

Nonparametric Inference for Ordered Alternatives in a One-Way Layout Using the SAS® System

Paul Juneau, Pfizer Global Research & Development, Ann Arbor, Michigan

ABSTRACT

The one-way layout is a commonly used experimental design in scientific research (Juneau, 2004). With such a design it is easy to test alternative statistical hypotheses of interest including all pair-wise comparisons of treatment effects or all pair-wise comparisons with a designated (control) group. Sometimes a natural ordering exists among the treatments in an experiment; e.g., the treatment levels in the study may represent increasing doses of a single agent and an investigator may conjecture, *a priori*, that the response will be decreasing (or increasing) with increasing levels of dose. It could also be the case that an investigator might wish to study if the response increases up to a certain treatment level and then begins to decrease over the remaining ordered levels. SAS® affords the user the ability to analyze data in both of these settings.

The author will review the analysis of monotonic responses in a one-way layout using the Jonckheere-Terpstra statistic in PROC FREQ and then introduce a set of original macros to analyze non-monotonic, or “umbrella” responses with the SAS® system. He will briefly review the statistical theory of these two approaches and demonstrate the necessary code to perform the data analysis on simulated and actual data sets from scientific research.

INTRODUCTION: THE ONE-WAY LAYOUT SETTING

Consider a simple experiment set up in the form of a one-way layout: experimental units (e.g., subjects) are randomly assigned to two or more treatment groups and each unit receives its designated treatment. The treatment groups could consist of two or more distinct agents or increasing doses of a single treatment (as in the case of later stage drug discovery studies or many standard toxicology investigations). After each unit is treated, a continuous measurement is collected (often called the *response* of the treatment). Formally, let X_{ij} = the response of the j th experiment unit ($1 \leq j \leq n_i$) to the i th treatment ($1 \leq i \leq k$):

$$\begin{array}{ll}
 \text{Group 1: } n_1 \text{ units treated with Treatment 1:} & X_{11}, X_{12}, \dots, X_{1n_1} \\
 \text{Group 2: } n_2 \text{ units treated with Treatment 2:} & X_{21}, X_{22}, \dots, X_{2n_2} \\
 & \vdots \\
 & \vdots \\
 \text{Group } i: n_i \text{ units treated with Treatment } i: & X_{i1}, X_{i2}, \dots, X_{in_i} \\
 & \vdots \\
 & \vdots \\
 \text{Group } k: n_k \text{ units treated with Treatment } k: & X_{k1}, X_{k2}, \dots, X_{kn_k}
 \end{array}$$

with $\sum_{j=1}^k n_j = N$, the total sample size for the experiment. Let τ_i be the effect of the i th treatment. Then, we can express the relationship between the treatment and response with a simple linear model:

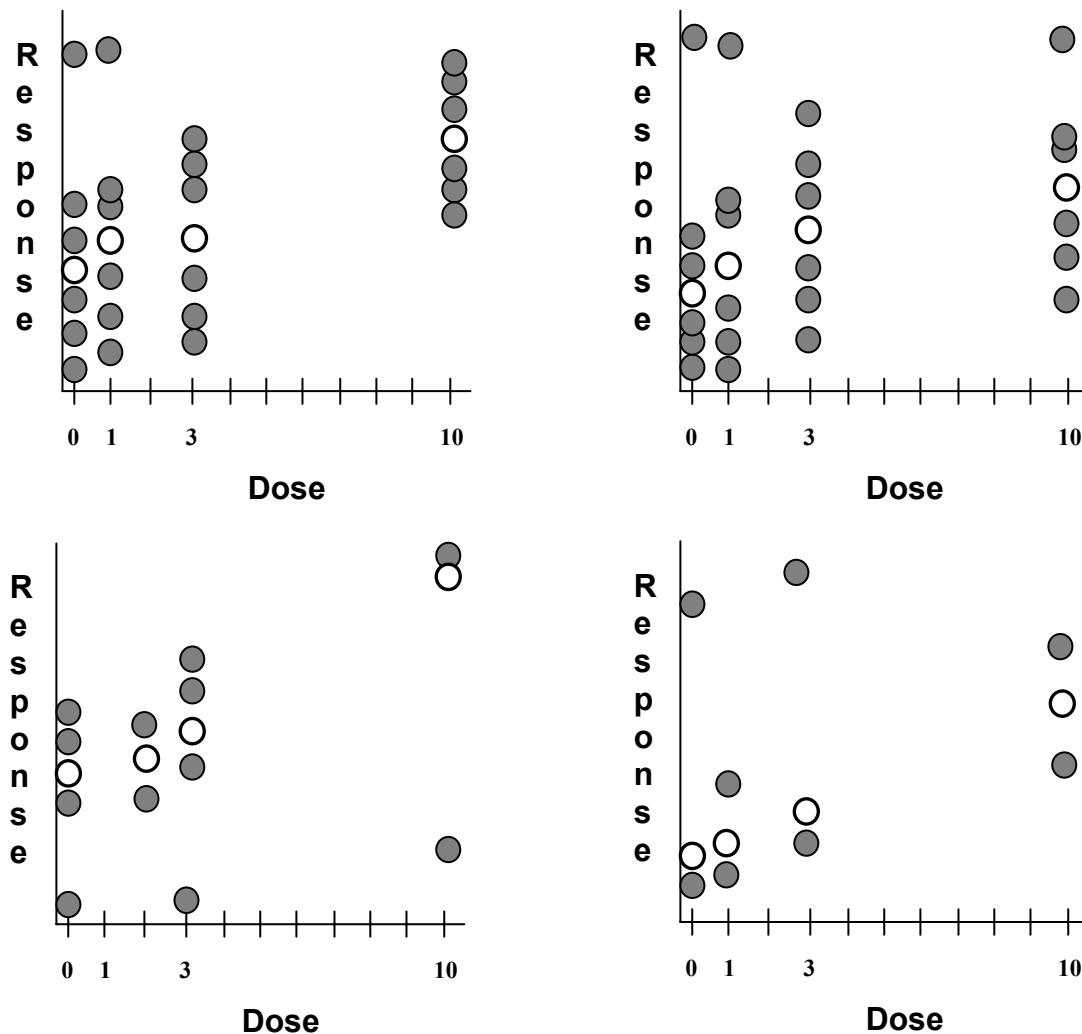
$$X_{ij} = \tau_i + \varepsilon_{ij} \quad (\text{M1})$$

where ε_{ij} represents the measurement error associated with the j th unit receiving the i th treatment. We will assume that the ε_{ij} 's are independent and identically distributed by virtue of the original random assignment of units to treatments.

MONOTONIC RESPONSES – THE JONCKHEERE-TERPSTRA STATISTIC SETTING

Typically, investigators are interested in performing some form of inference on the τ_i 's. It is common in pharmaceutical research for investigators to study a biological response to increasing doses of a single agent within the one-way layout paradigm. Using the notation from before, suppose that the k treatment levels may be ordered by a hypothesized monotonic response pattern. Without loss of generality, suppose that this is a monotonic non-decreasing pattern of response to the agent. Illustration 1 depicts examples of four possible response patterns for $k=4$ treatments, all of which are monotonic non-decreasing.

Illustration 1: Examples of Monotonic Non-Decreasing Response Patterns (Median Response = ○)



Then, we could mathematically express the relationship of the effects of the k treatments as $\tau_1 \leq \tau_2 \leq \tau_3 \leq \dots \leq \tau_i \leq \dots \leq \tau_k$. The desired statistical significance test would be:

$$H_0: \tau_1 = \tau_2 = \dots = \tau_i = \dots = \tau_k \text{ VS} \\ H_a: \tau_1 \leq \tau_2 \leq \tau_3 \leq \dots \leq \tau_i \leq \dots \leq \tau_k \text{ (H1),}$$

with at least one strict inequality existing amongst the k treatments. This test is known as the *Jonckheere-Terpstra* test (Jonckheere, 1954, Terpstra, 1952). How could we test H1?

Intuitively, it seems sensible to test each pair $\binom{k}{2} = \frac{k!}{(k-2)!2!} = \frac{k(k-1)}{2}$ of treatments in a “one-sided”

fashion corresponding to the hypothesized order of the treatments (e.g., assuming that increasing doses of the agent induce monotonically non-decreasing responses). If each of these “pair-wise” statistics is large (demonstrating evidence in favor of the hypothesized order of the treatments), it follows that the sum of the $\frac{k(k-1)}{2}$ statistics should

also be large. This is exactly how the Jonckheere-Terpstra statistic works. For each treatment pair τ_i and τ_l ($1 \leq i < l \leq k$), we calculate a Mann-Whitney statistic (Mann & Whitney, 1947). The test statistic for the Jonckheere-Terpstra test is the sum of all of these pair-wise Mann-Whitney one-sided test statistics.

Mathematically, if we denote the Mann-Whitney test statistic (incorporating an adjustment for ties) comparing τ_i and τ_l by $U_{\tau_i \tau_l}$, the Jonckheere-Terpstra statistic may be written as:

$$J = \sum_{i=1}^{l-1} \sum_{l=2}^k U_{\tau_i \tau_l} \quad (E1)$$

We will reject the null of H_1 , in favor of the alternative at level α , if J is larger than a specific cutoff value, say c_α . If ties exist amongst the responses, the test will only be approximately an α -level test. The cutoff of the Jonckheere-Terpstra test may be based on a large-sample (asymptotic) approximation or the exact distribution of the ranks.

To perform the large-sample approximation, we need to “adjust” the Jonckheere-Terpstra statistic so that we can invoke the Central Limit Theorem and use a Gaussian or normality-based cutoff. The form of the large sample statistic, *without* ties amongst the responses, is:

$$J_{lsa} = \frac{\sum_{i=1}^{l-1} \sum_{l=2}^k U_{\tau_i \tau_l} - \left\{ \frac{N^2 - \sum_{i=1}^k n_i^2}{4} \right\}}{\sqrt{\frac{N^2(2N+3) - \sum_{i=1}^k n_i^2(2n_i+3)}{72}}} \quad (E2)$$

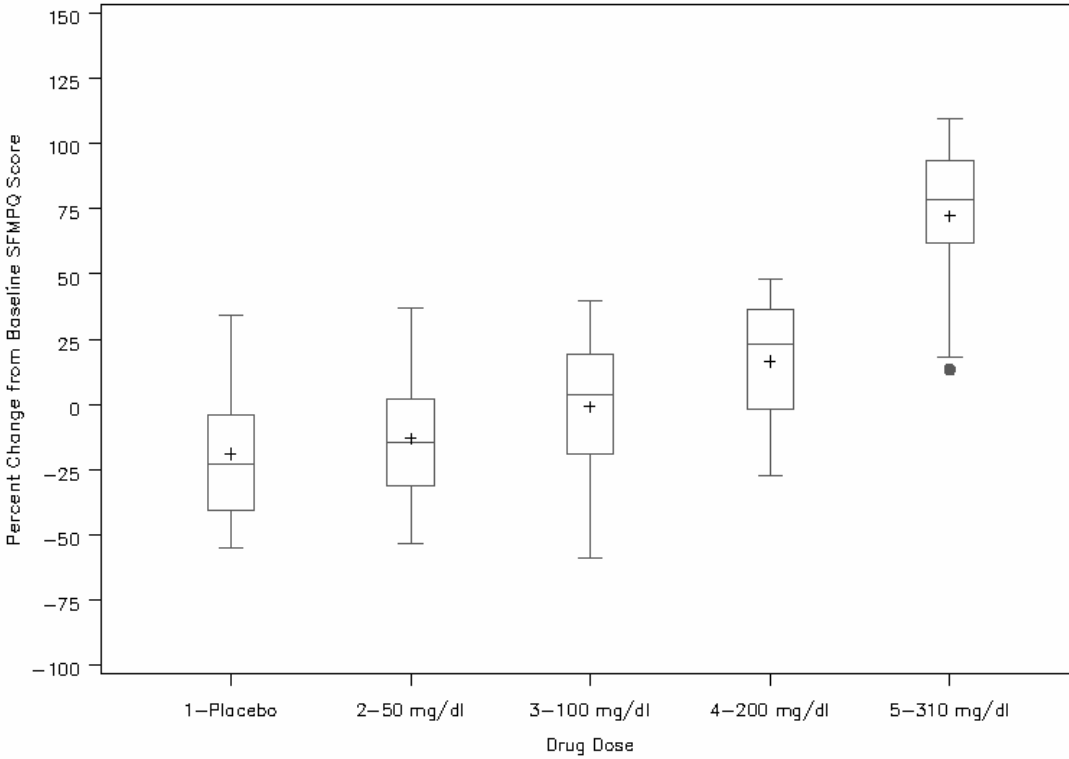
The null hypothesis expressed in H_1 is rejected at level α if $J_{lsa} \geq z_\alpha$, where z_α is the quantile from a standard Gaussian (normal) distribution. If ties exist amongst the responses, the denominator of E2 must be adjusted to accommodate their effect on the variance under the null hypothesis. For the mathematical expression of the variance in the presence of ties, please see a reference like Hollander and Wolfe's *Nonparametric Statistical Methods*, second edition. SAS will perform both the large-sample approximate and exact versions of the Jonckheere-Terpstra test via the FREQ procedure. Unfortunately, at the time of this writing, SAS does not perform such an analysis in JMP. Let's now look at an example from medical research to make the entire discussion “concrete”.

Consider a double-blind, randomized, Phase II clinical trial designed to investigate a potential analgesic. One hundred twenty-five subjects were randomized to one of the following treatment groups with equal probability: placebo, or 50 mg/dl, 100 mg/dl, 200 mg/dl, or 310 mg/dl of the test agent. Subjects received a single dose of treatment and were followed for 3 hours. Each subject's pain response was measured using the *Short-Form McGill Pain Questionnaire* before treatment (at baseline) and afterwards. A percent change from baseline was calculated and used as the primary endpoint. The objective of the study was to establish a dose-response relationship for the potential analgesic agent. The pain research physicians postulated that the response to the treatment would be monotonic by virtue of knowledge of the compound's response in pre-clinical studies and because of its similarity to other compounds in its pain-relief class.

The response data from the trial are summarized below in Illustration 2 via box plots.

Illustration 2: Phase II Dose-Response Analgesia Study

Percent Change from Baseline Short-Form McGill Pain Questionnaire After Treatment with a Novel Analgesic by Dose



Group	Sample Size	Mean (%)	Median (%)	Standard Deviation
Placebo	25	-18.80	-22.98	25.07
50 mg/dl	25	-13.13	-14.57	26.11
100 mg/dl	22	-0.85	3.78	25.47
200 mg/dl	25	16.65	23.38	24.26
310 mg/dl	25	72.35	78.57	29.13

The table of summary statistics below Illustration 2 provides us with some interesting information. Three measurement values were not recorded for three patients in the 100 mg/dl group. Moreover, as the dose of the compound increases, the data tend to be slightly skewed downward. The box plot above the table of summary statistics also provides an indication that at least one extreme value (towards the lower side of the percent change in scores) existed in a group in the study: the “dot” in the plot for the 310 mg/dl group represents a value of 13.47%. Most of the scores for this group are greater than 64%. The presence of the tailing responses to therapy with increasing doses makes a case for the use of nonparametric statistical methods (absence of symmetry).

The clinical investigators in this study are interested in a hypothesis test like H1:

$$H_0: \tau_{\text{placebo}} = \tau_{50 \text{ mg/dl}} = \tau_{100 \text{ mg/dl}} = \tau_{200 \text{ mg/dl}} = \tau_{310 \text{ mg/dl}} \text{ VS}$$

$$H_a: \tau_{\text{placebo}} \leq \tau_{50 \text{ mg/dl}} \leq \tau_{100 \text{ mg/dl}} \leq \tau_{200 \text{ mg/dl}} \leq \tau_{310 \text{ mg/dl}}.$$

To achieve this objective, we can apply the Jonckheere-Terpstra test. The sample size is sufficiently large that the large sample approximation should perform satisfactorily.

PERFORMING THE JONCKHEERE-TERPSTRA TEST IN SAS WITH PROC FREQ

The Jonckheere-Terpstra test is performed via SAS' PROC FREQ. The code to perform the large sample approximate test is:

```
proc freq data=ANALGESIA;  
  tables GROUP*PCIPS/noprint JT;  
run;
```

The procedure call looks like standard PROC FREQ code: the only difference is that the "JT" option needs to be invoked to perform the desired Jonckheere-Terpstra test. Note the order of the variables in the tables statement and the invocation of the "noprint" option. The grouping or classification variable must occur first to perform the appropriate test. Without the "noprint" option, the FREQ PROCEDURE output will include a large table containing columns for every percent change in pain score. The only advantage that the author sees in printing out the table is it affords the user the ability to look for ties in a large data set (the column totals for all columns should be "1" if no ties exist). For our purposes, we will not include this table. The output for the test is shown in Illustration 3.

Illustration 3: Analysis Results for the Phase II Analgesia Study (Jonckheere-Terpstra Statistical Test Output from PROC FREQ)

Percent Change from Baseline Short-Form McGill Pain Questionnaire
After Treatment with a Novel Analgesic by Dose

The FREQ Procedure

Statistics for Table of GROUP by PCIPS

Jonckheere-Terpstra Test	
Statistic	4825.0000
Z	8.3656
One-sided Pr > Z	<.0001
Two-sided Pr > Z	<.0001

Effective Sample Size = 122
Frequency Missing = 3

From the output, we see the statistic's value (4825.0000), the asymptotic statistical test's value ($Z = 8.3656$), the number of missing values (3) and the p -value (<0.0001). From this example, we would conclude that a statistically significant monotonic dose-response relationship exists over the analgesia study (from placebo-dosed through 310 mg/dl-dosed subjects).

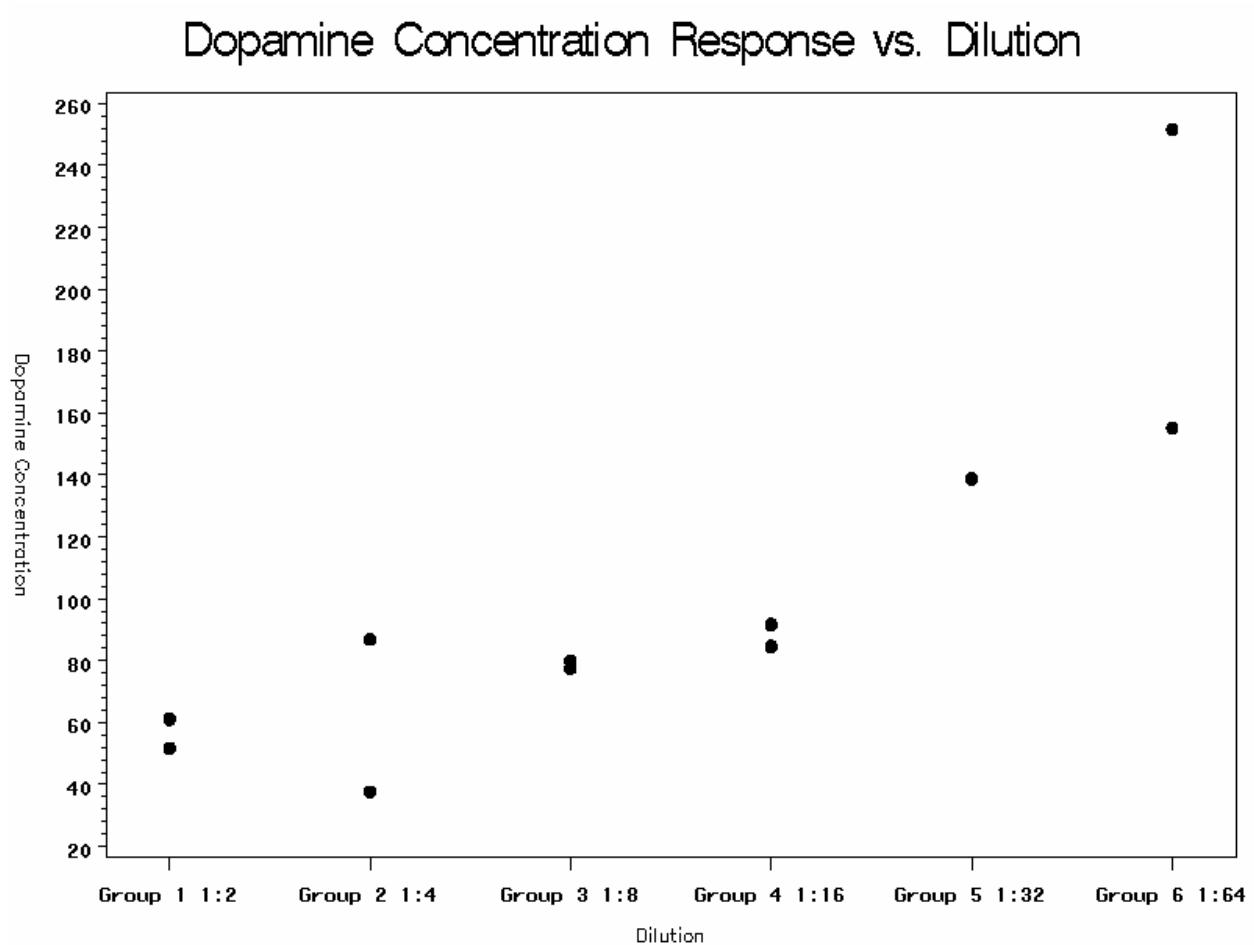
SMALL SAMPLE SIZES AND THE EXACT FORM OF THE JONCKHEERE-TERPSTRA TEST

Let's now look at an example using the exact form of the Jonckheere-Terpstra (JT) test. Suppose that we are consulting with a neuroscientist interested in research studying the effect that exposure to a particular strain of bacteria has on the dopamine concentrations in PC12 cells. She creates a set of serial dilutions of the bacterial solutions ranging from 1:2 to 1:64. She is interested in studying dopamine response as the dilution increases from a ratio of 1:2 to 1:64. She hypothesizes that the response will be monotonic as the dilution increases from 1:2 to 1:64. Once again, we are interested in a statistical significance test like H_1 :

$$H_0: \tau_{1:2} = \tau_{1:4} = \tau_{1:8} = \tau_{1:16} = \tau_{1:32} = \tau_{1:64} \text{ VS} \\ H_a: \tau_{1:2} \leq \tau_{1:4} \leq \tau_{1:8} \leq \tau_{1:16} \leq \tau_{1:32} \leq \tau_{1:64} \text{ (with at least one strict inequality).}$$

Let's now look at the results of the experiment. The dopamine responses are plotted versus the ordered treatments in Illustration 4.

Illustration 4: Scatterplot for the Dopamine Concentration/Dilution Experiment



Upon studying the data, it becomes clear that we are dealing with a very small study. Note that the experimental design consists of at most two replications per dilution. The total sample size for this study is 11 samples. This small sample size warrants the use of an exact test. We will once again use SAS' PROC FREQ procedure to test the monotonicity of the response relative to the specified order. The SAS code to perform the analysis of this example of a small sample setting is shown below:

```
data DA;
  input DILUTION $1-15 DOPAMINE_CONCENTRATION;
  datalines;
Group 6 1:64 155.294
Group 6 1:64 251.765
Group 5 1:32 138.824
Group 4 1:16 91.765
Group 4 1:16 84.706
Group 3 1:8 80.000
Group 3 1:8 77.647
Group 2 1:4 37.647
Group 2 1:4 87.059
Group 1 1:2 61.176
Group 1 1:2 51.765
;
run;
```

```

proc gplot data=DA;
  title "Dopamine Concentration Response vs. Dilution";
  plot DOPAMINE_CONCENTRATION*DILUTION/vaxis=axis1 haxis=axis2;
  symbol v=dot;
  axis1 label=(a=-90 f=simplex "Dopamine Concentration");
  axis2 label=(f=simplex "Dilution") offset=(5,5);
run;
quit;

proc freq data=DA;
  tables DILUTION*DOPAMINE_CONCENTRATION/JT noprint;
  exact JT;
run;

```

The main difference between the procedure call for this analysis and the one in the Phase II Clinical Trial Example presented previously is the insertion of the exact statement specifying the exact form of the Jonckheere-Terpstra statistics ("JT"). The results of the analysis are shown in Illustration 5.

Illustration 5: Output from the Exact JT Test Performed on the Dopamine Concentration Response Dilution Experiment

```

Dopamine Concentration Response vs. Dilution

The FREQ Procedure

Statistics for Table of DILUTION by DOPAMINE_CONCENTRATION

Jonckheere-Terpstra Test

Statistic (JT)          45.0000
Z                      3.1623

Asymptotic Test
One-sided Pr > Z        0.0008
Two-sided Pr > |Z|      0.0016

Exact Test
One-sided Pr >= JT      4.040E-04
Two-sided Pr >= |JT - Mean| 8.081E-04

Sample Size = 11

```

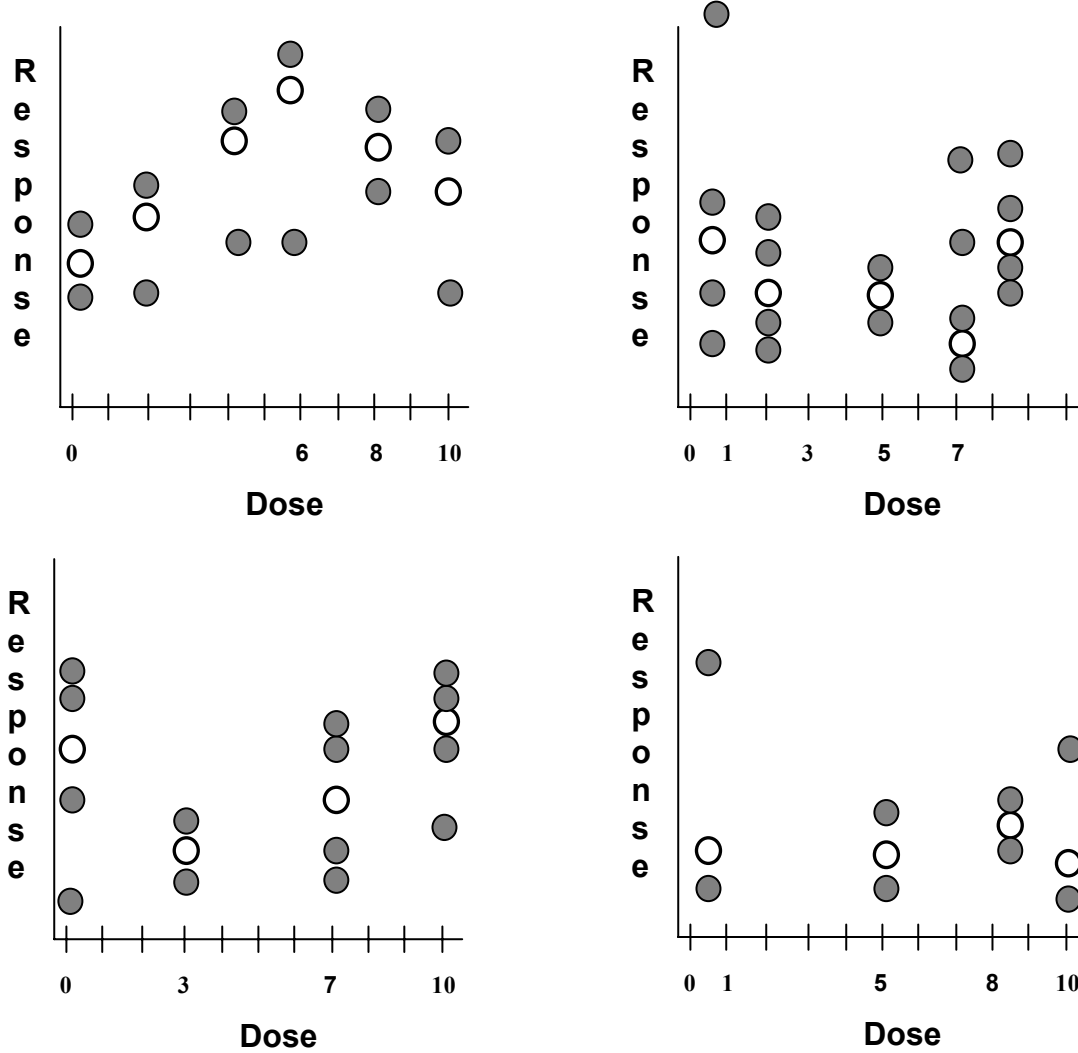
The output above suggests that a statistically significant monotonic trend relationship exists between the dopamine concentration response and the six ordered dilutions ($p \approx 0.0004$). For this analysis, the program ran very quickly, even with the request of the execution of an exact test. This is not always the case. The author has experience using the exact form of the JT test in other settings where the calculations took a few hours to complete. SAS does warn the user of the possibility of a long run-time in the log when s/he uses PROC FREQ to perform an exact JT test.

NON-MONOTONIC ("UMBRELLA") TRENDS

UMBRELLA ALTERNATIVES WITH THE PEAK CONJECTURED, A PRIORI

In the previous section, a nonparametric approach to the analysis of monotonic trend data within the paradigm of a one-way layout was introduced. It is possible that other forms of dose response could exist because of the particular biology under examination. Consider the dose response patterns shown in Illustration 6.

Illustration 6: Examples of Non-Monotonic Response Patterns (Median Response = ○)



Note that for the examples shown in Illustration 6 we cannot express the conjectured (alternative hypothesis) relationship between the treatment levels as we did previously (H1), as the relationship is either $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4, \tau_4 \geq \tau_5 \geq \tau_6$ (upper left panel of Illustration 6), $\tau_1 \geq \tau_2, \tau_2 \approx \tau_3, \tau_5 \geq \tau_3 \geq \tau_4$ (upper right panel), $\tau_1 \geq \tau_2, \tau_2 \leq \tau_3 \leq \tau_4$ (lower left panel) or $\tau_1 \approx \tau_2, \tau_2 \leq \tau_3, \tau_3 \geq \tau_4$ (lower right panel). Among the k treatments, there exists a treatment, say, τ_p , (for $1 < p < k$), which represents the typical largest or smallest median response; i.e., a response (trend) reversal exists amongst the k ordered treatments. Without loss of generality, let's restrict our discussion to situations like those illustrated in the upper left and lower right panels of Illustration 6. In these circumstances, τ_p corresponds to the "peak" median response amongst the k treatments. Conjectured alternatives, where a peak is postulated, are called *umbrella alternatives* (Mack & Wolfe, 1981).

It is important to stress, at this point in our discussion, that the umbrella alternatives are conjectured, *without* examination of the data *a priori*. It is *not a recommended practice* to set up a one-way layout with ordered treatments, conduct an experiment, examine the results, and then formulate the alternative hypothesis of interest, *after the fact*. This latter practice is really a form of "data-snooping" and should be avoided. Later, we will discuss the setting in which knowledge of peak's location is slight to none, requiring an analytical method for the determination of the peak, followed by a test of statistical significance.

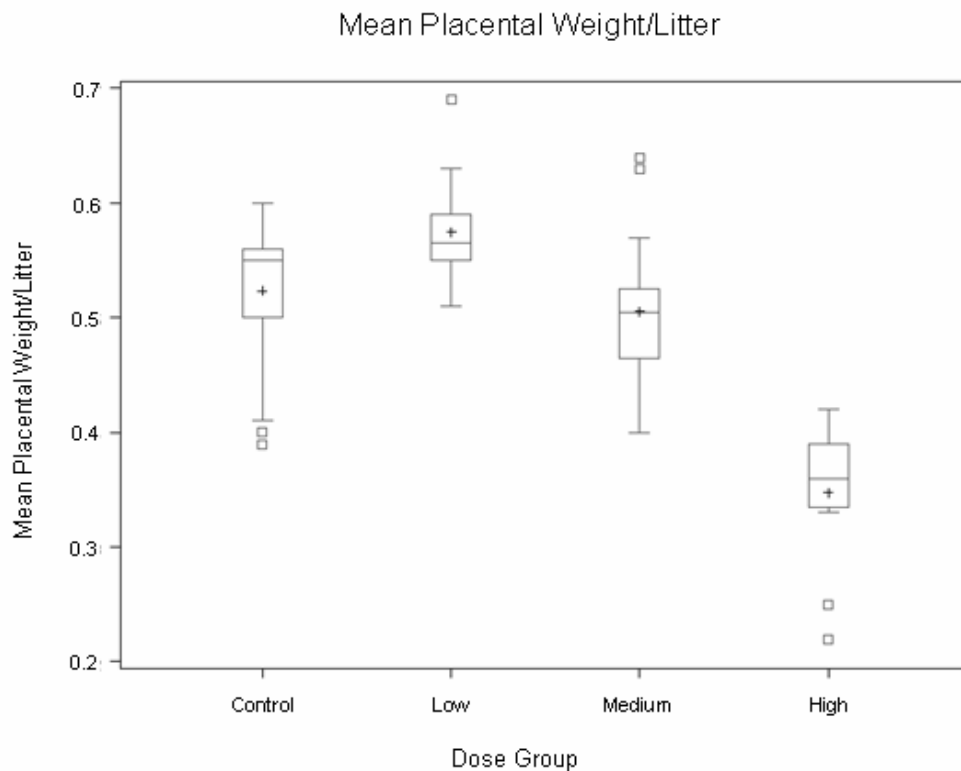
Are umbrella alternatives a nice "theoretical statistics" discussion, or does biology exist that requires their application? What are the settings where biological researchers might wish to test the alternative of a "concave down" (trend reversal) pattern in a set of experimental units' responses to a collection of ordered treatments?

Trend reversals have been reported in drug binding studies (Martin, *et. al.*, 1997), modeling of animal behavior (Rodriguez, *et. al.*, 1994), hypertension studies (Balt, *et. al.*, 2001) and cancer investigations (Teeguarden, *et. al.*, 1999). The field of toxicology is also a very rich area to observe “inverted U-shaped” dose response curves (Calabrese and Baldwin, 2001).

Let us appeal to the rich area of toxicology with the next example. One branch of toxicology that is critical in the development of a new compound is a prenatal developmental toxicity study (EPA 712-C-98-207). In such a study, subjects are randomized to one of several groups and a test substance is administered to gravid animals from implantation until one day prior to the anticipated birth of the offspring. Just prior to the expected delivery date, the pregnant females are terminated and the contents of their uteri examined. One measure of particular interest might be the mean placental weight of all the animals in a litter collected for each dosed female animal.

For a specific case, consider a study that consisted of four groups: one group of subjects received the vehicle alone and the three remaining received one of three increasing doses of a single agent. Moreover, let’s suppose that the researchers conjectured, *a priori*, that the “peak” response would be at the low dose and the decrease for the middle and high dose groups. Illustration 7 shows the results of the study.

Illustration 7: Results from Mean Placental Weight/Litter Study



Note the conjectured “trend reversal” in dose response is supported by the empirical results: the response increases from the control group to the low dose group and then “trails off” or decreases in the medium and high dose groups. The hypothesis of interest would then be:

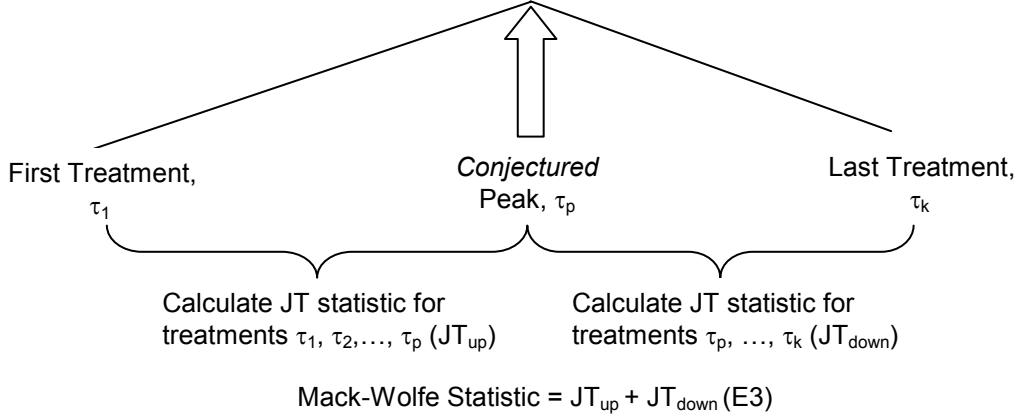
$$H_0: \tau_{\text{Control}} = \tau_{\text{Low}} = \tau_{\text{Medium}} = \tau_{\text{High}} \text{ VS}$$

$$H_a: \tau_{\text{Control}} \leq \tau_{\text{Low}} \text{ \& } \tau_{\text{Low}} \geq \tau_{\text{Medium}} \geq \tau_{\text{High}}, \text{ with at least one strict inequality (H2).}$$

Mack and Wolfe (1981) derived a statistical method to test an alternative as specified above. How does the procedure work?

The procedure is very elegant because of its simplicity: it is simply the sum of a Jonckheere-Terpstra statistic calculated for all treatment levels ordered up to, and including, the conjectured peak and another Jonckheere-Terpstra statistic determined using the conjectured peak and all treatments levels “descending down” from the peak. Illustration 8 shows the determination of the Mack-Wolfe statistic in a schematic diagram.

Illustration 8: Calculating the Mack-Wolfe Statistic



To test a hypothesis like H2, we would reject the null of H2 in favor of the alternative if the Mack-Wolfe Statistic ($M_w(p)$) is greater than or equal to some cutoff:

Reject H_0 : $\tau_{Control} = \tau_{Low} = \tau_{Medium} = \tau_{High}$, at level α , in favor of

H_a : $\tau_{Control} \leq \tau_{Low}$ & $\tau_{Low} \geq \tau_{Medium} \geq \tau_{High}$, with at least one strict inequality if $M_w(p) \geq c_\alpha$.

The cutoff for the statistic, c_α , can be either for an exact or a large sample (asymptotic) test. The exact test is computationally intense and involves determination of all permutations of the statistic under the null. A large sample approximate test can be found by an application of the Central Limit Theorem:

$$M_w(p)_{asymptotic} = \frac{M_w(p) - \text{Expected Value Under } H_0}{\sqrt{\text{Variance Under } H_0}},$$

where: $\text{Expected Value Under } H_0 = \{(\sum_{i=1}^p n_i)^2 + (\sum_{i=p}^k n_i)^2 - \sum_{i=1}^k n_i^2 - n_p^2\} / 4$ and

$$\begin{aligned} \text{Variance Under } H_0 = & \{2[(\sum_{i=1}^p n_i)^3 + (\sum_{i=p}^k n_i)^3] + 3[(\sum_{i=1}^p n_i)^2 + (\sum_{i=p}^k n_i)^2] - \\ & \sum_{i=1}^k n_i^2 (2n_i + 3) - n_p^2 (2n_p + 3) + \\ & 12n_p \sum_{i=1}^p n_i \sum_{i=p}^k n_i - 12n_p^2 N\} / 72 \quad (\text{E4, E5, \& E6}). \end{aligned}$$

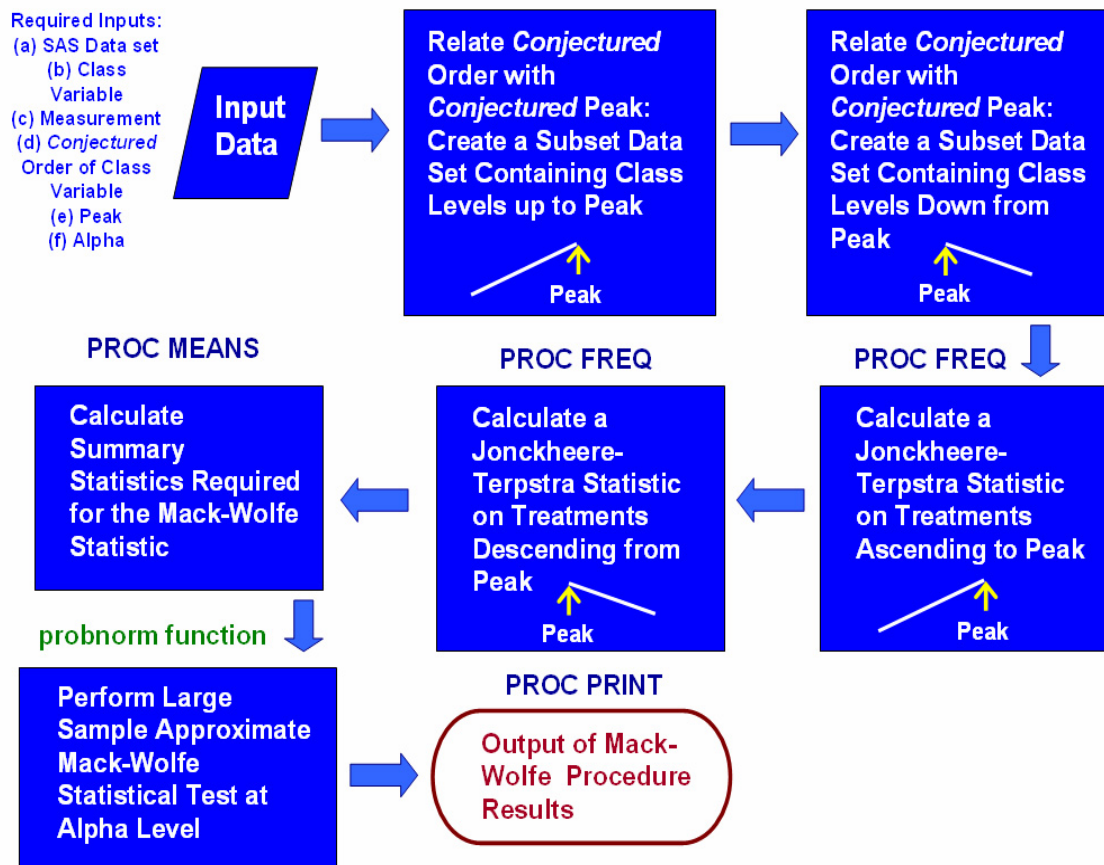
Thus, we reject the null hypothesis of H2 in favor of the alternative if $M_w(p)_{asymptotic} \geq Z_\alpha$, where Z_α represents the quantile from a standard Gaussian (normal) distribution. The large sample approximate version of the Mack-Wolfe test can be carried out in the SAS[®] system using an original macro developed by the author called %MWUSPK.

SUMMARY OF THE %MWUSPK MACRO

The %MWUSPK macro requires the user to input the name of a SAS data set containing a class variable and an analysis variable (the measurement to be tested by the Mack-Wolfe test). The user also needs to include the conjectured order of the treatments, the conjectured peak level of the classification or grouping variable and the level of statistical significance for the test when s/he invokes the macro.

The macro's first function is to look at the input string of treatment levels and determine the number of treatment levels present. The author used a macro found on the *SAS Technical Support Web Page* under FAQ 1617: (<http://support.sas.com/faq/016/FAQ1617.html>) to achieve this objective. Records with missing values are then removed from consideration. The macro then needs to reconcile the conjectured order against the conjectured peak and organize datasets to perform the two necessary Jonckheere-Terpstra statistic tests using PROC FREQ. These results are combined with some summary statistics produced by PROC MEANS to yield the Mack-Wolfe statistic. The large sample (asymptotic) version of the test is performed by determining the p-value with the PROBNORM function and comparing this p-value with the significance level specified by the user. The final results are printed out with PROC PRINT. Illustration 9 summarizes the macro flow for %MWUSPK.

Illustration 9: Flow of the %MWUSPK Macro



```
%MWUSPK(GROUPS, GROUP, MPLACWT, 1Control 2Low 3Medium 4Maximum, 2Low, 0.05);
```

The first parameter of the macro (&DSN) is the name of the SAS data set containing the classification variable (the second parameter, &GROUP) and the response variable (the third parameter, &VAR). The fourth parameter of the macro is the conjectured order of the levels of the classification or grouping variable (&order). The fifth parameter is the "peak" level of the "umbrella alternative": the level of the treatment effect, where responses "ascend to" and "descend from" as shown in Illustration 8. The final parameter of %MWUSPK macro is the nominal level of statistical significance for the test (&ALPHA).

Illustration 10 shows the results of the analysis of the mean placental weight/litter data set.

Illustration 10: Output from %MWUSPK Analysis of the Mean Placental Weight/Litter

Mack-Wolfe Test for Umbrella Alternatives
Testing 4 Groups in the Conjectured Order: 1Control 2Low 3Medium 4Maximum
and Class Level 2Low as the Peak Response
(0.05 Level of Statistical Significance)

Reference Mack, G.A. and Wolfe, D.A., JASA 76: 175-181

	Large Sample	
Mack-Wolfe Statistic	Approximate p-Value	Statistically Significant?
5.617	p < 0.0001	YES

The titles of the output identify the test, the number of groups tested, their conjectured order, the conjectured peak response and the nominal level of statistical significance. The output also contains the value of the Mack-Wolfe statistic, its corresponding large-sample approximation p -value and a statistical significance indicator comparing the resultant p -value with the user-specified level of statistical significance. In this example, we see that the p -value is much smaller than one ten-thousandth; results of this magnitude are summarized with the expression “ $p < 0.0001$ ”. Larger p -value results are printed out rounded to four decimal places.

UMBRELLA ALTERNATIVES WITH THE PEAK UNKNOWN, A PRIORI

As stated previously, it is unwise to conduct an experiment, examine the data visually, pick the “peak” of the umbrella, perform the Mack-Wolfe test, and make an inference. However, in exploratory settings (e.g., discovery biology or toxicology studies), the peak may be unknown, *a priori*. An analytical procedure to estimate the location of the peak response based upon the data followed by an application of the Mack-Wolfe test would be a more reasonable approach to the analysis of data hypothesized to have non-monotonic response.

How does this procedure work? We begin by calculating Mann-Whitney statistics between a group j ($1 \leq j \leq k$) and the remaining $N_{|j} = \sum_{d=1}^k n_d - n_j$ observations. Thus, for k levels of a classification or grouping variable, we would have k Mann-Whitney statistics. Let's denote these statistics as U_j . We then transform each U_j as follows:

$$U_j^T = [U_j - \{1/2 * n_j(N - n_j)\}] / \sqrt{n_j(N - n_j)(N + 1)/12} \quad (E7).$$

The location of the peak response is then found by the j that corresponds to $\max(U_1^T, U_2^T, \dots, U_j^T, \dots, U_k^T)$.

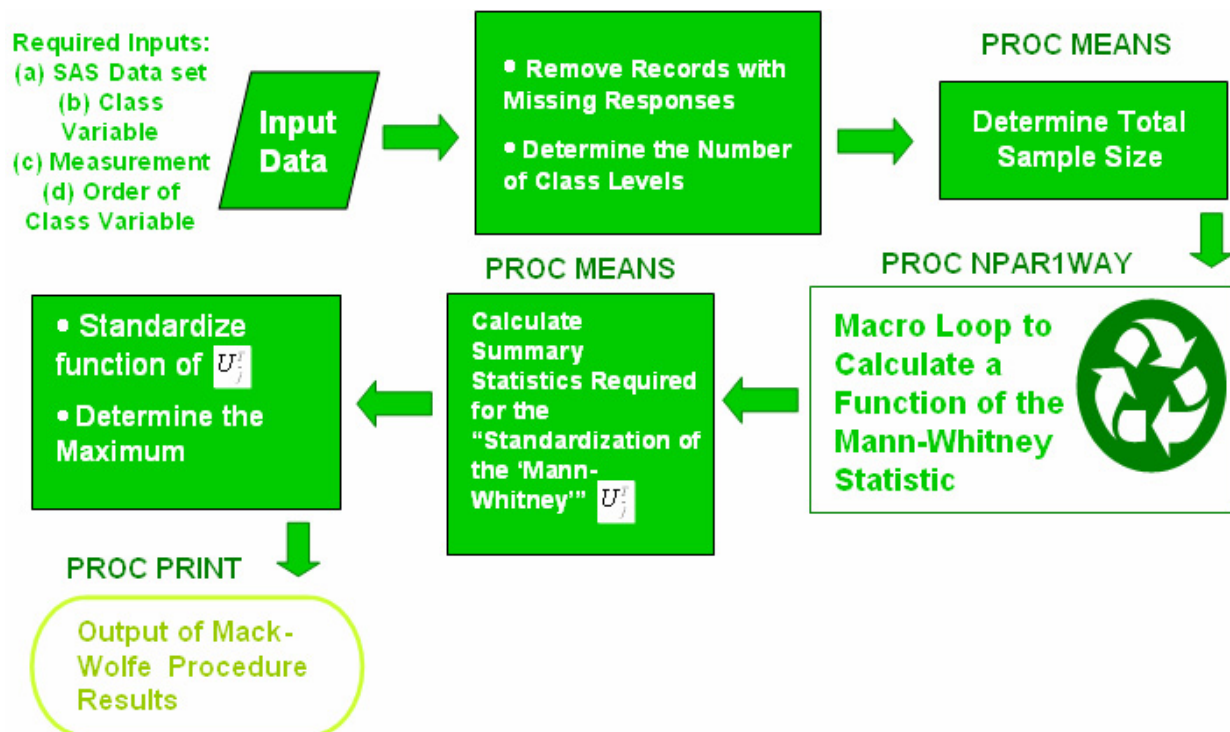
It is possible that $\max(U_1^T, U_2^T, \dots, U_j^T, \dots, U_k^T)$ may not be unique; i.e., it is possible that ξ ties could exist for the k U_j^T statistics. If ties exist, the Mack-Wolfe statistic test, as stated in E3 or E4, must be calculated for the ξ “prospective peaks” and the test becomes reject the null hypothesis in favor of the alternative if $\text{mean}_{\xi}(U_j^T) \geq c_{\alpha}$.

From a data analyst's point of view, what is needed is a SAS program to complement the already existing %MWUSPK macro. The author has developed some code to identify the peak (or “tied for contention” peaks) called %MWUSPU. The analyst needs to run %MWUSPU first, determine the peak, and then run the %MWUSPK code for the number of ties (i.e., ξ times), “average” the statistic values, and compare the mean value with a Gaussian (normal) cutoff.

SUMMARY OF THE %MWUSPU MACRO

Macro %MWUSPU requires the user to input the name of a SAS data set containing a classification or grouping variable and a response. The user must also input an order of the levels of the classification variable. The macro first removes all records containing a missing response value from analysis. The next function of the macro is to determine the number of levels of the grouping variable. Once again, this is accomplished by the same code available from SAS (<http://support.sas.com/faq/016/FAQ1617.html>) as was employed in the macro %MWUSPK. The macro uses PROC MEANS to determine the total number of non-missing values. The macro then performs the required pair-wise tests with PROC NPAR1WAY and the “wilcoxon” option. (It can be shown that the Wilcoxon Rank Sum [Wilcoxon, 1945] and the Mann-Whitney statistic are functions of one another. A nice description of the relationship is found in Hollander and Wolfe’s *Nonparametric Statistical Methods*, second edition, pp. 117-118). The macro then invokes PROC MEANS again for necessary “standardization” summary statistics for each U_j^T , determines the k values of U_j^T and the maximum. The macro then prints out (via PROC PRINT) the group number (order number from the original macro input) corresponding to $\max(U_j^T)$ if the value is unique; if not, it reports all group numbers whose values for U_j^T are tied as $\max(U_j^T)$. Illustration 11 provides a graphical summary of the %MWUSPU macro’s flow.

Illustration 11: Flow of the %MWUSPU Macro



Let’s use the %MWUSPU macro to re-analyze the mean placental weight/litter data.

The call of the %MWUSPU macro is very similar to the original call for the known peak case:

```
%MWUSPU (GROUPS, GROUP, MPLACWT, 1Control 2Low 3Medium 4Maximum);
```

The first parameter of the macro (&DSN) is the name of the SAS data set containing the classification variable (the

second parameter, &GROUP) and the response variable (the third parameter, &VAR). The fourth parameter of the macro is the conjectured order of the levels of the classification or grouping variable (&order). Illustration 12 shows the results of the running the macro on the mean placental weight/litter data set.

Illustration 12: Output from %MWUSPU Analysis of the Mean Placental Weight/Litter

```
Mack-Wolfe Test for Umbrella Alternatives -PEAK UNKNOWN CASE
Testing 4 Groups (User Specified Order: 1Control 2Low 3Medium 4Maximum)
to Estimate the Peak Response Location
```

```
Estimated
Location
of
Peak
```

```
2
```

The macro's output identifies the test and the case ("Peak Unknown"). It also specifies the number of groups and the levels of the grouping or classification variable in the order specified by the user. The output reads "2" for the location of the peak. If we look at the user specified order we would identify the low dose group as the location of the peak (it is the second in the list). Only one value is printed, and thus, the maximum standardized function of the Mann-Whitney statistic is unique. If we had not run macro %MWUSPK (as we did in the first example in this section), we could run it a single time at this point and perform the desired inference directly from the output.

CONCLUSION

A one-way layout is a commonly used experimental design in scientific research, and sometimes a natural ordering exists among the treatments in a study. This paper considers two possible types of ordered alternatives: monotonic or non-monotonic ("umbrella"). The Jonckheere-Terpstra test is a nonparametric, or distribution free, means to test the results of a study where the trend is considered monotonic. PROC FREQ is an easy means to perform either an exact or approximate version of this test in SAS. Examples of non-monotonic responses also exist in the scientific literature suggesting a need to analyze responses that increase to a certain treatment level ("the peak") and then decrease over the remaining levels. Such a need is met by the Mack-Wolfe statistical procedure. The author has demonstrated that the Mack-Wolfe procedure can be programmed in SAS and provides an easy to use macro (%MWUSPK). In some arenas of exploratory research, it may be the case that little is known *a priori* about the response, but the mechanism of action is such that an umbrella alternative is conjectured among the treatment groups. The author has also developed a macro to estimate the location of a (the) conjectured peak response(s) analytically, the %MWUSPU macro. This macro is also easy to use and may be "coupled" with the %MWUSPK macro to perform a complete analysis.

REFERENCES

- Balt, J. C., Mathy, M-J, Pfaffendorf, M., and van Zwieten, P.A. "Inhibition of facilitation of sympathetic neurotransmission and angiotensin II-induced pressor effects in the pithed rat: comparison between valsartan, candesartan, eprosartan and embusartan" *Journal of Hypertension* **19** (12), pp. 2241-2250 (2001).
- Calabrese, E.J. and Baldwin, L.A. "The Frequency of U-Shaped Dose Responses in the Toxicological Literature" *Toxicological Sciences* **62**, pp. 330-338 (2001).
- United States EPA, Prevention, Pesticides and Toxic Substances (7101), "Health Effects Test Guidelines OPPTS 870.3700 Prenatal Developmental Toxicity Study". (EPA 712-C-98-207 August 1998)
- Hollander, M. and Wolfe, D.A. *Nonparametric Statistical Methods*, 2/e. New York: John Wiley & Sons, 1999: pp. 202-234.
- Jonckheere, A.R. "A distribution-free k-sample test against ordered alternatives" *Biometrika* **41**, pp.133-145 (1954).
- Juneau, P.L. "Simultaneous nonparametric inference in a one-way layout using the SAS® System" *Proceedings of the PharmaSUG 2004 Annual Meeting*. Available at: <http://www.lexjansen.com/pharmasug/2004/statistics&pharmacokinetics/sp04.pdf>.

Mack, G.A. and Wolfe, D.A. "K-sample rank tests for umbrella alternatives" *Journal of the American Statistical Association* **76**, pp. 175-181 (1981).

Mann, H.B. and Whitney, D.R. "On a test of whether one of two random variables is stochastically larger than the other" *Annals of Mathematical Statistics* **18**, pp.50-60 (1947).

Martin, C., Berridge, G. Higgins, C.F., and Callaghan, R. "The multi-drug resistance reversal agent SR33557 and modulation of vinca alkaloid binding to P-glycoprotein by an allosteric interaction" *British Journal of Pharmacology* **122**, pp. 765-771 (1997).

Rodriguez, W.A., Horne, C.A., Mondragon, A.N., and Phelps, D.D. "Comparable dose-response functions for the effects of glucose and fructose on memory" *Behavioral Neural Biology* **61** (2), pp. 162-169 (1994).

Teeguarden, J.G., Dragan, Y.P., Singh, J., Vaughn, J., Xu, Y.H., Goldsworthy, T. and Pitot, H.C. "Quantitative analysis of dose- and time-dependent promotion of four phenotypes of altered hepatic foci by 2,3,7,8-tetrachlorodibenzo-p-dioxin in female Sprague-Dawley rats" *Toxicological Sciences* **51**, pp. 211-223 (1999).

Terpstra, T.J. "The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking" *Indagationes Mathematicae* **14**, pp. 327-333 (1952).

Wilcoxon, F. "Individual comparisons by ranking methods" *Biometrics* **1**, pp. 80-83 (1945).

ACKNOWLEDGMENTS

The author would like to thank the expert editorial review of Dr. Dianne Camp of the William Beaumont Research Institute and Drs. Thomas Vidmar and Christy Chuang-Stein of Pfizer Global Research & Development – Michigan Laboratories (Ann Arbor Campus).

RECOMMENDED READING

An excellent "all purpose" reference on nonparametric statistics is Hollander and Wolfe's *Nonparametric Statistical Methods* (Second Edition).

CONTACT INFORMATION

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

Paul Juneau
Pfizer Global Research & Development
2800 Plymouth Road
Ann Arbor, Michigan 48105
Work Phone: 734-622-1791
Fax: 734-622-3153
E-mail: paul.juneau@pfizer.com

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.