. The ASE estimate of 0.1762 is based on the length of the confidence interval. The exact confidence limits are 0.00 and 1.33.

Output 64.4.1 Hodges-Lehmann Estimate of Location Shift

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Time
Classified by Variable Stim

Stim	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	13	110.50	130.0	11.004784	8.500
2	6	79.50	60.0	11.004784	13.250

Average scores were used for ties.

Hodges-Lehmann Estimation

Location Shift   0.3500

Type	98% Confidence Limits		Interval Midpoint	Asymptotic Standard Error
Asymptotic (Moses)	0.0000	0.8200	0.4100	0.1762
Exact	0.0000	1.3300	0.6650	

Example 64.5: Exact Savage Multisample Test

A researcher conducting a laboratory experiment randomly assigned 15 mice to receive one of three drugs. The survival time (in days) was then recorded.

The following SAS statements create the data set Mice, which contains the observed survival times for the mice. The variable Treatment denotes the treatment received. The variable Days contains the number of days the mouse survived.

```
data Mice;
   input Treatment $ Days @@;
   datalines;
1 1 1 1 1 3 1 3 1 4
2 3 2 4 2 4 2 4 2 15
3 4 3 4 3 10 3 10 3 26
;
```

The following statements request a Savage test of the null hypothesis that there is no difference in survival time among the three drugs. Treatment is the CLASS variable, and Days is the analysis variable. The SAVAGE option requests an analysis of Savage scores. The SAVAGE option in the EXACT statement requests exact p-values for the Savage test. Because the sample size is small, the large-sample normal approximation might not be adequate, and it is appropriate to compute the exact test.

PROC NPAR1WAY tests the null hypothesis that there is no difference in the survival times among the three drugs against the alternative hypothesis of difference among the drugs. The SAVAGE option specifies an analysis based on Savage scores. The variable Treatment is the CLASS variable, and the variable Days is the response variable. The EXACT statement requests the exact Savage test.

```
proc npar1way savage data=Mice;
   class Treatment;
   var Days;
   exact savage;
run;
```

Output 64.5.1 shows the results of the Savage test. The exact p-value is 0.0445, which supports a difference in survival times among the drugs at the 0.05 level. The asymptotic p-value based on the chi-square approximation is 0.0638.

Output 64.5.1 Savage Multisample Exact Test

```
                          The NPAR1WAY Procedure

             Savage Scores (Exponential) for Variable Days
                   Classified by Variable Treatment

                          Sum of      Expected       Std Dev          Mean
Treatment        N        Scores      Under H0      Under H0         Score
--------------------------------------------------------------------------
1                5     -3.367980           0.0      1.634555     -0.673596
2                5      0.095618           0.0      1.634555      0.019124
3                5      3.272362           0.0      1.634555      0.654472

                    Average scores were used for ties.

                         Savage One-Way Analysis

                  Chi-Square                        5.5047
                  DF                                     2
                  Asymptotic Pr >  Chi-Square       0.0638
                  Exact      Pr >= Chi-Square       0.0445
```

References

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7 (1), 131–177.

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, Second Edition, New York: John Wiley & Sons.

Agresti, A. (2002), *Categorical Data Analysis*, Second Edition, New York: John Wiley & Sons.

Agresti, A., Mehta, C. R., and Patel, N. R. (1990), "Exact Inference for Contingency Tables with Ordered Categories," *Journal of American Statistical Association*, 85, 453–458.

Agresti, A., Wackerly, D., and Boyett, J. M. (1979), "Exact Conditional Tests for Cross-Classifications: Approximation of Attained Significance Levels," *Psychometrika*, 44, 75–83.

Bishop, Y., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Conover, W. J. (1999), *Practical Nonparametric Statistics*, Third Edition, New York: John Wiley & Sons.

Gail, M. and Mantel, N. (1977), "Counting the Number of $r \times c$ Contingency Tables with Fixed Margins," *Journal of the American Statistical Association*, 72, 859–862.

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

Hajek, J. (1969), *A Course in Nonparametric Statistics*, San Francisco: Holden-Day.

Halverson, J. O. and Sherwood, F. W. (1930), "Investigations in the Feeding of Cottonseed Meal to Cattle," *North Carolina Agr. Exp. Sta. Tech. Bulletin*, 39, 158pp.

Hodges, J. L., Jr. (1957), "The Significance Probability of the Smirnov Two-Sample Test," *Arkiv for Matematik*, 3, 469–486.

Hodges, J. L., Jr. and Lehmann, E. L. (1983). "Hodges-Lehmann Estimators," in *Encyclopedia of Statistical Sciences*, vol. 3, ed. S. Kotz, N. L. Johnson, and C. B. Read, New York: John Wiley & Sons, 463–465.

Hollander, M. and Wolfe, D. A. (1999), *Nonparametric Statistical Methods*, Second Edition, New York: John Wiley & Sons.

Kiefer, J. (1959), "K-Sample Analogues of the Kolmogorov-Smirnov and Cramer-von Mises Tests," *Annals of Mathematical Statistics*, 30, 420–447.

Lehmann, E. L. (1963). "Nonparametric Confidence Intervals for a Shift Parameter," *Annals of Mathematical Statistics*, 34, 1507–1512.

Mehta, C. R. and Patel, N. R. (1983), "A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables," *Journal of American Statistical Association*, 78, 427–434.

Mehta, C. R., Patel, N. R., and Senchaudhuri, P. (1991), "Exact Stratified Linear Rank Tests for Binary Data," *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E. M. Keramidas, ed.), 200–207.

Mehta, C. R., Patel, N. R., and Tsiatis, A. A. (1984), "Exact Significance Testing to Establish Treatment Equivalence with Ordered Categorical Data," *Biometrics*, 40, 819–825.

Owen, D. B. (1962), *Handbook of Statistical Tables*, Reading, MA: Addison-Wesley.

Quade, D. (1966), "On Analysis of Variance for the k-Sample Problem," *Annals of Mathematical Statistics*, 37, 1747–1758.

Randles, R. H. and Wolfe, D. A. (1979), *Introduction to the Theory of Nonparametric Statistics*, New York: John Wiley & Sons.

Sheskin, D. J. (1997), *Handbook of Parametric and Nonparametric Statistical Procedures*, Boca Raton, FL: CRC Press.

Valz, P. D. and Thompson, M. E. (1994), "Exact Inference for Kendall's S and Spearman's ρ with Extensions to Fisher's Exact Test in $r \times c$ Contingency Tables," *Journal of Computational and Graphical Statistics*, 3 (4), 459–472.

Subject Index

Syntax Index

O

P

S

V

W

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **`yourturn@sas.com`**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **`suggest@sas.com`**.

SAS® Publishing Delivers!

Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.

SAS® Press

Need to learn the basics? Struggling with a programming problem? You'll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

support.sas.com/saspress

SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – **free** on the Web.
- Hard-copy books.

support.sas.com/publishing

SAS® Publishing News

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

support.sas.com/spn

sas THE POWER TO KNOW®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2009 SAS Institute Inc. All rights reserved. 518177_1US.0109