

Using Analysis of Variance to Analyze Toxicology Data

Peggy T. Konopacki, Hazleton Laboratories
George W. Pasdirtz, Hazleton Laboratories

Introduction

At Hazleton Laboratories, we have developed a SAS* procedure for analysis of variance (ANOVA) that can identify violations of assumptions yet still produce usable estimates of experimental effects. The goal was to give our toxicologists statistical tools in which they can have some confidence. In this paper, we present the same tutorial (assumptions of analysis of variance, tests for violations of assumptions, and appropriate estimators) that we give to our study directors. Our object is to show how to use the tools on an actual data problem.

Over time, we have worked with two approaches to providing an automated ANOVA procedure. The first approach (Draper and Hunter, 1969) used a series of transformations (log, square root, arcsin, etc.), which continued until either the normal-theory assumptions were met or the procedure failed. In practice, we found that the procedure can fail, often due to outliers (extreme values) in the data set (Thakur, Trutter, and Korte, 1983) and some of the transformations that did work could not be explained scientifically. Instead, we began using another approach, based on the work of Conover and Iman (Conover, 1980; Conover and Iman, 1981; Iman, 1988; Iman and Conover, 1988, 1989), which converts to a nonparametric method based on the ranks of the data when normal-theory is not appropriate.

The procedure first runs the usual ANOVA. Next, a test for normality and equality of variance is performed. If either test fails, each dependent variable is ranked from largest to smallest with PROC RANK. Then, PROC GLM is run on the ranks. In one-way ANOVA, the statistical tests produced by the rank transformation (RT) procedure are known to be monotonically related to their normal-theory counterparts. Monotonicity means that one procedure will be a good approximation to another.

We present justification, competing tests, and code below. We will first explain how to construct an ANOVA model, its assumptions and how we perform a standard analysis at Hazleton. We will also present clinical pathology data on platelet counts from a rodent toxicology study conducted at Hazleton.

ANOVA Specification

We favor a general linear model presentation (the textbook sum of squares notation is equivalent) because it is more easily extended to complicated designs and is used throughout the SAS/STAT documentation. In matrix notation

$$y = XB + e$$

where y is an $n \times 1$ vector of dependent variables, X is an $n \times k$ design matrix, B is a $k \times 1$ vector of parameters to be estimated, and e is an $n \times 1$ vector of unknown errors. Visually,

$$\begin{matrix} y_{11} \\ y_{22} \\ y_{33} \\ \vdots \\ y_{nk} \end{matrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & & & \vdots \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

each y value has an observation number (left subscript) and a treatment group assignment (right subscript). One treatment group is eliminated to obtain a full-rank model (all effects are independent of the grand mean).

The manner in which the X matrix is specified is somewhat flexible. For example,

$$\begin{matrix} y_{11} \\ y_{22} \\ y_{33} \\ \vdots \\ y_{nk} \end{matrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

hypotheses tests can be coded as $[-1,0,1]$ group comparisons or average differences (the mean group difference can be found by multiplying each y value by either $[-1,0,1]$ and summing). The coding above compares each treatment group with a control (group 1). The tests can be generated either directly in the data step (see Freund, 1989) using logical functions

d1=(group=1)-(group=2);
d2=(group=1)-(group=3);

or through the CONTRAST statement on PROC GLM. The data step coding has the advantage of producing independent tests even in situations that are not estimable by CONTRAST.

If the treatment groups receive different dose levels, the dose values themselves could be used. In the example above, the experimental groups might actually have received 100-mg/kg, 150-mg/kg, 200-mg/kg, and 300-mg/kg doses of a drug, which would be coded as

$$\begin{matrix} y_{11} \\ y_{22} \\ y_{33} \\ \vdots \\ y_{nk} \end{matrix} = \begin{bmatrix} 1 & 100 \\ 1 & 150 \\ 1 & 200 \\ \vdots & \vdots \\ 1 & 300 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

and lead to a linear regression trend test.

The SAS code for the ANOVA model (the residuals are saved for testing assumptions) is

```
proc glm ;
  class group;
  model y = group;
  output r = yresid;
```

For the Hazleton data (see Appendix 2), there were significant group differences ($p \leq 0.0064$). Note that the CLASS statement produces the [0,1] design matrix. If it were eliminated, a trend test on the group levels would be generated.

ANOVA Assumptions

The assumptions necessary to perform parametric ANOVA and regression trend tests, however, are somewhat different.

- The model must be properly specified. For ANOVA, the group average (rather than the median or mode) is assumed to be the correct measure of central tendency. For a simple trend test, the relationship must be linear.
- The X variables are measured without error. For the ANOVA model, group assignment is always unambiguous and fixed. For the trend test, the toxicologist must be concerned with whether the animals received the actual doses specified. If not, the assignment of the dose level might be in error. The ranks of the doses, however, can be substituted and are always fixed.
- The variance of the errors is a constant. This assumption can be violated in two ways. First, the experimental groups might not respond similarly to treatment. For example, animals in the high-dose groups might respond more variably to treatment than those in the low-dose groups. Second, if certain factors in a study (animals, doses, facilities, etc.) are not randomly assigned, a separate source of variation not common to all animals could be introduced.
- The errors are uncorrelated. If the same animal is repeatedly observed (for example, for body weight) a correlation pattern could be introduced.
- The errors are normally distributed. The familiar bell-shaped curve might not always be applicable, particularly for some blood chemistry or hematology data.

We need to test these assumptions and shift as necessary to a nonparametric approach, meaning that no inference can be made to a population parameter (for example, true average body weight gain for all animals taking a drug). Instead, we are just testing for observed group differences. Given that animals are not randomly sampled from an entire species for toxicological studies, statistical inference might be a questionable enterprise in any event.

The assumptions we can test are normality and homogeneity of variance. The other ones must be left to the judgment of the study director, who would then report only the nonparametric results if the other assumptions are violated.

Testing for Normality

The test for normality is based on a normal probability (or quantile-quantile) plot (Conover, 1980, Ch. 6). If the ANOVA residuals come from a normal distribution, the empirical quantiles (percentage points of the observed sample distribution) should be highly correlated with similar quantiles from the standard normal distribution.

The probability plot takes the sorted data values on the y-axis plotted against

$$\Phi^{-1}((r_i - 3/8)/(n + 1/4))$$

where r_i is the rank of the data value, n is the sample size, and Φ^{-1} is the inverse of the standard normal distribution function.

The Shapiro-Wilk W statistic (Shapiro and Wilk, 1965) provides a quantitative measure and appropriate statistical distribution for departures from the normal probability plot. The W statistic and the normal probability plot can be obtained in SAS from

```
proc univariate plot normal;
  var yresid;
  run;
```

For the Hazleton data (see the x-y plot in Appendix 2), we see some departure from normality in the tails, enough for the distribution to be judged non-normal by this test ($p < W = 0.0001$).

Alternative empirical distribution function (EDF) statistics (Stephens, 1974; Conover, 1980 Ch. 6) are also available and have comparable power (the ability to detect effects when they actually exist). W, however, is more widely accepted.

Homogeneity of Variance

Levene's test (Levene, 1960) for homogeneity of variance and its extension to the linear model (Draper and Hunter, 1969) are based on the following reasoning.

We can construct a test for different variances by squaring the residuals (the Mean Square Error, *MSE*)

$$\frac{e^2}{n-m} = \frac{(y - XB)^2}{n-m} = MSE = \frac{(y - \bar{y})^2}{n-1} = \sigma^2$$

to obtain a variance that is similar to a group variance. Through simulation studies, however, Levene found that the absolute value of the residuals

$$|e| = |y - XB|$$

was more sensitive to heterogeneous variances and also relatively insensitive to departures from normality.

We obtain Levene's test in SAS by taking the absolute value of the residuals

```
data b;
  set _last_;
```

```
yabs = abs(yresid);
```

```
proc glm data=b;
  class group;
  model yabs = group;
run;
```

and running the same ANOVA model. The Hazleton data are heterogeneous ($p \leq 0.0003$).

The competitor for this test is Bartlett's test (see Winer, 1971: 208-210), which some clients request. However, the test is more sensitive to non-normality and does not generalize to the linear model. Other competitors are described in Conover (1980, 5.3).

Nonparametric Estimation

To estimate an RT-1 model (the RT-2 model ranks within groups and does not have similar properties, see Conover and Iman, 1981), first rank the dependent variable using

```
proc rank data=_last_;
  var y;
  ranks yranks;
```

and then run the standard ANOVA. The justification for using the RT model is that

- Numerical values of test statistics derived from ranks can be algebraically (monotonically) related to test statistics derived from raw scores (Conover and Iman, 1981).

- The exact distribution of ranks is based on permutation distributions (Lehmann, 1975) where

$$\Pr(r) = \frac{N!}{n!}^{-1}$$

indicates samples of N things taken n at a time, where n is the number of experimental groups. This is equivalent to

$$\frac{N!}{n_1!n_2!\dots n_k!} \text{ or } \frac{N!}{(k!)^N}$$

for equal- n control groups. Thus, the probability of a given rank is always known ahead of time given the sample size, a result which does not hold for the normal distribution. If there is no group assignment, the expression simplifies to $1/N!$ (the inverse of N -factorial), which can be computed in SAS using $p = \text{gamma}(1-N)$.

The models are *distribution free* since the distribution of the ranks is always determined by the permutation distribution (except in some complicated cases, see Conover and Iman, 1981) rather than by the actual distribution of the data.

- Tests for many nonparametric procedures can be derived by randomly permuting or shuffling the data (Edgington, 1987) with computer algorithms. The RT model is analogous and easier to compute.

- Permutation distributions become normal (as a result of the central limit theorem) at moderate sample sizes. What is crucial to invoking the central limit theorem, however, is

that the sample estimate is not biased. The RT model is less biased in non-normal samples (Conover and Iman, 1981).

- The loss in power due to ranking is minimal (Lehman, 1975) and the resulting p -values may be very similar ($p \leq 0.005$ as compared with $p \leq 0.0064$ for the parametric model in the Hazleton data).

The competitors to RT models are various nonparametric procedures (the Wilcoxon-Mann-Whitney test, the Kruskal-Wallis test, the Wilcoxon signed ranks test, the Friedman test, and others), which Conover and Iman (1981) show are similar to an RT procedure.

Nonparametric Trend Test

When the grouping variable is coded on an ordinal scale (or reformatted from the actual dose levels), the trend test involves simply removing the CLASS statement from

```
proc glm data=_last_;
  model yranks = group;
```

The nonparametric trend test is preferable in an automated situation because we do not need to investigate departures from linearity (Iman and Conover, 1989). The Terpstra-Jonckheere test (Thakur, 1984) is a competitor, but does not generalize directly to the linear model.

Post Hoc Comparison Among Means

To make comparisons among group means while controlling the experiment-wide error rate (the probability of a jointly significant result), we use the Tukey-Kramer test for all possible comparisons, Dunnett's test for all comparisons against a control, or planned comparisons for more than one control group.

The choice of these techniques was based on the results of a number of simulation studies (Dunnett, 1955, 1964, 1980) that indicated that the two techniques perform best for controlling error rates, effects of unequal n 's, and power.

The SAS code in the PROC GLM step is

```
means group/dunnett alpha=.01;
means group/tukey alpha=.01;
means group/dunnett alpha=.05;
means group/tukey alpha=.05;
```

where we test at both the $p \leq 0.05$ and the $p \leq 0.01$ level. For safety studies, we also include a $p \leq 0.10$ level, to meet agency requirements. The results from the Hazleton data are not presented to conserve space.

Competition for Dunnett's test is data step coding (Freund, 1989), which has more power but is more difficult to apply in an automated program. We use data step coding for multiple control groups or other complicated designs.

Using the SAS Code

The SAS program we run is presented in Appendix 1. The printout is scanned for a significant result from either the

Shapiro-Wilk test or Levene's test. If there are significant departures from normality or heterogeneous variances, the results from the nonparametric ANOVA are used. The nonparametric trend test is reported if called for in the study protocol.

References

- Conover, W. J. (1980), *Practical Nonparametric Statistics* (New York: Wiley).
- _____ and R. L. Iman (1981), "Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics," *The American Statistician*, 35, 124-133.
- _____ (1982), "Some Aspects of the Rank Transformation in Analysis of Variance Problems," *SUGI '82 Proceedings*, 676-680.
- Draper, N. R. and W. G. Hunter (1969), "Transformations: Some Examples Revisited," *Technometrics*, 11, 23-40.
- Dunnnett, C. W., (1955), "A Multiple Comparison Procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association*, 50, 1096-1121.
- _____ (1964), "New Tables for Multiple Comparisons with a Control," *Biometrics*, 20, 482-491.
- _____ (1980), "Pairwise Multiple Comparisons in the Homogeneous Variance, Unequal Sample Size Case," *Journal of the American Statistical Association*, 75, 789-795.
- Edgington, E. S. (1987), *Randomization Tests, Second Edition* (New York: Marcel Dekker).
- Freund, R. J. (1989), "Some Additional Features of Contrasts," *SUGI 14 Proceedings*, 42-50.
- Iman, R. L. (1988), "The Analysis of Complete Blocks Using Methods Based on Ranks," *SUGI 13 Proceedings*, 970-978.
- _____ and W. J. Conover (1989), "Monotone Regression Utilizing Ranks," *SUGI 14 Proceedings*, 1310-1311.
- Lehmann, E. L. (1975), *Nonparametrics: Statistical Methods Based on Ranks* (Oakland, CA.: Holden-Day).
- Levene, H. (1960), "Robust Tests for Equality of Variances," in *Contributions to Probability and Statistics*, (eds.) I. Olkin et al., Ch. 25 (Stanford, CA: Stanford University Press) 278-292.
- Shapiro, S. S. and M. B. Wilk (1965), "An analysis of variance test for normality (complete samples)," *Biometrika*, 52, 591-611.
- Stephens, M. A. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, 69, 730-737.
- Thakur, A. K., J. Trutter, and D. Korte (1983), "Classical Parametric (P) vs. Nonparametric (NP) Significance Testing in Toxicity Studies," *The Toxicologist*, 3.
- Thakur, A. K. (1984), "A FORTRAN program to perform the nonparametric Terpstra-Jonckheere test," *Computer Programs in Biomedicine*, 18, 235-240.

Winer, B. J. (1971), *Statistical Principles in Experimental Design, Second Edition* (New York: McGraw-Hill).

*SAS is a registered trademark of the SAS Institute, Inc., Cary, NC, USA.

Authors

Peggy T. Konopacki
George W. Pasdirtz, Ph.D.
Hazleton Laboratories
3301 Kinsman Boulevard
Madison, Wisconsin 53704

Appendix 1. SAS Code

```
options pagesize=60;
title1 'Study No. xxxx-xxx'; run;
proc sort; by group;
proc print; by group;
proc means; by group; var y; run;
title2 'Parametric ANOVA'; run;
proc glm data=_last_;
  class group; model y = group;
  means group/dunnett alpha=.01;
  means group/tukey alpha=.01;
  means group/dunnett alpha=.05;
  means group/tukey alpha=.05;
  output r = yresid; run;
title2 'Test for Normality'; run;
proc univariate plot normal;
  var yresid; run;
title2 'Test for Homogeneity of
Variance (Levene ANOVA)'; run;
data b;
  set _last_; yabs = abs(yresid);
proc glm data=b;
  class group;
  model y2 yabs = group; run;
title2 'Nonparametric ANOVA'; run;
proc rank data=_last_;
  var y; ranks yrank;
proc glm data=_last_;
  class group;
  model yrank = group;
  means group/dunnett alpha=.01;
  means group/tukey alpha=.01;
  means group/dunnett alpha=.05;
  means group/tukey alpha=.05; run;
title2 'Nonparametric Trend Test'; run;
proc glm data=_last_;
  model yrank = group; run;
```

Appendix 2. Sample Output

Parametric ANOVA

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1182130.741	295532.685	3.83	0.0064
Error	91	7026904.217	77218.728		
Corrected Total	95	8209034.958			
R-Square					
C.V.					
Root MSE					
Y Mean					
0.144004					
33.41022					
277.8826					
831.729167					

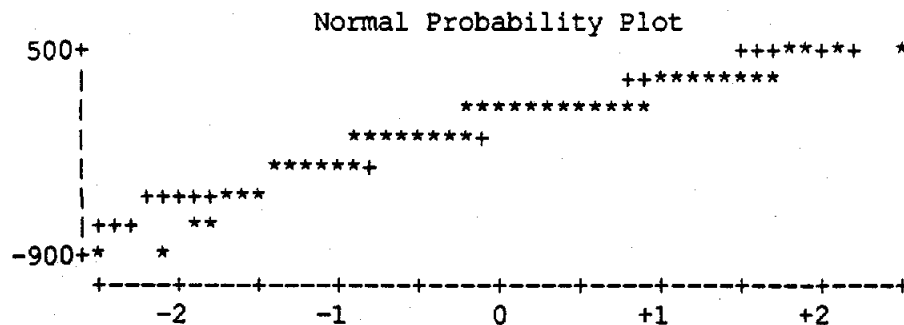
Variable=YRESID

Moments

N	96	Sum Wgts	96
Mean	0	Sum	0
Std Dev	271.9695	Variance	73967.41
Skewness	-0.87534	Kurtosis	1.686311
USS	7026904	CSS	7026904
CV	.	Std Mean	27.75777
T:Mean=0	0	Prob> T	1.0000
Sgn Rank	299	Prob> S	0.2768
Num ^= 0	96		
W:Normal	0.931403	Prob<W	0.0001

UNIVARIATE PROCEDURE

Variable=YRESID



Test for Homogeneity of Variance (Levene ANOVA) 86
13:32 Tuesday, July 2, 1991

General Linear Models Procedure

Dependent Variable: YABS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	714305.9389	178576.4847	5.94	0.0003
Error	91	2736568.1260	30072.1772		
Corrected Total	95	3450874.0649			

R-Square	C.V.	Root MSE	YABS Mean
0.206993	89.84987	173.4133	193.003404

Nonparametric ANOVA 13:32 Tuesday, July 2, 1991
89

General Linear Models Procedure

Dependent Variable: YRANK RANK FOR VARIABLE Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	10992.05752	2748.01438	3.99	0.0050
Error	91	62723.94248	689.27409		
Corrected Total	95	73716.00000			

R-Square	C.V.	Root MSE	YRANK Mean
0.149114	54.13202	26.25403	48.5000000

Nonparametric Test for Trend
General Linear Models Procedure

Dependent Variable: YRANK RANK FOR VARIABLE Y

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
GROUP	-6.67371617	-3.58	0.0005	1.86203122