

ANSWERS TO COMMONLY ASKED STATISTICS QUESTIONS

June 1995 Edition

Mike Stockstill, Technical Support Statistician
SAS Institute Inc.

The purpose of this paper is to address some common "how to" and "what does this mean" users' questions collected by the statistics group in Technical Support. Product discussion includes SAS/IML®, SAS/INSIGHT®, SAS/STAT®, SAS/OR®, SAS/QC®, SAS/ETS®, SAS/LAB®, JMP®, and base SAS® software. The emphasis is on what can be done with the current releases rather than with future releases, although comments are provided where appropriate. Understanding the paper requires little or no technical or statistical expertise. The paper should be most useful to site representatives who face a lot of stat-oriented questions. Topics addressed are alphabetized by procedure as much as possible, and a miscellaneous section follows at the end.

The paper itself is based on our internal notes to each other and is not really intended to stand alone as any sort of documentation. We all contribute, and information is constantly updated. We use this information to help resolve as many calls as possible on initial contact and hope that you can share this with your fellow users.

Please note that this is not intended to be a complete description of any product, procedure or macro. It does not replace any documentation. Instead, browse over it, and use it as a reference tool.

The best way to make use of this is by storing it as an ASCII file on your computer somewhere, and then use your 'find' or 'search' commands to look for text of interest. That is how we use it, and it works well. You may obtain an ASCII version of this paper one of two ways:

- ☐ anonymous ftp: connect to /techsup/download/samples/stat_and_imsi and download stat_qa.txt
- ☐ SIBBS: download the stat file named stat_qa.txt

Occasionally, the answer to a question might make reference to TS###. This refers to a Technical Support Note or TS Note. These are not available as a collection; however, we do provide them individually free of charge on an as-needed basis. Please do not request one unless you have a specific need — that way we can continue to provide them without charge.

Other contributors to this paper include: David Schlottzauer (who originated the file), Eddie Routten, Liz Shaw, Duane Hayes, Phil Gibbs, Annette Sanders, Annie Dudley, Cathy Maahs-Fladung, Donna Woodward, and Kathleen Kiernan.

General, non-PROC-specific or Multi-PROC Questions

Q: Where are the sample library programs?

A: ☐ WITHIN SAS ON UNIX:

!sasroot/samples/<prod>/<filename>.sas, where <prod> is the product name (STAT, IML, etc). Example: %inc '!sasroot/samples/stat/bartlett.sas';

- ☐ WITHIN SAS ON PC: !sasroot/<prod>/sample/<filename>.sas, where <prod> is as above. Example: %inc '!sasroot/stat/sample/bartlett.sas'; Also, under 6.08 and later on PCs, click on HELP: Sample programs to get to them.

☐ WITHIN SAS ON THE VAX:

SASSAMPLES:[<prod>]<filename>.sas, where <prod> is as above. Example: %inc 'sas\$samples:[stat]bartlett.sas';

- ☐ MAINFRAMES: contact your sas site representative to find out the location of the sample libraries. Example: if the sample library is located in 'SAS.SAS608.SAMPLE', use: %inc 'sas.sas608.sample(bartlett)';

Q: Where are the data and programs from the "SAS System for..." books stored locally and how can users get them?

A: The following compressed (*.zip) and uncompressed (*.sas) files are available via anonymous ftp, SIBBS, and the Worldwide Web:

- ☐ forecast.sas—from *SAS System for Forecasting Time Series*
- ☐ forecast.zip linmod.sas—from *SAS System for Linear Models*
- ☐ linmod.zip reg.sas—from *SAS System for Regression, 2nd edition*
- ☐ reg.zip statgdat.sas—from *SAS System for Statistical Graphics*
- ☐ statgdat.zip (the data sets)
- ☐ statgraf.sas (the macro code)
- ☐ statgraf.zip

Compressed files are smaller and faster to download, but will have to be expanded after download. For example, the commands to use on UNIX to download the file forecast.zip are:

- ☐ ftp ftp.sas.com cd techsup/download/samples/bookdata/forecast
get forecast.zip

To download these files from the Web connect to
<http://www.sas.com/techsup/download/sibbs/stat/>

Q: What are and where are the problem alert letters?

A: These are letters that are sent to all SAS Representatives which document problems involving incorrect output or data integrity. They are sent 3-4 times a year, depending on need. Copies of all letters going back to 1991 are on the external Web server, and can be viewed by connecting to: <http://www.unx.sas.com/techsup/download/palerr/>

Q: Is SAS Institute on the World Wide Web (WWW)? How can I access it?

A: Yes. Using a Web browser such as Mosaic, Netscape, MacWeb, etc., the SAS Institute home page is at the URL: <http://www.sas.com/>. You can also use your Web browser to download files from our ftp server and bulletin board (SIBBS) by connecting to: <http://www.sas.com/techsup/download/>

See the README file for information on the various subdirectories and their contents.

Q: How do I access the TS Notes (TS Docs) from the Internet?

A: Only some are available online. They can be downloaded via ftp from the directory: techsup/download/technote. They can also be downloaded from the Web in the directory: <http://www.unx.sas.com/techsup/download/technote/>. Note: Some files are ASCII files, while some are postscript files that will have to be sent to a postscript printer to be usable.

Q: Does SAS do bootstrapping? jackknifing? crossvalidation?

A: These are not specific statistics. Bootstrap and jackknife are methods that can be used for estimating the standard error of any statistic. Crossvalidation can be used to estimate the error rate of a prediction rule. Bootstrap resampling and permutation resampling can be used for adjusting the p-value of a test when assumptions are not met or as a better alternative to Bonferroni adjustment when multiple testing is done. PROC MULTTEST can use bootstrap or permutation resampling to adjust the p-values of the t-test, Cochran-Armitage trend test or Fisher exact test. PROC DISCRIM uses crossvalidation to obtain nearly unbiased estimates of the classification error rates in a discriminant analysis.

Q: Are there any Internet newsgroups available for posting discussions on statistical consulting/analysis?

A: You might post your question to one of these newsgroups:

- ☐ sci.stat.consult—for statistical consulting
- ☐ sci.stat.math—for mathematical statistics questions
- ☐ comp.soft-sys.sas—for SAS questions or to one of these LISTSERV mailing lists via Email (note that these may also be available as newsgroups): SAS-L STAT-L

Q: What does this message mean: "ERROR: Unable to access message ###.###" ?

A: Messages like this can occur with ANY procedure. It means that the procedure was unable to find the message file and indicates an installation problem. The message file contains the text of Notes, Warnings and Error messages as well as the text of table and column headers used in the printed output.

Q: Can I do power analysis / determine the sample size needed ?

A: There is a macro called %POWER that will. It uses PROC GLM's OUTSTAT dataset to compute the power of a statistical effect test or to estimate the sample size required to provide a significant effect test. Since the macro uses the OUTSTAT dataset as the input dataset, only those tests that use the MSE as an error term will be computed appropriately. It may be necessary to do some modifications to the macro in the case of a REPEATED or RANDOM statement. %POWER macro is available via anonymous ftp, SIBBS or the Worldwide Web (WWW):

- ❑ WWW connect to:
<http://www.sas.com/techsup/download/sibbs/stat/power.sas>
- ❑ anonymous ftp: [ftp ftp.sas.com get /users/ftp/techsup/download/sibbs/stat/power.sas](ftp://ftp.sas.com/get/users/ftp/techsup/download/sibbs/stat/power.sas)
- ❑ SIBBS: download power.sas from the STAT area.
[/nfs/TECH2.pc.sas.com/tech2/sibbs/tbbs/download/stat/power.sas](http://nfs/TECH2.pc.sas.com/tech2/sibbs/tbbs/download/stat/power.sas)

To use this macro, see documentation in the header, or see "%Power: A Simple Macro for Power and Sample Size Calculations" by Kristin Latour, *SUGI 17 Proceedings*, 1992, pp.1173-1177.

Q: Is there a way to test for homogeneity of variances for a 1-way ANOVA?

A: One option is to use the sample library program bartlett.sas, to do Bartlett's test. Also, the %HOMOVAR macro performs the O'Brien, Brown-Forsythe, Levine, Bartlett, and Welch Anova F tests for homogeneity of variance. The macro requires base SAS and SAS/STAT software (Release 6.06 or later). %HOMOVAR is available via anonymous ftp, SIBBS or the Worldwide Web (WWW):

- ❑ WWW:
<http://www.sas.com/techsup/download/sibbs/stat/homovar.sas>
- ❑ anonymous ftp:
[ftp ftp.sas.com get /users/ftp/techsup/download/sibbs/stat/homovar.sas](ftp://ftp.sas.com/get/users/ftp/techsup/download/sibbs/stat/homovar.sas)
- ❑ SIBBS: download homovar.sas from the STAT area. To use, see documentation in header or see the *SUGI 17 Proceedings*, 1992 pp 1178-1182.

Q: Does SAS have something to generate values and estimate parameters from a Johnson distribution?

A: A program is available that generates values from the Johnson Su, Sb and Sl systems of distributions. It also contains a macro that chooses the proper system and estimates the parameters of the distribution. It then draws a hi-resolution histogram and overlays the fitted Johnson distribution. However, no test of the model fit is done. SAS/IML and SAS/QC are required (uses PROC CAPABILITY).

The program is available electronically:

- ❑ WWW:
<http://www.sas.com/techsup/download/sibbs/stat/johnsys4.sas>
- ❑ anonymous ftp:
[ftp ftp.sas.com get /users/ftp/techsup/download/sibbs/stat/johnsys4.sas](ftp://ftp.sas.com/get/users/ftp/techsup/download/sibbs/stat/johnsys4.sas)
- ❑ SIBBS: download johnsys4.sas from the STAT area.

QUESTIONS REGARDING SPECIFIC PROCEDURES/MACROS

ADX MACROS (SAS/QC)

Q: The Reference Guide doesn't show examples. Are there any?

A: *SAS/QC Software: Usage & Reference, Version 6, First Edition* provides complete documentation, including introductory examples and advanced examples for Release 6.10. In general, this book can be used for all current releases of SAS/QC.

ADX MENU SYSTEM (SAS/QC)

Q: Where is the ADX MENU SYSTEM documented?

A: Official documentation is now available in the book "SAS/QC Software: ADX Menu System for Design of Experiments" There is also an older yellow booklet titled "ADX Menu System Examples," which is available as TS212.

Q: What are the requirements to use the ADX MENU SYSTEM?

A: SAS/QC and SAS/STAT are required. If SAS/GRAPH* is available, you will be able to produce hi-resolution graphs. If SAS/FSP* is available, you will be able to add response information from within ADX. Otherwise, you will need to write and submit a small DATA step program to add response data. Consequently, SAS/FSP is highly desirable.

Q: If you already have a design, how can you import it for analysis?

A: If the design is one of the types that ADX can create, then you can import it by clicking on File: New design: Add existing data set. Next you fill in the resulting window.

Q: Can you display the aliasing structure of the design?

A: Yes. How you do this depends on whether or not you have saved the design. If you are creating a design, then you can see the aliasing structure by selecting Examine: Parameters: Examine Aliasing in NOPROMPT mode (select the NOPROMPT button if you are using the prompting windows). This is described on page 28 of the ADX menu system book.

If you have already saved the design and entered response data, then in NOPROMPT mode, select the MODEL button to display the "Change Model" window and select "Second order." Select File: End to close the "Change Model" window. Now enter PROMPT mode by selecting Help: Prompt and then select Examine Effect Estimates. This shows any aliasing involving main effects and 2-way interactions. Note that the response data are not used, but they must exist. Consequently, random data will do.

You don't need ADX to get the aliasing structure of a design. Simply run it through PROC REG using anything as response data. If some of the effects listed in the model are aliased, REG prints a note that the model is not full rank and prints the linear dependencies it finds. This report is the aliasing structure. Note that PROC REG cannot tell you if a 3-way interaction is aliased with a main effect unless both the main effect and the 3-way interaction are in the model. So, be sure to list all effects of interest in the model. Since PROC REG does not accept MODEL terms like A*B, it is easier to use PROC GLM with the undocumented option, ALIASING, along with the E option on the MODEL statement. For instance, if you have a 5-factor design, and you want the aliasing structure showing up to 3-way interactions, use the following code:

```
proc glm; model y=x1|x2|x3|x4|x5|3 / aliasing e; run;
```

PROC ANOVA (SAS/STAT)

Q: How can I get multiple comparisons on interaction means?

A: See "Strategies for Performing Multiple Comparisons on Means" by Jenny Kendall, *SUGI 18 Proceedings*, 1993, pages 1283-1289. This is also available as TS282.

PROC ARIMA (SAS/ETS)

Q: Is there any additional documentation?

A: *SAS/ETS Software: Applications Guide 1, Versin 6, First Edition* provides examples of fitting univariate Box-Jenkins models in PROC ARIMA. Examples of estimating and forecasting transfer function and intervention models are also included. For a more theoretical reference, see *SAS System for Forecasting Time Series*.

Q: What do I need to do if I get the warning: more values of input variable are needed?

A: You get this message when you are forecasting either a regression model or a transfer function/intervention model and you do not supply future values of your input variable(s) for your LEAD= observations. You can supply future values by either concatenating them to the bottom

of your DATA= data set, or by fitting a model to your input series with IDENTIFY and ESTIMATE statements. Note: For intervention models (i.e. models with 0,1 dummy variables), you must supply values for your dummy variables in the DATA= data set.

Q: Is there any way to get the parameter estimates in an output data set?
A: As of Release 6.07.03, there are two ESTIMATE statement options that allow you to output the parameter estimates to data sets. While they contain the same information, the data sets are structured differently. Therefore, you can choose the one most appropriate for your application needs. The options are OUTEST= and OUTMODEL=. These options are documented in either *SAS/ETS User's Guide, Version 6, Second Edition*, or in SAS Technical Report P-231, *SAS Software, Summary of Changes and Enhancements, Release 6.07*.

Q: Is there any way to get the fit statistics in an output data set?
A: As of Release 6.07.03, the OUTSTAT= option on the ESTIMATE statement allows you to output the fit statistics such as AIC, SBC, SSE, and the log likelihood. This is documented in *SAS/ETS User's Guide, Second Edition*, or in SAS Technical Report P-231.

PROC AUTOREG (SAS/ETS)

Q: Can it fit ARCH or GARCH models?

A: As of Release 6.08, the AUTOREG procedure may be used to fit the following variations of GARCH models: generalized ARCH (GARCH), integrated GARCH (IGARCH), exponential GARCH (EGARCH) and GARCH-in-mean (GARCH-M). This functionality is documented in *SAS/ETS User's Guide, Second Edition*.

Q: For GARCH models, what distributions are supported for the error term?
A: Currently, only the standard normal distribution is supported. However, we are planning to support other distributions such as Student's t in a future release (most likely Release 6.11). Suggestions for alternative distributions are welcomed.

Q: Is it possible to fit a Cochran-Orcutt model?
A: No. The Cochran-Orcutt method is similar to the default Yule-Walker method for first-order autocorrelation, except that the Yule-Walker method retains information from the first observation. To fit this model, specify the NLAG=1 option on the MODEL statement.

PROC CALIS (SAS/STAT)

Q: Is multi-group analysis supported?

A: Not currently. However it is planned for Version 7.

PROC CANDISC (SAS/STAT)

Q: Which variables are most important (are best discriminators)?

A: PROC CANDISC (or PROC DISCRIM with the CANONICAL option) can tell you how important each variable is via the structure coefficients. The structure coefficients are just the correlations between the original variables and the canonical variables (what some people call 'loadings'). The discriminant function coefficients from PROC DISCRIM are NOT the best way to determine variable importance since correlation among the variables can make their interpretation in that way misleading.

PROC CAPABILITY (SAS/QC)

Q: The Reference Guide doesn't show examples. Are there any?

A: Yes. *SAS/QC Software: Usage & Reference, Version 6, First Edition Volumes 1 and 2* provide complete documentation, including introductory examples and advanced examples for Release 6.10. In general, this book can be used for all current release of SAS/QC.

Q: Can I overlay more than one curve from the same family of distributions onto a histogram?

A: No, not in PROC CAPABILITY. You can use PROC GREPLAY to replay multiple plots on top of each other. This feature is scheduled to be implemented with Version 7.

Q: Can I test for Uniform distribution?

A: Yes. The Uniform distribution is a special case of the Beta

distribution when $\alpha=\beta=1$. Use the options A=1 and B=1 with the BETA distribution option to test for uniformity. Use the THETA= and SIGMA= options to establish the lower and upper limits of support. Set THETA= to the lower limit. Set SIGMA= to the range between the upper and lower limits. For instance, to test that the input data are distributed U(10,25), use the option: BETA(A=1 B=1 THETA=10 SIGMA=15)

PROC CATMOD (SAS/STAT)

Q: Is there any additional documentation?

A: Yes. See the new book, *Categorical Data Analysis Using the SAS System*.

Q: Are there any other recommended references?

A: Agresti (1990) *Categorical Data Analysis*, published by Wiley provides a very good survey and talks about SAS some. Freeman (1987), *Applied Categorical Data Analysis* is also very good, and perhaps a bit lower-level than Agresti. He also shows some SAS code. There are several Sage publications that might be instructional.

Q: What could be causing jobs to keep running out of memory or taking a lot of time?

A: As the number of parameters in the model increases, so does the amount of memory and time needed. The number of parameters in the model is a function of the number of variables and the number of levels in each variable. Generally, CATMOD is not intended to handle ANY variables with many levels.

Q: Why do the parameter estimate signs seem backwards? How can I fix it?

A: PROC CATMOD models the probability of the LOWER response level. For example, if no RESPONSE statement is specified, for a binary response, Y, with levels 0 and 1, PROC CATMOD models a quantity (called a logit) that increases and decreases with $\Pr(Y=0)$. This causes the parameter estimate signs to be reversed from modeling $\Pr(Y=1)$. If you want to model $\Pr(Y=1)$, then do this:

```
PROC SORT; BY DESCENDING Y;
PROC CATMOD ORDER=DATA ....
```

Q: Parameter estimates—which goes with which independent variable level? Why aren't there as many parameters as levels of my independent variable?

A: See Example 1 in the PROC CATMOD chapter. Variable level ordering is given by the default POPULATION PROFILE section or by the ONEWAY option (both shown on page 479). Parameters are on top of page 481. Parameter 2 is the effect of the first level of TRT (a), parm 3 is the effect of the 2nd level of TRT (b), parm 4 is the effect of the 3rd level of TRT (c). Notice that there is no parameter for the last level of TRT (d). That is because of the sum-to-zero constraint of the model. The effect of the last level is constrained to be the negative of the sum of the other parameters, i.e. $-(\text{parm1}+\text{parm2}+\text{parm3})$.

Q: What do I do about infinite parameters or redundant effects messages?

A: These happen when some parameters of the model are detected to be infinite. Parameters can become infinite when there are more parameters in the model than the data can support or when some variable(s) perfectly predict the response. Infinite parameters commonly happen when there are zeros in the table that PROC CATMOD analyzes. To see this table, use the FREQ option on the MODEL statement. By reducing the number of effects on the MODEL statement and/or the number of levels per variable (both of which will reduce the number of parameters in the model), you reduce the size of the table and remove these zeros. Looking at the table should help you find the least disruptive way to remove the most zeros. If any variable on the MODEL statement has many levels (variables in the DIRECT statement, if used, are likely candidates here), try categorizing it to reduce its number of levels.

Q: What does this error message mean, and what can I do to fix it: The response functions are linearly dependent since the number of functions per population, x, is greater than or equal to the number of response

levels, y, in population z (where x, y and z are integers)?

A: It means the data are too sparse for the model specified. One way to fix this is to decrease the number of populations. The number of populations is a function of the number of independent variables appearing on the MODEL statement AND the number of distinct values appearing in each of these variables. The number of populations can be decreased by either dropping independent variables or merging (also called "collapsing") distinct values of one or more independent variables. The obvious place to start is to collapse the values of the variable with the most values. Using the FREQ option on the MODEL statement will print the underlying 2-way table that PROC CATMOD is trying to analyze. Each row is a population and each column is a response level. This problem occurs because of too many zeros appearing in the table. The distribution of zeros in the FREQ table should suggest the best way to collapse or drop variables. Note: Another way to reduce the number of zeros in the table would be to reduce the number of columns by collapsing the response variable values.

PROC CLUSTER (SAS/STAT)

Q: How can I tell how many clusters are in my data set?

A: The best place to look for information regarding the number of clusters is in the "Introduction to Clustering Procedures" chapter in the *SAS/STAT User's Guide Version 6, Fourth Edition, Volumes 1 and 2*. See also the papers by Milligan & Cooper, and Cooper & Milligan referenced there. You can examine several statistics in the printed output to determine the number of clusters if you specify the options PSEUDO and CCC. To interpret:

pseudo T statistic: start at the top of the printed output, and look for the first 'relatively' large value, then move back up one cluster.

pseudo F: look for a 'relatively' large value.

CCC: values greater than 2 or 3 indicate good clusters, values between 0 and 2 indicate potential clusters (but should be taken with caution), large negative values may indicate outliers.

R-Square: look for a value that explains as much variance as you think appropriate. Milligan and Cooper demonstrated that changes in the R-Square are not very useful for estimating the number of clusters, but it may be useful if you are interested solely in data reduction.

Q: How can I tell which observations go in which cluster?

A: First run PROC CLUSTER and create an outtree= data set. Once you have decided how many clusters you want, run PROC TREE on the outtree= data set, specify N= the number of clusters, and OUT= the name of a data set to create. The created data set will have a new variable named CLUSTER, the values of which indicate the cluster member for each observation.

Q: Can I cluster binary data?

A: PROC CLUSTER treats all raw data as continuous. If you want to cluster binary data, create a distance matrix (you may be able to use the %distance macro to do this for releases 6.07.03 and later) as you think it should be defined. Then use this as input to PROC CLUSTER. PROC FASTCLUS will not accept distance data.

Q: How many observations can I cluster?

A: PROC CLUSTER is not practical for very large data sets. See example 3 in PROC CLUSTER—About half way through the example, it shows how to form preliminary clusters with PROC FASTCLUS, how to cluster these preliminary clusters with PROC CLUSTER, and then how to reassign the original observations. How BIG a data set has to be before it is considered BIG depends on how much memory your machine has, how long you are willing to wait, and how much money you are willing to spend on CPU.

Q: I've already clustered one set of data and I want to take a new data set and put the observations in the same clusters. How do I do this?

A: There are two ways. One uses PROC FASTCLUS, the other uses PROC DISCRIM.

- PROC FASTCLUS: First, compute the means from each of your current clusters to use as a SEED= data set in PROC FASTCLUS. Then run FASTCLUS with DATA= your new data set, SEED= the

data set of means from your current clusters, MAXITER=0, REPLACE=NONE, and MAXC= the number of clusters you have. Be sure to use a VAR statement listing the same variables as you used in PROC CLUSTER.

- PROC DISCRIM: Use your data set of cluster-identified observations as the DATA= data set and your data set of new observations as the TESTDATA= data set. List the cluster-identifying variable on the CLASS statement. Use the same VAR statement as you used in PROC CLUSTER. See example 5 in the PROC DISCRIM chapter.

PROC CORRESP (SAS/STAT)

Q: How can I correct the error message: "DIMENS=2 is too large for a R by C contingency table" (for some integers R and C)?

A: Specify DIM=1 on the PROC CORRESP statement. The default value of 2 is too large when there are only 2 rows or columns. See the DIM= option in PROC CORRESP.

Q: Does SAS have a procedure that will do detrended correspondence analysis?

A: No.

Q: The documentation mentions the importance of equal axis scales (so that one inch on the vertical axis covers the same data range as one inch on the horizontal axis). How can this be done? Is there something to make creating correspondence plots easy?

A: Yes. %EQUATE is a simple macro that illustrates how to equate axes. %PLOTIT is a complex macro that creates plots of labeled points typically used to display the results of correspondence analysis, principal components analysis, multidimensional scaling and preference mapping. It equates axes in the process of making such plots. %EQUATE requires SAS/GRAPH. %PLOTIT requires release 6.10 or later and can produce either printer plots or hi-res plots with SAS/GRAPH. Both are available via anonymous ftp, SIBBS or the Web: 6.08 STAT sample library, members equate.sas (for %EQUATE) or crsplot.sas (for %PLOTIT).

- WWW connect to:

<http://www.sas.com/techsup/download/sibbs/stat/> and getting equate.sas (for %EQUATE) or crsplot.sas (for %PLOTIT).

- anonymous ftp: ftp ftp.sas.com get /users/ftp/techsup/download/sibbs/stat/ equate.sas, crsplot.sas
- SIBBS: download equate.sas or crsplot.sas from the STAT area.

To use, see documentation in header. For %PLOTIT, see TS259 which discusses and illustrates its use, or equivalently, refer to Observations 4Q94 (Vol 4., No. 1) pg 23.

PROC CUSUM (SAS/QC)

Q: The Reference Guide doesn't show examples. Are there any?

A: Yes. "SAS/QC Software: Usage & Reference, Version 6, First Edition, Volumes 1 and 2" provides complete documentation, including introductory examples and advanced examples for Release 6.10. In general, this book can be used for all current release of SAS/QC.

PROC DISCRIM (SAS/STAT)

Q: How can I display/output the discriminant function coefficients (linear or quadratic)?

A: If METHOD=NORMAL and POOL=YES (the default, resulting in linear functions) or POOL=NO (resulting in quadratic functions), the coefficients are stored in the OUTSTAT= data set. Also, when the functions are linear, there is a page of printed output labeled 'Linear Discriminant Function.' When METHOD=NP, no functions are used and therefore they cannot be displayed or output. With METHOD=NP, an observation is classified by determining the group for which it has the maximum density (if using R=) or the largest number of the nearest neighbors (if K= is used). See pp 682-683 of *SAS/STAT User's Guide, Fourth Edition, Volume 1*.

Q: Why do I get more discriminant functions than I expected?

A: People often expect that when they have 2 groups (classes) they should get 1 function. For classification purposes, this gets very cumbersome when there are >2 classes. Therefore, DISCRIM ALWAYS creates as many functions as classes. However, if you want only 1

function for 2 classes, try CANDISC. It will give one fewer function than classes, but the functions cannot be easily used for classification. See "how to classify observations" below.

Q: How do I classify observations?

A: Given the discriminant functions, all you do is take the dot product of the observation with each of the functions. Whichever group's function gives the largest result is the group into which the observations is classified. But you don't have to do this yourself unless you don't want to use SAS. Example 5 on pages 760-761 of the *SAS/STAT User's Guide Vol. 1* shows how to use the TESTDATA= option to let the procedure do the work. The general method is also discussed on page 698.

Q: Why must I classify according to the LARGEST value? Aren't the dot product values the distances of the observations to the group means?

A: No. The dot product values are just scores. They are related to the distances, but are different, partly, by a factor of -2. As a result, maximizing these scores minimizes the distances.

Q: Which variables are most important (are the best discriminators)?

A: Use PROC STEPDISC to help select the best discriminating variables. Once a set of variables is selected, PROC CANDISC (or PROC DISCRIM with the CANONICAL option) can tell you how important each variable is via the structure coefficients. The discriminant function coefficients from PROC DISCRIM are NOT the best way to determine variable importance since correlation among the variables can make their interpretation in that way misleading.

Q: Can I use categorical variables?

A: The normal theory method (METHOD=NORMAL, the default) assumes multivariate normality. Using categorical variables violates this assumption rather strongly. Logistic modeling is a better, and simpler, approach. Use PROC LOGISTIC for a 2-group situation or PROC CATMOD for more than 2 groups. If data sparsity causes problems with PROCs CATMOD or LOGISTIC, or if discriminant analysis is required, try the non-parametric methods available with METHOD=NPAR. Then use either K= for the k-nearest neighbor method, or R= for the kernel density method.

DISTANCE macros

There are currently two versions of %DISTANCE macro:
DISTANCE.SAS: 6.09 SAS/STAT sample library; doc: internal
DISTNEW.SAS: post-6.10 (in 6.11 SAS/STAT sample library); doc: TS404 (postscript; not via email)

Q: What are the differences between the two versions of the macro?

A: The DISTANCE macro found in DISTNEW.SAS is newer and recommended. Basic syntax information is in the header, but not enough to really take advantage of the macro. Additional documentation (TS404) is available.

Q: What do the %DISTANCE and %DISTNEW macros do?

A: These macros compute various measures of distance, dissimilarity, or similarity between the observations (rows) of a SAS data set. These proximity measures are stored as a lower triangular (or square, optionally in DISTNEW.SAS) matrix in an output data set that can then be used as input to the CLUSTER, MDS, or MODECLUS procedures. The input data set may contain numeric or character variables or both, depending on which proximity measure is used.

Q: What do the macros require?

A: The XMACRO set of macros and base SAS are required, unless the STD=AGK(p) or STD=L(p) options are used, in which case SAS/STAT is needed. If you want to standardize variables (STD= option), or replace missing values (MISSING=, OPTIONS=REPLACE or OPTIONS=REONLY), then the STDIZE macro is needed too.

Q: How do I get the macros?

A: DISTNEW.SAS (newer and preferred): Note: If you have a postscript printer, you can download and print the distnew.doc file. If you don't, you may request TS404. The code is in: 6.11 (or later) STAT sample library, member distnew.sas and xmacro.sas WWW connect to:

- ❑ <http://www.sas.com/techsup/download/sibbs/stat/> and get distnew.doc, distnew.sas and xmacro.sas (optionally stdize.sas)
- ❑ anonymous ftp:
ftp ftp.sas.com get
/users/ftp/techsup/download/sibbs/stat/distnew.doc get
/users/ftp/techsup/download/sibbs/stat/distnew.sas get
/users/ftp/techsup/download/sibbs/stat/xmacro.sas and (optionally, if want to standardize variables): get
/users/ftp/techsup/download/sibbs/stat/stdize.sas
- ❑ SIBBS: download distnew.doc, distnew.sas, xmacro.sas and stdize.sas from the STAT area. To use %DISTANCE, see the header for basic, syntax or the documentation by printing the postscript file, distnew.doc. Or see TS404.

DISTANCE.SAS or DISTOLD.SAS (older):

STAT sample library, member distance.sas (6.09 or 6.10) or distold.sas (6.11 or later) and xmacro.sas.

- ❑ WWW connect to:
<http://www.sas.com/techsup/download/sibbs/stat/> and getting distance.sas and xmacro.sas (optionally stdize.sas)
- ❑ anonymous ftp: ftp ftp.sas.com get
/users/ftp/techsup/download/sibbs/stat/distance.sas get
/users/ftp/techsup/download/sibbs/stat/xmacro.sas and (optionally, if want to standardize variables): get
/users/ftp/techsup/download/sibbs/stat/stdize.sas
- ❑ SIBBS, download distance.sas, xmacro.sas and stdize.sas from the STAT area. To use %DISTANCE, see documentation in its header.

PROC EXPAND (SAS/ETS)

Q: I need to compute a moving average of a variable on my SAS data set. Is there an easy way to do this?

A: Yes. In Release 6.08 of the SAS System, PROC EXPAND in the SAS/ETS product can be used to make a variety of data transformations. These transformations include: leads, lags, weighted and unweighted moving averages, moving sums and cumulative sums, to name a few. For the complete list of available transformations, see pages 411-412 of the *SAS/ETS User's Guide, Second Edition*. The following example illustrates how to compute a centered, 5-term moving average of a variable X. The resulting moving average is stored as variable MA_X in the OUT= data set. Variable X is also copied to the OUT= data set.

```
proc expand out=ma method=none;
  convert x = ma_x / transform=(movave 5);
run;
```

If you do not have access to the SAS/ETS product, you can compute a moving average in the DATA step. A program in the SAS Sample Library called MAVERAGE illustrates one possible method of doing this in the DATA step. An alternative method is described in *SAS Language and Procedures: Usage 2, Version 6, First Edition*, on pages 223-227. Neither of these methods are as efficient or straightforward as the one described above using PROC EXPAND.

PROC FACTEX (SAS/QC)

Q: The Reference Guide doesn't show examples. Are there any?

A: Yes. "SAS/QC Software: Usage & Reference, Version 6, First Edition" provides complete documentation, including introductory examples and advanced examples for Release 6.10. In general, this book can be used for all current release of SAS/QC.

Q: How can I completely randomize the output plan?

A: Use the RANDOMIZE option on the OUTPUT statement. However, if you are also using either the DESIGNREP= or POINTREP= options, then RANDOMIZE won't do it. After you output the plan, add a random variable and then sort on it. For example:

```
data outplan; set outplan; random=ranuni(123); run;
proc sort; by random; run;
```

PROC FASTCLUS (SAS/STAT)

Q: Is it possible to weight the variables?

A: Yes. Actually, this happens implicitly since variables with larger

variances have larger effects on the clusters (see pg 832 in SAS/STAT User's Guide Vol. 1). To make the variables have equal effects, standardize them so that they have the same variance. To weight them differentially, standardize them so that their variances reflect the weighting that you want. You can use PROC STANDARD with the S= option to do this.

FORECASTING MENU SYSTEM (SAS/ETS)

1st available: exp'tal: 6.07.03. doc.: TS292 (new interface starting with 6.10)

Q: How do I start the Forecasting Menu System?

A: (pre- Release 6.10) In SAS/ASSIST®, select Planning tools, then Forecasting, or type FORECAST on the command line, or via pmenus: Globals->Invoke Application->Forecasting

Q: Where is the Forecasting Menu System documented?

A: Because it is experimental software, it is currently undocumented. A copy of the preliminary documentation is available upon request through Technical Support.

Q: Why do I get the message, "Check ID variable" when I try to specify my input data set?

A: It could be due to several possible reasons: The ID variable is not a SAS date variable. There are multiple occurrences of the same ID value on the data set. (This could happen if you are attempting to do BY-group processing. BY-group processing is not supported at this time.) The ID variable, which is a SAS date variable, does not have a format associated with it. Note: This list is not exhaustive, but it contains the most commonly reported problems that trigger this message.

PROC FREQ (Base SAS and SAS/STAT)

Q: How do I input cell counts instead of raw data?

A: Use the WEIGHT statement to specify a variable containing cell counts. See any of the examples at the end of the PROC FREQ chapter.

Q: How do I test the equality of proportions from two independent samples?

A: The Pearson chi-square test, available using the CHISQ option on the TABLES statement of PROC FREQ will do this. For example, if you want to compare the proportions of males and females responding YES to a question and the proportions are: MALE 30 out of 100; FEMALE 40 out of 100. Arrange the data as a 2x2 table:

	YES	NO
MALE	30	70
FEMALE	40	60

and run in PROC FREQ by entering these cell counts (using WEIGHT statement as explained above). Note that the Pearson test is a test of independence of the row and column variables (gender and response here). But this hypothesis can be shown to be equivalent to the hypothesis of equality of proportions. Also note that the same thing can be done if you have more than two proportions from independent samples.

Q: How do I test the equality of proportions from DEPENDENT samples?

A: Two common situations are:

- Each subject is asked several yes/no questions and the only data available are the proportions responding yes. (If the cell counts of the entire 2x2x...x2 table are available, Cochran's Q can be done as described elsewhere in this file.)
- Each subject is categorized into one level of a multilevel variable and you want to compare the proportions in each level. To test the hypothesis that the proportions are all equal, or take on values that you specify, search on "goodness of fit" in this file.
- The article by Berry and Hurtado in *Observations*, 3Q94 (Vol. 3, No. 4) presents SAS/IML programs that give confidence intervals for the difference in dependent proportions.

Q: How do I get a chi-square test on a one-way (1xC or Rx1) table? or... How do I compute a goodness of fit test for given expected counts?

A: PROC FREQ does not do any tests on one-way tables. If you want to test that the cell frequencies in a one-way table are equal or follow some distribution that you specify, request TS176 that describes how to compute the goodness of fit chi-square test. TS176 shows several ways of doing this. Also, PEARSON.SAS in the SAS/STAT sample library will give chi-square tests of equal cell frequencies.

Q: How do I compute agreement statistics such as McNemar's test, Cochran's Q or kappa statistics?

A: In Release 6.10 or later, the AGREE option on the TABLES statement computes all of these statistics. In earlier releases, page 129 of the *SAS/STAT User's Guide, Version 6, Fourth Edition, Volumes 1 and 2* shows how to compute McNemar's test for matched pairs data using the CMH option. The same data arrangement is needed to get Cochran's Q which is appropriate when there are more than two in a matched set (or more than two repeated measurements). The only difference is that the row variable (the middle variable on the TABLES statement) will have more than two levels and you should use the CMH2 option instead of CMH1. The second CMH statistic (labeled 'Row Means Scores Differ') is Cochran's Q. Before Release 6.10, kappa (and weighted kappa) statistics can be computed with PROC CATMOD, but it is not at all easy. A complex RESPONSE statement must be written for each situation. A SUGI paper (in *SUGI 12 Proceedings* (1987), page 1149, by Robert Terry) discusses how this is done. It is also available as TS118.

Q: How do I interpret kappa (from the AGREE option)?

A: Kappa measures the strength of agreement of the row and column variables, which typically represent the same categorical rating variable as applied by two raters to a set of subjects or items. When there is perfect agreement, all cell counts off the diagonal are zero and kappa is one. Kappa is zero when there is no more agreement than would be expected under independence of the row and column variables. Landis and Koch (*Biometrics*, 1977) give this interpretation of the range of kappa: <=0 Poor; 0-.2 Slight; .2-.4 Fair; .4-.6 Moderate; .6-.8 Substantial; .8-1 Almost perfect.

Q: How do I do "multiple comparisons" or tests on subtables if I get a significant Pearson chi-square?

A: There are no "multiple comparison" methods per se. However, you could use the CONTRAST statement in PROC CATMOD to test for row (or column) differences. Also, correspondence analysis is often used as a way to visualize the non-independence in a table. See the examples in the PROC CORRESP chapter. Or, you can just do additional Pearson chi-square tests on subtables. The WHERE statement is helpful for selecting subsets of rows or columns. Define formats to merge rows or columns. Note: It is possible to partition the likelihood ratio statistic using certain rules of partitioning the table into a set of 2x2 tables. These rules are outlined in Agresti, *Categorical Data Analysis*, 1990, Wiley.

Q: What is the interpretation of the Right- and Left-tailed p-values given with Fisher's exact test for 2x2 tables?

A: If the Left probability value is small, the null hypothesis of equal row probabilities is rejected in favor of the alternative hypothesis that the probability of being in column 1 is less in row 1 than in row 2. Equivalently, the null hypothesis of equal column probabilities is rejected in favor of the alternative that the probability of being in row 1 is less in column 1 than in column 2. If the Right probability value is small, the null hypothesis of equal row probabilities is rejected in favor of the alternative hypothesis that the probability of being in column 1 is greater in row 1 than in row 2. Equivalently, the null hypothesis of equal column probabilities is rejected in favor of the alternative that the probability of being in row 1 is greater in column 1 than in column 2.

FUNCTIONS (Base SAS - Probability, Random number, Quantile)

Q: Is there a value that will return PI?

A: There is no PI() function, but you can get its value with arcos(-1). However, the precision is not unlimited. On an HP700, for example, it is correct to 14 decimal places.

Q: Is there a function that will give the factorial of a number?

A: The GAMMA function can be used to obtain the factorial of a number. For positive integers, GAMMA(X) is (X-1)!. Thus, to find 6! in a DATA step, use an assignment statement such as: SIXFACT = GAMMA(7);

Q: Are there functions for the inverses of COSH, SINH, and TANH (the hyperbolic cosine, sine, and tangent)?

A: No, but they are very easy to obtain using existing SAS functions in the DATA step:

```
inv(cosh): arcosh_x=log(x+sqrt(x**2-1));
inv(sinh): arsinh_x=log(x+sqrt(x**2+1));
inv(tanh): artanh_x=0.5*log((1+x)/(1-x));
```

(where x is the value from COSH, SINH or TANH, respectively). Inverses can be obtained similarly for COTH, SECH and CSCH. See *SAS Communications*, 4th quarter, 1993.

Q: Is there a function that can be used to determine the number of possible combinations of n objects, selected r at a time? This problem is often referred to as combinatorics.

A: Since the number of possible combinations of n objects, selected r at a time, is given by: $n!/(r! * (n-r)!)$, the GAMMA function can be used. (See discussion above regarding factorials). However, this only works when n is relatively small since the computations can overflow rather quickly. The exact limit before the GAMMA function overflows is machine dependent (it is, for example, greater than about 57! — on an IBM mainframe; smaller on a VAX). Another approach would be to use the EXP and LGAMMA functions. In general,

```
A! B!
----- = EXP(LGAMMA(A+1)+LGAMMA(B+1)-
C! D!      LGAMMA(C+1)-LGAMMA(D+1)).
```

Thus, the number of ways of selecting five things, two at a time, is given by: $5!/(2! 3!) = 10$, and can be computed either by:

```
METHOD1 = GAMMA(6)/(GAMMA(3)*GAMMA(4)); or
METHOD2 = EXP(LGAMMA(6)-LGAMMA(3)-LGAMMA(4)).
```

Q: How many random numbers can you generate before it repeats? That is, what is the period of the random number generators?

A: $(2^{**31})-1$. See page 1059 of the *SUGI 12 Proceedings* (1987) for an article on these functions.

PROC GENMOD (SAS/STAT)

1st available: Release 6.09 (experimental in Release 6.08; production in 6.08 4th maintenance = 6.08 TS415); documentation: SAS Technical Report P-243

Q: What are some recommended references?

A: The McCullagh and Nelder reference at the end of the PROC GENMOD chapter is very good. It has many examples. The Dobson reference provides a good, low-level, introductory reference.

Q: Can PROC GENMOD do GAMs (Generalized Additive Models)?

A: No. Nor does anything else in SAS at this time. However, you might want to play with a combination of PROC TRANSREG and PROC GENMOD if you want to do something like GAMs.

Q: Can PROC GENMOD do GEEs (Generalized Estimating Equations)?

A: This capability will be added to PROC GENMOD in a maintenance release to Release 6.11. For now, there is a user-written SAS macro using SAS/IML that will. You can get it via anonymous ftp from statlab.uni-heidelberg.de (or 129.206.113.100). After connecting, change (cd) to the directory pub/statlib/GEE/GEE1 and get the files GEE1_2_0.SAS and GEE1_2_0.DOC. It can also be obtained via Email by sending a message containing only the text GET GEE1 to stat-request@statlab.uni-heidelberg.de. Note: **THIS PROGRAM IS NOT SUPPORTED BY SAS INSTITUTE; IT IS SUPPORTED BY THE AUTHOR**, who is available via Email at UST403@unidozr.hrz.uni-dortmund.de.

Q: Can PROC GENMOD fit data using a negative binomial distribution?

A: Yes, see GENMODX3.SAS in the 6.10 SAS/STAT sample library which shows a user-provided example.

Q: Why do the p-values for a 1 DF effect in the Analysis of Parameter Estimates table and in the LR Statistics for Type 3 Analysis differ so much? Which should I trust?

A: The test in the Analysis of Parameter Estimates table is a Wald test, while the test in the LR Statistics for Type 3 Analysis table is a likelihood-ratio test. Generally, the Wald test is known to be less powerful than the LR test when the effect is large. Under these circumstances, the Wald statistic becomes too small. Generally, the LR statistic is preferred.

%GLMMIX macro

1st available: post-Release 6.08 by request. Release 6.10 SAS/STAT sample library. doc: internal.

Q: What does %GLMMIX do?

A: The macro uses iteratively reweighted likelihoods to fit the model (see Wolfinger and O'Connell, 1993, "Generalized Linear Mixed Models: A Pseudo-Likelihood Approach," *Journal of Statistical Computation and Simulation*, 48, 233-243). By default, %glimmix uses maximum likelihood to find the parameter estimates if there are no random components, and restricted/residual maximum likelihood (REML) if there are. The macro calls PROC MIXED iteratively until convergence, which is decided using the relative deviation of the parameter estimates or the estimated covariance matrix, depending upon if there are no random components. An extra-dispersion scale parameter is estimated by default.

Q: What does %GLMMIX require?

A: Release 6.08 SAS/STAT and SAS/IML.

Q: How do I get %GLMMIX?

A: Release 6.10 (or later) STAT sample library, member glimmix.sas.

- ❑ WWW connect to:
<http://www.sas.com/techsup/download/sibbs/stat/> and get glimmix.sas and glimex.sas.
- ❑ anonymous ftp: ftp ftp.sas.com get /users/ftp/techsup/download/sibbs/stat/glimmix.sas get /users/ftp/techsup/download/sibbs/stat/glimex.sas
- ❑ SIBBS: download glimmix.sas and glimex.sas from the STAT area.

To use, see documentation in header. Examples of use are in glimex.sas (and following the macro definition in the sample library member, glimmix.sas).

PROC GLM (SAS/STAT)

Q: Are there additional references and documentation for PROC GLM?

A: Yes, see "SAS System for Linear Models, Third Edition," the DETAILS section of PROC GLM, and the introductory chapters of the *SAS/STAT User's Guide, Fourth Edition*.

Q: How can I output the parameter estimates to a data set?

A: In release 6.07.03 or later, use PROC MIXED with a MAKE statement selecting the 'SolutionF' table for output. For example, you can output the parameter of the following GLM analysis with MIXED:

```
proc glm; class group; model y=group x*group /
noint solution;
proc mixed; class group; model y=group x*group /
noint solution;
make 'SolutionF' out=parms; ;
```

Prior to 6.07.03, the options you have are either to run PROC PRINTTO or code your own dummy variables and use PROC REG with the OUTEST= option. Which option is better depends on the user, the number of CLASS variables and levels, and the complexity of the model.

Q: How can I get multiple comparisons on interaction means?

A: See "Strategies for Performing Multiple Comparisons on Means" by Jenny Kendall, *SUGI 18 Proceedings*, 1993, pages 1283-1289. This is also available as TS282.

Q: How can I obtain multiple comparisons for my repeated measures?

A: See "Performance of Multiple Comparisons in Repeated Measures Designs Under Nonsphericity" by Steve Thomson *SUGI 15 Proceedings*, 1990, pp. 1365-1370. It is also available as TS235.

Q: My hand computations for the expected mean squares do not match. Why might this happen?

A: See pp. 197-198 of "SAS System for Linear Models, Third Edition." There is disagreement as to the best method for computing expected mean squares. These pages explain the reasoning for choosing the method that SAS uses. Relevant references are given.

Q: My CONTRAST (or ESTIMATE) is non-estimable.

A: See "Using the GLM Procedure and the CONTRAST Statement: A Very Basic Approach" by Donna Fulenwider *SUGI 14 Proceedings*, 1989, pp 152-162. It is also available as TS116. The most common error encountered here is not assigning 'main effect' coefficients, when they are involved in the comparison. There is also a good discussion in "SAS System for Linear Models, Third Edition."

Q: What are some strategies for handling insufficient memory errors, or some kind of system ABEND while running PROC GLM.

A: See SAS Notes 5567 and 6177. These notes give information on how to compute the amount of memory needed. Add about 2.5M for GLM overhead.

Q: When should you use PROC MIXED?

A: Any time you have:

- ☐ a RANDOM statement.
- ☐ a TEST statement.
- ☐ a nested term that is likely a random term on the MODEL statement.
- ☐ an E= option on LSMEANS, MEANS, or CONTRAST statement.
- ☐ the SOLUTION option on the MODEL statement to get parameter estimates which you want in a data set.
- ☐ ESTIMATE statement output that you want in a dataset.
- ☐ a repeated measures problem with a covariate, or missing values on the subjects

(To get started with PROC MIXED, see SAS Technical Reports P-229, *SAS/STAT Software: Changes and Enhancements, Release 6.07*, and P-242, *SAS Software: Changes and Enhancements Release 6.08*. In general, all RANDOM effects will go on the RANDOM statement and NOT on the MODEL statement in PROC MIXED. Use the MAKE statement to write statistics to data sets.)

Q: Why are my parameter estimates biased in GLM?

A: This occurs as a necessary solution to an over-parameterized model. If there are exact linear dependencies (a.k.a. redundancies or exact collinearity) among the parameters in the model, the parameter estimates in PROC GLM will be biased. Note that this is more severe than the high correlations among variables referred to as multicollinearity.

The CLASS statement always induces a linear dependency between the dummy variables it creates for an effect and the intercept. Linear dependencies cause the X'X matrix to be singular (less than full rank), resulting in no unique inverse. A generalized inverse is computed resulting in parameter estimates which are only one of an infinite number of possible solutions. The estimates are regarded as BIASED. The only way to obtain "unbiased" or unique parameter estimates is to: Find the linear dependencies among the parameters and remove them from the model. PROC REG prints a report of linear dependencies and may be easier than using PROC GLM's estimable function. (See *SAS/STAT Vol.2, Fourth Edition*, pg 1416 for an example).

Depending on the complexity of the model, convert from an EFFECTS model to a MEANS model. There are several ways to convert from an EFFECTS model to a MEANS model. Here is one example:

EFFECTS: CLASS T B; MODEL Y=T B T*B;

MEANS: It is necessary to create a new variable in the dataset, for example TRT=10*T+B.

CLASS TRT; MODEL Y=TRT/NOINT;

Note: Including the NOINT option does NOT always guarantee a MEANS model. The necessity of the NOINT option depends on the model specified.

For more detail, see the following references: "SAS System for Linear Models, Third Edition", page 6, "Linear Models" by Shayle Searle, or "Analysis of Messy Data" by Milliken & Johnson.

Q: How is an LSMEAN computed, and how can I output the p-values constructed from the PDIFF table in PROC GLM?

A: There is a paper available from Technical Support that demonstrates this process. It is available upon request.

%ICE macro

Q: What does %ICE do?

A: The %ICE macro computes non-parametric survival curves from interval-censored data. Confidence intervals for the survival curves are also calculated.

Q: What is required for %ICE?

A: SAS/IML, Release 6.08 or later (requires nlp routines). It also requires the utility macros found in the file xmacro.sas.

Q: How do I get %ICE?

A: from Release 6.11 (or later) IML sample library, members ice.sas and xmacro.sas.

- ☐ WWW by connect to:
http://www.sas.com/techsup/download/sibbs/stat/ and get ice.sas and xmacro.sas.
- ☐ anonymous ftp: ftp ftp.sas.com get
/users/ftp/techsup/download/sibbs/stat/ice.sas get
/users/ftp/techsup/download/sibbs/stat/xmacro.sas
- ☐ SIBBS: download ice.sas and xmacro.sas from the STAT area. To use, see documentation in header.

SAS/IML

Q: Can PROC IML be run in batch mode?

A: Yes, it can be run in batch, but it will still execute statement-by-statement.

Q: How large of a matrix can I define in my proc iml job?

A: Below are the Hard-coded Matrix Size Limitations in SAS/IML Software by Platform and SAS Release:

MS-DOS, PC-DOS	6.03, 6.04	4095 elements
UNIX	6.03	32767 rows x 32767 cols
OS/2*	6.06 3	2767 rows x 32767 cols
Windows	6.08	32767 rows x 32767 cols
OS/2	6.08	limited only by memory
UNIX	6.07, 6.09	limited only by memory
Windows NT	6.09	limited only by memory
MVS, CMS, VMS	6.06 and higher	limited only by memory

Starting with 6.10, *ALL* SAS/IML releases will be unlimited (except by available memory).

SAS/INSIGHT

Q: Are there any limitations to SAS/INSIGHT?

A: There must be sufficient memory to hold the data set in memory. In addition to this limitation are the following hard-coded limits:

- ☐ maximum number of variables = machine's largest INT value (usually 32767).
- ☐ the maximum number of observations = machine's largest LONG value (***32767 on Windows 3.1 (a 16-bit system)***; about 1 billion on 32-bit systems like OS/2, Win32s, UNIX)

Increasing available memory can speed INSIGHT up because when memory is tight, it has to do a lot of paging in and out to disk when doing analyses/graphs/etc.

Q: How do I get parameter estimates for a polynomial fit?

A: If you fit a model with X in it and then use the menu selection, CURVES POLYNOMIAL 2 (to get a quadratic fit, for instance), the curve is shown but the parameters are not. To get the parameters (Releases 6.08 and 6.09), select ANALYZE FIT(Y X). In the FIT window, select the Y and X variables as before, then select the X variable in the rightmost window and click on the EXPAND button to add terms up to the order indicated by the value below the EXPAND button. You can change the EXPAND value. Then click on OK or RUN to fit the model. The parameters will be displayed. Starting in release 6.10, there is a "Parameter Estimates" check box in the CURVES POLYNOMIAL dialog box that, if selected, will add a table of the parameter estimates for the requested model.

%INTRACC macro

Q: What is required for %INTRACC?

A: Base SAS and SAS/STAT.

Q: What does %INTRACC do?

A: This macro calculates the six intraclass correlations discussed in Shrout and Fleiss "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, 1979, 86, 420-428. Additionally it calculates two intraclass correlations using formulae from Winer "Statistical Principles in Experimental Design." It also calculates the reliability of the mean of N ratings where N is a parameter of the macro, using the Spearman-Brown prophecy formula. Therefore you can examine the effect that obtaining more raters would have on the reliability of a mean.

Q: How do I get %INTRACC?

A: From Release 6.10 (or later) STAT sample library, member intracc.sas.

- ☐ WWW connect to:
<http://www.sas.com/techsup/download/sibbs/stat/> and get intracc.sas
- ☐ anonymous ftp: ftp ftp.sas.com get
/users/ftp/techsup/download/sibbs/stat/intracc.sas
- ☐ SIBBS: download intracc.sas from the STAT area. To use, see documentation in header.

%ITEM macro

Q: What is required for %ITEM?

A: Base SAS, 6.04 or later. If running Release 6.04 under DOS, you will probably need expanded memory.

Q: What does %ITEM do?

A: The %ITEM macro computes descriptive statistics for analysis of data from a multiple-choice test. Each observation contains the answers from one subject to a set of questions ("items"). The data are compared to an answer key to determine which answers are correct. The score for each subject is computed as the number of correct answers. The output is very similar to that from the ITEM procedure in the SUGI Supplemental library, but several incorrect statistics have been fixed.

Q: How do I get %ITEM?

A: from Release 6.07.03 (or later) STAT sample library, member item.sas.

- ☐ WWW connect to:
<http://www.sas.com/techsup/download/sibbs/stat/> and get item.sas
- ☐ anonymous ftp: ftp ftp.sas.com get
/users/ftp/techsup/download/sibbs/stat/item.sas
- ☐ SIBBS: download item.sas from the STAT area. To use, see documentation in header. Examples follow the macro definition.

JMP

Q: How do I do XXXXX? Can JMP do XXXXX?

A: Check with JMP Help by clicking on the Apple Menu (the little apple symbol on the menu bar), and then selecting "About JMP." The "Stat Guide" button provides an extensive list of the statistical capabilities of JMP and tells how to access each one. Other buttons give information about other aspects of JMP.

Q: Why does my output from JMP differ from that of PROC GLM's output?

A: Differences comparing PROC GLM from SAS to JMP will occur when you have unequal cell sizes and random effects. This is due to the different parameterizations and the two algorithms used to calculate the statistics. JMP's algorithm is computationally more efficient, but GLM's is more general. Both are correct, just different. To answer any comparison questions, point to Appendix A, (pg. 537 *JMP User's Guide*, V2, or pg 536 *JMP Statistics and Graphics Guide*, V3, or pg. 550 *JMP Statistics and Graphics Guide*, V3.1).

Q: Is there a way to rank a column in my data table?

A: Yes. Select Analyze→Distribution of Y→Save Ranks, or Save Ranks Averaged from the 'S' pop-up menu.

Q: Is there a way to standardize a column in my data table?

A: Yes. You can create a formula in the calculator, or more easily go through Analyze→Distribution of Y→Save Standardized.

Q: Can I get discriminant function scores or function coefficients?

A: Yes. See pp. 252-3 of *JMP Statistics and Graphics Guide*, V3, or pp. 255-6 of *JMP Statistics and Graphics Guide*, V3.1 for a general description and detailed formulas. If you do a Save Discriminant Formulas from the 'S' menu, you'll get additional columns in the data table with the Mahalanobis distances and posterior probabilities. The values in the Dist variables are the Mahalanobis distances of each observation to the group centroids and could be called the scores on the discriminant functions. Each of these columns has a formula associated with it. If you look at the formula for the Dist variables, you will see the coefficients of the functions used in JMP (which are not the same as those in SAS).

Q: Can I calculate a geometric mean?

A: Not automatically, but the formula is easily created in the calculator.

Q: Can I test my data for normality?

A: Yes. Go through the 'Distribution of Y' platform. Click 'Test Dist is Normal' from the pop-up menu beside the column's reveal button.

Q: Can I get the kurtosis and skewness?

A: Not automatically, but the formulas are easily created in the calculator.

Q: Is the Mann-Whitney U Test available in JMP?

A: Yes. It is equivalent to the Wilcoxon 2-sample (or Kruskal Wallis k-sample) Test. This is obtained from the 'Fit Y by X' platform, fitting a continuous Y and a nominal or ordinal X, then selecting 'Nonpar-Wilcoxon' from the plot's pop-up menu.

Q: Can I perform a multinomial logit analysis in JMP?

A: No.

Q: Is there a way to get standardized regression coefficients?

A: Yes. You must standardize your variables prior to running the regression. See above for methods of standardizing columns.

Q: My Nonlinear model isn't converging. Why?

A: JMP is very sensitive to initial estimates. If these values are not good, the model may not converge. Also, your data may not support the number of parameters you are trying to estimate.

Q: Can I get a Receiver Operating Characteristic (ROC) curve in JMP?

A: No.

Q: Can I get the process capability index CPK?

A: Yes. From the 'Distribution of Y' output, select 'Set Spec. Limits...' from the pop-up menu and fill in the dialog.

Q: How do I interpret the saved spline coefficients table?

A: The coefficients are the A(intercept), B(linear), C(quadratic), and D(cubic) coefficients for the variable (x-O), where O is the variable for the knot at the beginning of the interval.

Q: I want the medians for by-groups. Can I get them?

A: Yes. This is easily obtained in the calculator, by incorporating nested 'if' statements and by using the 'quantile' function.

Q: I have nominal and ordinal variables, and I don't understand how JMP is performing the stepwise regression analysis. Is there any information on this?

A: See pages 207-208 in the Version 3 documentation or pages 209-210 in the Version 3.1 documentation.

Q: How do I perform a paired t-test?

A: Create two continuous columns in the data table with the pre/post data. Go Through 'Fit Y by X' platform, choose one column as the Y, and the other as the X. Then, from the Fitting pop-up menu, select 'Paired t-Test.'

Q: Can I get Fisher's Exact test?

A: Yes, but for 2-by-2 tables only. Fit Y by X, with nominal Y and X.

Q: Can I get a pdf plot?

A: Yes, if you know the parameters of the pdf. This can be done in the calculator, and there are two examples in the SAMPLE DATA that demonstrate how this is done: 'Normal Density Compare' and 'Normal to Gamma Compare'.

Q: If I change my columns from ordinal to nominal, or vice versa, my parameter estimates do not match. Why?

A: This is due to the different parameterizations JMP uses for the two column types. See Appendix A in the *JMP Statistics and Graphics Guide*, V3 or V3.1 for more information on the exact parameterizations.

Q: I want to test for equal variances in my groups. Can I?

A: Yes, through the 'Fit Y by X' platform. Select continuous Y, and nominal or ordinal X, and choose 'Unequal Variances' in the Analysis pop-up menu. You will get the following tests: O'Brien(.5), Brown-Forsythe, Levene and Bartlett.

Q: Can I get Geisser-Greenhouse and Huynh-Feldt adjusted F-tests?

A: Yes, within the MANOVA fitting personality in 'Fit Model'. Select multiple Y's, check 'Univariate Tests Also' and click 'Repeated Measures.'

SAS/LAB

Q: What is required?

A: Base SAS and SAS/GRAPH, release 6.08 or higher. SAS/FSP is optional, but highly recommended for data entry.

PROC LIFEREG (SAS/STAT)

Q: How do I get the shape and scale parameters for a Weibull distribution from LIFEREG?

A: PROC LIFEREG will print out a scale and an intercept parameter, but these are from an accelerated failure time model (see 'introduction' in the LIFEREG documentation for details). The intercept is called mu, and the scale is called sigma on the LIFEREG output. These parameters may be transformed to produce the Weibull parameters alpha and gamma, as shown in the 'distributions allowed' section. $\Gamma = 1/\sigma$ and $\alpha = \exp(-\mu/\sigma)$. Note that different authors use different forms of the probability density function. Therefore, you might have to do a bit of algebra to make $g(t)$ look like the pdf with which you are familiar.

PROC LIFETEST (SAS/STAT)

Q: How can I output (insert any statistic here) to a data set?

A: If it can't be output in the OUTTEST= or OUTSURV= data sets, there is an experimental version of LIFETEST available that can use the output delivery system (ODS) in Release 6.08 on VMS, OS/2, and Windows.

Q: What is the formula for the confidence interval for the median (or other quartiles)?

A: See SAS Note 7000: Beginning in release 6.08, there is a new option in PROC LIFETEST which will request a confidence interval for each quartile. It is documented in SAS Technical Report P-242, *SAS Software Changes and Enhancements, Release 6.08*, on page 113, and is referred to as the ALPHAQT= option. The computational formula used can be

found in the following reference: R. Brookmeyer and J. Crowley (1982) "A confidence interval for the median survival time," *Biometrics* 38, 29-41. See page 31 of the reference for the actual formula.

PROC LOGISTIC (SAS/STAT)

Q: Are there any recommended references or additional documentation for PROC LOGISTIC?

A: The Hosmer and Lemeshow reference at the end of the LOGISTIC chapter is probably the best and is fairly readable. A book of examples is available from Book Sales: "Logistic Regression Examples Using the SAS System, Version 6, First Ed." Data sets and code from this book are available for downloading. See the instructions on the inside front cover. They are also available in the STAT area of SIBBS in the file logistex.sas. Also available from Book Sales is *Categorical Data Analysis Using the SAS System*, TS274. "Some issues in using PROC LOGISTIC for binary logistic regression," is also useful.

Q: I am getting backwards parameter estimate signs — Why? How can I fix it?

A: PROC LOGISTIC models the probability of the LOWER response level. For example, for a binary response, Y, with levels 0 and 1, LOGISTIC models a quantity (called a logit) that increases and decreases with $\Pr(Y=0)$. This causes the parameter estimate signs to be reversed from modeling $\Pr(Y=1)$. If you want to model $\Pr(Y=1)$, then use the DESCENDING option on the PROC LOGISTIC statement (Release 6.07.03 and later), or do:

```
PROC SORT; BY DESCENDING Y;
PROC LOGISTIC ORDER=DATA ...
```

Q: Odds ratios—How do I get them? How are they computed/interpreted? How can I get confidence intervals for them?

A: Odds ratios are automatically printed starting in Release 6.07.03. Confidence intervals can be obtained using the RISKLIMITS option. See SAS Technical Report P-229, *SAS/STAT Software: Changes and Enhancements, Release 6.07*, to see how they are computed.

Q: Predicted probabilities—How do I get them? How are they computed/interpreted? How can I get confidence intervals for them?

A: Predicted probabilities, for each observation, can be output to a data set using the OUTPUT statement (OUTPUT OUT=<dsn> P=<name>). Upper and lower confidence limits can also be added to the output data set (U= and L= on OUTPUT statement). They are computed as shown on page 1091 of *STAT/STAT User's Guide, Fourth Edition, Volume 2*.

Q: Why can't I reproduce CTABLE results using the predicted probabilities in the OUTPUT data set?

A: CTABLE uses the bias-adjusted predicted probabilities as described on page 1092 of *SAS/STAT User's Guide, Fourth Edition, Volume 2* under "Calculation Method" while the predicted probabilities in the OUTPUT data set are not adjusted.

Q: Stepwise (or forward or backward) selection — Why is it taking so long?

A: Using the SELECTION= option can take a *LOT* of time. The time needed depends on the number of variables and the entry and removal criteria (SLE= and SLS=). As you add variables, increase the SLE= setting, and decrease the SLS= setting, you tend to increase the number of steps in the selection process; this requires more time. Time also depends on the data itself. Two data sets with the same number of variables and observations, using the same entry and removal criteria, can require vastly different numbers of steps in the selection process. For example one data set could have only one significant variable, requiring one step; the other could have many significant variables that are correlated requiring many entry and removal steps. Also, if the size of the input data set is too large to be held in memory, the amount of time will be *MUCH* greater than if it can be held in memory. If there are a large number of observations, try doing the model selection on a random subset of the data. Then, verify the model on the remaining data. If there are a large number of variables (say, over 30), try running separate, 1-variable models for each to eliminate those that have no association with the response. Then, do the model selection on those that remain. Also,

try to remove any variable that is highly correlated with any other variables since it will not contribute anything new to the model.

Q: How can I classify/score new observations?

A: Add the new observations to the original data set, setting the response to missing for all of the new observations. PROC LOGISTIC ignores all observations with missing response values (so the fit will be identical to using just the original data set). Then use the OUTPUT statement to request the OUT= data set and predicted probabilities (P=varname). The OUT= data set will contain predicted probabilities for all observations, including the new ones. PROC SCORE will NOT do the scoring.

Q: What is a "multinomial logit" model (or "discrete choice" or "McFadden's conditional logit" model) and how do I fit it?

A: These are models for data in which the response is usually a set of choices and therefore, is nominally- scaled. Also, at least some of the independent variables in these models indicate characteristics of the choices (cost, size, attractiveness) instead of characteristics of the subject or chooser, as is usually the case. The effect of an independent variable is conditional on the subject choosing between two alternatives, and it depends on the distance between the variable's values that were assigned by the subject to the two alternatives. Such models can ONLY be fit with PROC PHREG. The book "Logistic Regression Examples Using the SAS System, Version 6, First Ed" shows examples of how to do it (alternatively, request TS273). Note that in models that can be fit with PROC CATMOD and PROC LOGISTIC, the independent variables are always characteristics of the chooser, not of the choices. In this case, PROC CATMOD can be used for a nominally- or ordinally- scaled response. PROC LOGISTIC can be used for an ordinally- scaled response and is preferable to PROC CATMOD if any of the independent variables are continuous.

Q: I am getting messages about NON-CONVERGENCE or SEPARATION (complete or quasicomplete). What do they mean? How can I fix it?

A: This is a common problem. It is caused by some parameters of the model becoming theoretically infinite. This can happen when the model perfectly predicts the response or if there are more parameters in the model than can be estimated by sparse data where "sparse" means that there are few or no repeats of each setting of the covariates. PROC LOGISTIC does not converge unless the largest change among the parameters is small; this will never happen when a parameter is infinite. To fix this, reduce the number of variables and/or change continuous variables to categorical. There is no way to know exactly which variable to eliminate/categorize — it's a trial and error sort of thing. Or, instead of reducing or altering the variables, you could use an alternative rule to decide when to stop the iterations. A way to do this is presented in the tutorial on logistic regression by Ying So in the *SUGI 18 Proceedings* (1993). Alternatively, refer to TS450 - available by fax and ftp (as a postscript file).

Q: What is the difference between a WEIGHT and a FREQ statement? Why do all of my effects become significant if I include a WEIGHT statement?

A: A WEIGHT statement should be used when you want to unequally weight the contributions of observations to the likelihood function which is maximized by the procedure. Weights are often used to adjust for the use of stratified, rather than random, sampling in a study (see below). The WEIGHT statement does not change the value of N (sample size) used in formulas for some statistics (listed in the WEIGHT statement documentation). Use the FREQ statement to enter summarized data in which only the unique combinations of independent variable settings are stored, along with a variable indicating the number of times each occurred in the raw data. This method is exactly like entering just the cell counts of a contingency table as opposed to raw data. The values MUST be integers; weights may be fractional. The FREQ statement DOES affect N. Using a given integer-valued variable as a WEIGHT variable, or as a FREQ variable, results in the same parameter estimates and standard errors. However, inflating the values of a FREQ or WEIGHT variable (such as multiplying them by a constant) will drive all effects toward significance. For this reason, weights are typically

'normalized' so that they sum to the actual sample size. You can request that the procedure normalize your weights by using the undocumented NORMWT option on the PROC statement.

Q: If I draw samples from each response level (i.e. stratified sampling as in a retrospective study), do I have to weight my data to make the results meaningful?

A: The odds ratio and its confidence limits are unaffected by the sampling scheme, so no weighting is needed to report them. Except for the intercept, the parameter estimates and their standard errors are also OK. However, since the intercept is affected, any computation based on the full set of parameters is incorrect, such as the predicted event probabilities, differences or ratios of event probabilities (the ratio is called the relative risk), and false positive and negative rates. If the event probabilities are very small in each group to be compared, then the odds ratio provides an estimate of the relative risk. See Agresti's book, "Categorical Data Analysis," Wiley, 1990, pages 8-18 for more information on the relative risk. For more information on handling data from retrospective studies, see chapter 6 of the Hosmer and Lemeshow reference or chapter 7 of D. Collett's book, "Modelling Binary Data" (1991), Chapman & Hall.

PROC LP (SAS/OR)

Q: What size problem will PROC LP handle? How many variables and constraints can I have in my problem?

A: There is no definite answer to this question. The size of the problem is very data dependent. We have run 20,000 variable by 20,000 constraint problems successfully, but have had difficulty solving much smaller problems. If there are integer or binary variables, the problem becomes much more difficult to solve. A large integer problem is on the order of 2000 variables.

PROC MACONTROL (SAS/QC)

Q: The Reference Guide doesn't show examples. Are there any?

A: Yes. "SAS/QC Software: Usage & Reference, Version 6, First Edition" provides complete documentation, including introductory examples and advanced examples for Release 6.10. In general, this book can be used for all current releases of SAS/QC.

PROC MIXED (SAS/STAT)

Q: Are there additional references and documentation?

A: See TS260: "A Tutorial on Mixed Models," Technical Report P-229 (the primary documentation), or Technical Report P-242, which contains changes and enhancements. A general reference on mixed models would be the Southern Cooperative Series Bulletin 343, "Applications of Mixed Models in Agriculture and Related Disciplines," available from the Louisiana State University Experiment Station in Baton Rouge.

Q: Is it possible to reproduce my PROC GLM analysis with PROC MIXED?

A: This is highly dependent on the model. As a general rule, if you have a RANDOM or REPEATED statement, PROC MIXED will not match PROC GLM. See the PROC MIXED documentation.

Q: How can I get the sums of squares for my fixed effects?

A: You can't. PROC MIXED uses maximum likelihood methods and does not compute sums of squares. See the CONTRAST statement documentation for an explanation of how the F statistic is constructed.

Q: How can I get my parameter estimates?

A: For fixed effects, use the SOLUTION option on the MODEL statement. For random effects, use the SOLUTION option on the RANDOM statement.

Q: PROC MIXED is taking a very long time, or a lot of computer resources or is causing an ABEND of some kind. What might be causing this?

A: See SAS Notes 5568 and 6176. The MIXED procedure is potentially one of the biggest resource hogs in SAS/STAT.

Q: Where is my PROC MIXED output? Why is my PROC MIXED output going directly to the printer?

A: See SAS Note 7068. This is an output delivery system issue.

Q: Can PROC MIXED handle spatial error structures?

A: Yes, in either the RANDOM or REPEATED statement. The structures available (such as the POWER structure) are documented in Technical Report P-229, page 312.

PROC MODEL (SAS/ETS)

Q: Is there any way to impose bounds on the parameter estimates in my model?

A: Currently, the MODEL procedure does not support a BOUNDS statement, although one is planned for a future release. Several strategies are provided in the documentation for imposing bounds on the parameters. These are described on pp. 589-590 of *SAS/ETS User's Guide, Second Edition*.

Q: Can I do Generalized Method of Moments Estimation?

A: Yes. As of release 6.08, the MODEL procedure supports GMM estimation by specifying GMM on the FIT statement. The BARTLETT, PARZEN and Quadratic Spectral kernels are supported for this method. For the details, see pp. 555-557 of *SAS/ETS User's Guide, Second Edition*.

Q: Can I do Full Information Maximum Likelihood estimation?

A: Yes, as of release 6.08, the MODEL procedure supports FIML estimation by specifying FIML on the FIT statement. Several estimators of the variance covariance matrix of the parameter estimates are available. For the details, see pp. 558-560 *SAS/ETS User's Guide, Second Edition*.

Q: I used to use PROCs SYSNLIN and SIMNLIN, but now I can't find them in the documentation. What happened to these procedures?

A: In Version 6, PROCs SYSNLIN, SIMNLIN and MODEL were rolled into one new MODEL procedure. The FIT statement in PROC MODEL provides the same functionality as PROC SYSNLIN (i.e. parameter estimation) and the SOLVE statement provides the same functionality as PROC SIMNLIN (i.e. model simulation and forecasting). Although we encourage you to convert to the new syntax for future models you develop, PROC SYSNLIN and PROC SIMNLIN syntax is still recognized by the system. It is simply converted internally to the appropriate PROC MODEL code.

PROC MORTGAGE (SAS/ETS)

Q: What is the current status of this procedure?

A: The newer PROC LOAN in 6.07.03 and later supersedes PROC MORTGAGE and is more versatile.

PROC MULTTEST (SAS/STAT)

Q: How do I compute a trend test for proportions?

A: Suppose Y indicates response or no response and X is the dose amount of a drug. You can test that the proportion of response increases with dose using the Cochran-Armitage test in PROC MULTTEST (with CLASS X; TEST CA(Y);). This can also be done in PROC LOGISTIC (with MODEL Y=X;) by looking at the score test printed in the 'Criteria for Assessing Model Fit' table of the printed output.

Neural Network Macros: %TNN_NLP, %TNN_RUN (SAS/OR) and %TNN_MODL (SAS/ETS)

Q: What do the Neural Network Macros do?

A: These macros are intended as EXAMPLES of how artificial neural networks can be implemented with SAS software. These macros are NOT full-featured production-quality programs suitable for serious data analysis in a wide range of practical applications. The %TNN_NLP macro fits a standard feed-forward neural network with one hidden layer (a multilayer perceptron) as a nonlinear multivariate regression model by least squares using the SAS/OR procedure NLP for nonlinear programming. The %TNN_SIM macro simulates a network using parameter estimates (weights) previously computed by the %TNN_NLP

macro. The predicted values and residuals are stored in the OUT= data set. Various statistics measuring goodness or badness of fit are stored in the OUTFIT= data set. The %TNN_MODL model does the same thing as the %TNN_NLP and %TNN_RUN macros, but it uses the MODEL procedure in SAS/ETS instead of PROC NLP in SAS/OR.

Q: What do the Neural Network macros require?

A: SAS/OR release 6.08 or higher (PROC NLP). %TNN_MODL does not require SAS/OR, but does require SAS/ETS (PROC MODEL).

Q: How do I get the Neural Network macros?

A: From the sample libraries.

- ☐ WWW connect to:
<http://www.sas.com/techsup/download/sibbs/stat/> and get nlpx7.sas or model7.sas
- ☐ anonymous ftp: ftp ftp.sas.com get
/users/ftp/techsup/download/sibbs/stat/nlpx7.sas or
/users/ftp/techsup/download/sibbs/stat/model7.sas
- ☐ SIBBS: download nlpx7.sas or model7.sas from STAT area. To use, see documentation in headers.

Q: Is there any further information on implementing Neural Networks in SAS?

A: Files containing two articles (in Postscript form), macro code, and examples are available via anonymous ftp in the directory /pub/sugi19/neural/. Connect to the Institute's ftp server with the command: ftp ftp.sas.com These articles appear in the *SUGI 19 Proceedings* (1994). neural1.ps - Sarle, W.S. (1994), "Neural Networks and Statistical Models," *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute, pp 1538-1550. neural2.ps - Sarle, W.S. (1994), "Neural Network Implementation in SAS Software," *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute, pp 1551-1573. (Slightly revised version) plots.ps - Plots from the 2nd paper in high-resolution graphics. macros.sas - Macros from the 2nd paper. example.sas - Examples using the macros with the XOR and sine data. example.bls - Output from example.sas. example2.sas - Examples using the macros with the motorcycle data. example2.bls - Output from example2.sas.

PROC NLIN (SAS/STAT)

Q: How can I get an rsquare for my nonlinear model?

A: One way to calculate an rsquare for a nonlinear model is demonstrated in the sample program NLINRSQ.SAS, which is available in the SAS Sample Library. However, you should note that this estimator is no longer bounded by zero and one. There is also debate as to whether corrected sums of squares should be used. For a discussion of the properties of different estimators of rsquare see Kvalseth's paper, "A Cautionary Note About Rsquare," in *The American Statistician* (November, 1985). It discusses eight different estimators of rsquare, all of which are equivalent for a linear model with an intercept, in terms of both linear models without an intercept and nonlinear models.

%NLINMIX macro

Q: What does %NLINMIX do?

A: It fits nonlinear, mixed models.

Q: What is required?

A: SAS/STAT, release 6.08 or later. An optional macro that generates plots based on the fitted model uses PROC GPLOT, so SAS/GRAPH will be needed to use it.

Q: How do I get %NLINMIX?

A: From Release 6.10 (or later) STAT sample library, member nlinmix.sas.

- ☐ WWW connect to:
<http://www.sas.com/techsup/download/sibbs/stat/> and get nlinmix.sas and nlmixex.sas
- ☐ anonymous ftp: ftp ftp.sas.com get
/users/ftp/techsup/download/sibbs/stat/nlinmix.sas get
/users/ftp/techsup/download/sibbs/stat/nlmixex.sas
- ☐ SIBBS: download nlinmix.sas and nlmixex.sas from the STAT

area. To use, see documentation in header. Examples of use are in `nlinmix.sas` (and following the macro definition in the sample library member, `nlinmix.sas`).

PROC NLP (SAS/OR)

Q: What is the status on documentation and availability for PROC NLP?
A: PROC NLP is available as an EXPERIMENTAL procedure beginning in Release 6.08. The procedure solves nonlinear programming problems, i.e., optimizing a nonlinear objective function when they have linear or boundary constraints. There is an extended users guide available by contacting Technical Support.

Q: Can PROC NLP handle nonlinear constraints?

A: No. But beginning in Release 6.09, there are routines in PROC IML that will. There is an internal draft document that describes these routines that is available upon request from Technical Support. The PROC IML routines are `NLPNMS`, `NLPQUA`, `NLPCG`, `NLPQN`, `NLPDD`, `NLPTR`, `NLPNRA`, `NLPNRR`, `NLPLM`, and `NLPHQN`. Each routine performs the optimization using a different algorithm.

Q: Does SAS have a function or proc that will maximize my likelihood function?

A: PROC NLP will maximize a likelihood function as long as you can express the function in DATA step code. There are examples available, and it is not difficult to do.

PROC NPAR1WAY (SAS/STAT)

Q: Can I get non-parametric multiple comparisons from NPAR1WAY?
A: No, but this is a suggestion on the *SASware Ballot*. (Some users run PROC GLM on ranked data.)

Q: Can I get exact, 1-sample, non-parametric tests of location for small samples?

A: Yes. PROC UNIVARIATE gives the Wilcoxon signed-rank test and the sign test which are exact for small samples (the Wilcoxon uses an exact method when $N < 20$).

Q: Can I get exact, 2-sample, non-parametric tests of location for small samples?

A: Not with any procedure. However, exact, two-sample tests can be done with some SAS/IML modules as presented in the *Observations* 4Q93 (Vol. 3 No. 1) article, "A PROC IML program to obtain exact significance levels in the non-parametric, two-independent-samples, location problem". The PROC IML code implementing this is available via anonymous ftp at `ftp.sas.com`, in the file `/observations/v3n1/berry/permtest`.

PROC OPTEX (SAS/QC)

Q: The Reference Guide doesn't show PROC OPTEX examples. Are there any?

A: Yes. "SAS/QC Software: Usage & Reference, Version 6, First Edition" provides complete documentation, including introductory examples and advanced examples for Release 6.10. In general, this book can be used for all current release of SAS/QC.

Q: How can I completely randomize the output plan?

A: You can't with OPTEX. After you output the plan, add a random variable and then sort on it, for example: `data outplan; set outplan; random=ranuni(123); run; proc sort; by random; run;`

Q: Can I create balanced incomplete block designs (BIBDs)?

A: Starting with Release 6.10, you can use the new `BLOCKS` statement to create BIBDs. Use the `STRUCTURE=` option on `BLOCKS` to specify the number and size of the blocks.

Q: Some of my factors have more than two levels, but the design it creates only has two levels for these factors. How can I get it to use all levels?

A: These factors should be specified on a `CLASS` statement. Otherwise, they are treated as regression effects and the most efficient design for such effects uses only their smallest and largest values.

PROC PHREG (SAS/STAT)

Q: Is PROC PHREG available on the PC under Release 6.04?

A: Yes. The SAS site representative must call to request it.

Q: How do I fit a "multinomial logistic" or "discrete choice" or "McFadden's conditional logistic" model?

A: TS273 shows an example of how to do it.

PROC PLAN (SAS/STAT)

Q: How can I completely randomize the output plan?

A: You can't with PLAN. After you output the plan, add a random variable and then sort on it, for example: `data outplan; set outplan; random=ranuni(123); run; proc sort; by random; run;`

%PLOTIT macro

Q: What does %PLOTIT do?

A: %PLOTIT is a complex macro that creates plots of labeled points typically used to display the results of correspondence analysis, principal components analysis, multidimensional scaling and preference mapping. It equates axes in the process of making such plots. TS259 discusses and illustrates its use, or equivalently, refer to *Observations* 4Q94 (Vol. 4 No. 1), pg 23.

Q: How do I get %PLOTIT?

A: from STAT sample library (Release 6.08 or later), member `crsplot.sas`. NOTE: The following versions of %PLOTIT require Release 6.10 or later:

- ☐ WWW connect to:
`http://www.sas.com/techsup/download/sibbs/stat/` and get `crsplot.sas`
- ☐ anonymous ftp: `ftp ftp.sas.com get /users/ftp/techsup/download/sibbs/stat/crsplot.sas`
- ☐ SIBBS: download `crsplot.sas` from the STAT area.

PROC PROBIT (SAS/STAT)

Q: Is PROC PROBIT available on the PC under Releases 6.03 or 6.04?

A: It was not available when Release 6.03 was released. However, it was added to Release 6.03 when the 6.03 update was shipped. It was documented in Technical Report P-179. It has been in all releases since (including Release 6.04).

Q: Why are my parameter estimate signs backwards? How can I fix them?

A: PROC PROBIT models the probability of the LOWER response level. For example, for a binary response, Y, with levels 0 and 1, PROC PROBIT models a quantity that increases and decreases with $\Pr(Y=0)$. This causes the parameter estimate signs to be reversed from modeling $\Pr(Y=1)$. If you want to model $\Pr(Y=1)$, then do:

```
PROC SORT; BY DESCENDING Y;
PROC PROBIT ORDER=DATA ....
```

Q: How do I classify/score new observations?

A: Add the new observations to the original data set, setting the response to missing for all of the new observations. PROC PROBIT ignores all observations with missing response values (so the fit will be identical to using just the original data set). Then, use the `OUTPUT` statement to request the `OUT=` data set and predicted probabilities (`P=varname`). The `OUT=` data set will contain predicted probabilities for all observations, including the new ones. Do not use PROC SCORE.

Q: Why are some or all of the inverse confidence limits missing?

A: If the first parameter estimate (after `INTERCPT`) is not strongly significant, missing inverse confidence limits often result, and this is correct. Keep in mind that you're trying to get a confidence interval on the INDEPENDENT variable that would generate a given response. Think of a plot of response on the vertical axis and the first independent variable on the horizontal axis (any other independents are held fixed). Imagine a regression line with confidence bands around it. Now draw a horizontal line at the response of interest. Where the horizontal line hits the confidence bands defines the inverse confidence interval on the independent variable for that response. Now, if there is a strongly

significant slope, the horizontal line should intersect both confidence bands. But, if the slope is near zero (the regression line is flat), then the horizontal line may contact only one or neither band. When this happens, the confidence interval is just an upper or lower bound or is unbounded.

Q: Why can't I get the OUTEST= data set?

A: The OUTEST= data set cannot be generated if the CLASS statement is used. See SAS Note 6000 for a possible work-around.

Q: Can %PROBIT compute inverse Mill's ratio? Can %PROBIT do Heckman's two-stage model to correct for sample-selection?

A: No, but there is a user-contributed program that does both of these.

- WWW connect to:
<http://www.sas.com/techsup/download/sibbs/stat/> and get heckman.sas
- anonymous ftp: <ftp://ftp.sas.com/get/users/ftp/techsup/download/sibbs/stat/heckman.sas>
- SIBBS: download file heckman.sas from STAT area.

Q: What are some issues regarding LACKFIT tests or heterogeneity factors with respect to sample size or degrees of freedom?

A: Because continuous variables may appear in the model, the lack-of-fit statistics and the heterogeneity factor are not computed within replicate groups and then pooled. However, the data should be presorted by the covariates (i.e. independent variables) to correctly compute the statistics. Further, the LACKFIT tests are not good tests of lack-of-fit if the number of replicates at some settings of the covariates is small. The heterogeneity factor is an estimate of overdispersion and is simply the Pearson chi-square statistic divided by its degrees of freedom. This is appropriate when there is little or no replication. We intend to add a "replicates analysis" option in Version 7 that will compute pooled, within-replicate group LACKFIT tests and overdispersion. A good discussion of these issues is given in McCullagh and Nelder, 1989, "Generalized Linear Models, 2nd edition" on pages 118-122 and 126-128. Also, see the chapter on overdispersion in D. Collett's book, "Modelling Binary Data" (1991), Chapman & Hall.

Q: Does %PROBIT do bivariate probit models?

A: No.

PROJMAN MENU SYSTEM (SAS/OR)

Q: What documentation do I need for PROJMAN?

A: For release 6.08 and later, you should have: *SAS/OR Software: The PROJMAN Menu System, V6, 1st Ed.*, and *SAS/OR User's Guide: Project Management, V6, 1st Ed.* For Release 6.10, see also "SAS Software: Changes and Enhancements, Release 6.10." Other project management features available in SAS/OR are documented in "SAS/OR User's Guide, V6, 1st Ed." Possibly also helpful are: *SAS/OR Software: Project Management Examples, V6, 1st Ed.*, and *Project Management Using the SAS System, V6, 1st Ed.* Note: The "OR User's Guide" and the "OR User's Guide: Project Management" are sent when SAS/OR is licensed. With 6.10 licenses, the "Changes and Enhancements" document is also sent. Along with these is a reply card that allows you to get a copy of the "PROJMAN Menu System" book, free of charge.

PROC RANK (Base SAS)

Q: What percentile/quartile/decile is each observation in?

A: Use the RANK procedure (documented in the SAS Procedures Guide, Version 6, Third Ed. for this. Specify GROUPS=100 to get percentile ranks, GROUPS=10 for decile ranks and GROUPS=4 for quartile ranks.

PROC REG (SAS/STAT)

Q: The NOINT option on the MODEL statement causes a note that the rsquare has been redefined. How is it defined with NOINT?

A: The rsquare is still calculated as the regression sum of squares divided by the total sum of squares. However, when there is not an intercept in the model, the uncorrected, rather than corrected, sums of squares is used. The best reference for exactly how the calculations are being performed is the section on multiple correlation (pages 95-98) of

Searle's book "Linear Models" (1971, John Wiley & Sons). A good discussion of regression through the origin is also found in Section 2.4.5 (pages 33-38) of the "SAS System for Regression, Second Edition" (1991, SAS Institute Inc.). Note that PROC GLM calculates the rsquare for the NOINT model in the same way.

For a discussion of the properties of different estimators of rsquare see Kvalseth's paper, "A Cautionary Note About Rsquare," in the *American Statistician* (November, 1985). It discusses eight different estimators of rsquare, all of which are equivalent for a linear model with an intercept, in terms of both linear models without an intercept and nonlinear models.

The most important things to note are that the rsquare for a nointercept model is calculated differently from that for a linear model with an intercept and that the two should NOT be directly compared. For those wanting to make such a comparison, Myers suggests calculating an rsquare for the NOINT model as:

$$1 - (\text{error ss from noint model} / \text{corrected total ss}).$$

For further discussion, see pages 30-31 of Myers' book "Classical and Modern Regression with Applications" (1986, PWS Publishers).

Q: How do I get the parameter estimates (a.k.a. beta values or regression coefficients), t-statistics, p-values or confidence intervals into a SAS data set?

A: The OUTEST= option on the PROC REG statement creates a data set that contains the parameter estimates and optionally the variance-covariance matrix of the parameter estimates (if you also specify the COVOUT option on the PROC REG statement). If you want the t-statistics and probabilities, see the REGCL sample program. See pages 1389-1391 of the *SAS/STAT User's Guide, V6, Fourth Edition* for a discussion of what is available in this data set. Many enhancements to this data set are documented on page 526 of Technical Report P-229. For example, you can now get the rsquare, adjusted rsquare and many other stats in this data set.

Q: How can I get the standardized parameter estimates into the OUTEST data set?

A: Use PROC STANDARD to standardize the data first (PROC STANDARD MEAN=0 STD=1); then run PROC REG with the OUTEST= option specified on the PROC statement. When the data is standardized prior to going into PROC REG, the regular estimates are the same as those produced by the STB option when PROC REG is executed on the raw data.

Q: How can I get PROC REG to use the standard errors produced by the ACOV option in calculating my t-statistics?

A: If you specify single degree of freedom tests (using the TEST statement), PROC REG will compute the test using the regular standard errors and then will recompute the test using the standard errors from the ACOV matrix. It will be a chi-square, rather than a t, since the standard errors are asymptotic, rather than exact. See pages 820-821 of Hal White's (May, 1980) *Econometrica* paper for details.

Q: Can I apply an inequality constraint (for example, B1>0) to the parameters in PROC REG?

A: The RESTRICT statement in PROC REG only allows equality constraints. However, you can use the BOUNDS statement in PROC NLIN to apply inequality constraints. PROC NLIN solves linear models quickly!

Q: What exactly is the SPEC option testing?

A: The SPEC option is testing a joint null hypothesis of the following:

- the errors are homoskedastic
- the errors are independent of the regressors
- the model is correctly specified

Failure of any of these conditions can lead to a statistically significant test statistic. Essentially, the statistic is testing a joint hypothesis that the model's specification of the first and second moments of the dependent variable is correct. See the bottom of page 823 of Hal White's May 1980 *Econometrica* paper for more details (the SPEC option is Theorem 2).

Q: How does PROC REG/GLM/etc. compute this or that statistic?
A: See Chapter 1, "Introduction to Regression Procedures," in the *SAS/STAT User's Guide* for many formulas (confidence limits, standard errors, etc.) that are used by ALL regression procedures.

PROC SHEWHART (SAS/QC)

Q: The Reference Guide doesn't show examples. Are there any?
A: Yes. "SAS/QC Software: Usage & Reference, Version 6, First Edition" provides complete documentation, including introductory examples and advanced examples for Release 6.10. In general, this book can be used for all current release of SAS/QC.

Q: The standard deviation computed by PROC SHEWHART is different from the one computed by PROCs MEANS or UNIVARIATE, or by hand. Why?

A: The formulas used to compute variability are different. PROCs MEANS and UNIVARIATE use the usual 'sum of squared deviations from the mean' formula, whereas in control charting, an unweighted average of subgroup standard deviations or ranges is computed. See the Computational Details section in *SAS/QC Reference, Ver. 6, First Edition* for the formulas used in quality control.

Q: The limits from a LIMITS= data set are being ignored—why would this happen?

A: By default, control limits will be computed based on the input DATA= data set. If you want to use the limits from a LIMITS= data set, you must also specify the READLIMITS option on the chart statement. The reason you must tell PROC SHEWHART to use the limits is so that you can have multiple chart statements in one run of the procedure. Therefore you can choose which charts will use previously determined limits and which charts will have limits based on the raw input data.

Q: How do I control the number of pages PROC SHEWHART uses to print a chart? How can I put more subgroups on one page?

A: Use the NPANELPOS= option on any chart statement to tell PROC SHEWHART how many subgroup positions to put on a page. For example, if you have 100 subgroups and you want your control chart to come out on one page, use NPANELPOS=100 on the chart statement. If you want this same control chart to come out on two pages, use NPANELPOS=50. The default is NPANELPOS=50 for all charts except boxchart, where the default is 20.

Q: Can SAS do a Gage (or gauge) R&R analysis?

A: The GAGE application is part of the Release 6.10 SAS/QC Sample Library. It can be used to produce range charts and average charts, and it also performs the Average/Range and Variance Components Methods. Reference: "SAS Tools for Assessing Gage Repeatability and Reproducibility", Julie LaBarr, *SUGI 19 Proceedings* (1994). Requirements: SAS/QC is required. SAS/AF is required if you want to modify the application, but it is not needed to run it. SAS/STAT is required for the Variance Components Method.

SQC MENU SYSTEM (SAS/QC)

Q: How do I start up the SQC Menu System?

A: ☐ Releases 6.03 and 6.04: Type the following on the command line:
 af c=sashelp.sqc.qcmenu.program
☐ Releases 6.06, 6.07.01, 6.07.02: 1) Under SAS/Assist, select Planning Tools, then Quality Cntl or 2) type SQC on the command line
☐ Releases 6.07.03 and later: 1) Under Assist, select Data Analysis, then Quality Cntl or 2) type SQC on the command line or 3) via pmenus: Globals->Invoke Application->Data Analysis->Quality Control

%STDIZE macro

Q: What does %STDIZE require?

A: Just base SAS unless the METHOD=AGK(p) or METHOD=L(p) options are used, in which case SAS/STAT is needed, and Release 6.06 and higher.

Q: What does %STDIZE do?

A: The %STDIZE macro standardizes one or more numeric variables in a SAS data set by subtracting a location measure and dividing by a scale measure. A variety of location and scale measures are provided, including estimates that are resistant to outliers and clustering (see the METHOD= argument). You can also multiply each standardized value by a constant and add a constant.

Q: How do I get %STDIZE?

A: Stat sample library (6.09 or later): member stdize.sas

- ☐ WWW connect to:
<http://www.sas.com/techsup/download/sibbs/stat/> and get stdize.sas and xmacro.sas via
- ☐ anonymous ftp: ftp ftp.sas.com get
 /users/ftp/techsup/download/sibbs/stat/xmacro.sas get
 /users/ftp/techsup/download/sibbs/stat/stdize.sas
- ☐ SIBBS: download stdize.sas and xmacro.sas from STAT area. To use, see documentation in headers.

PROC STEPDISC (SAS/STAT)

Q: Can PROC STEPDISC print the discriminant functions?

A: No. Use PROC DISCRIM instead.

Q: Can PROC STEPDISC output the selected variables to PROC DISCRIM?

A: No.

Q: Which variables are most important (are best discriminators)?

A: Use PROC STEPDISC to help select the best discriminating variables. Once a set of variables is selected, PROC CANDISC (or PROC DISCRIM with the CANONICAL option) can tell you how important each variable is via the structure coefficients. The discriminant function coefficients from PROC DISCRIM are NOT the best way to determine variable importance since correlation among the variables can make their interpretation in that way misleading.

PROC TREE (SAS/STAT)

Q: Can I get high resolution dendrograms (tree diagrams)?

A: PROC TREE does not currently produce high resolution trees, but it will begin with release 6.11. Currently, there is a user-written macro that will work, and it was published *Observations*, 1Q95 (Vol. 4, No. 2). The code is available upon request.

%TREEDISC MACRO (SAS/IML)

Q: What does %TREEDISC require?

A: SAS/IML. SAS/OR is required to draw the tree, but the tree can be listed without SAS/OR. Release 6.08 or higher is recommended.

Q: What does %TREEDISC do?

A: The %TREEDISC macro generates a SAS data set which describes a decision tree computed from an input data set to predict a specified categorical dependent variable from one or more other predictor variables. The tree can be listed or drawn or used to generate code for a SAS DATA step to classify observations.

Q: How do I get %TREEDISC?

- ☐ WWW connect to:
<http://www.sas.com/techsup/download/sibbs/stat/> and get treedisc.sas and xmacro.sas.
- ☐ anonymous ftp: ftp ftp.sas.com get
 /users/ftp/techsup/download/sibbs/stat/treedisc.sas get
 /users/ftp/techsup/download/sibbs/stat/xmacro.sas
- ☐ SIBBS: download treedisc.sas and xmacro.sas from STAT area.

To use, see documentation in headers.

Q: What are the three sets of statistics printed at each node?

A: VAL: Stands for VALUES. The values you see are the values of the predictor variable by which the node was split. For the root (the trunk of the tree) the values of the dependent variable are given. Missing values for character variables are noted with an '*'. COU: Stands for COUNT. This category contains the number of observations in each category of

the dependent variable on the node. PVAL: Stands for PVALUES. The values listed here are the two smallest adjusted p values which could be used for further splits. If there are no p values present TREEDISC will not split again at that branch at that node.

Q: Why weren't some of my observations classified?

A: Whenever an observation could be assigned to more than one node, it is considered "tied." The TREEDISC variable TIE_ shows the value two or more. The observation is dropped from the analysis.

PROC TSCSREG (SAS/ETS)

Q: Is TSCSREG available under the 6.04 release?

A: An experimental version of the procedure is available upon request through Technical Support.

Q: Why do I get the warning: "phi matrix is singular" when I run PARKS method?

A: Any time the number of cross sections is greater than the number of time series observations per cross section, PROC TSCSREG will produce the warning. (This is analogous to running seemingly unrelated regression with fewer observations than equations in the model.) The only circumvention is to reduce the number of cross sections.

Q: PROC TSCSREG requires the same number of time series observations in each cross section. Is there any way around this restriction? I don't want to omit any of my data from the analysis.

A: There is no way around this current limitation in PROC TSCSREG. However, if you have access to the SAS/STAT product, then the MIXED procedure can be used to fit models that are very close to the Fuller-Battese and Da Silva methods in PROC TSCSREG. Following are examples of the corresponding specifications:

```
/* Fuller-Battese model in TSCSREG */
proc tscsreg data=a fuller;
  model y=x1 x2 x3 / covb;
  id state year;
run;
/* Fuller-Battese model in MIXED */
proc mixed data=a;
  class state year;
  model y=x1 x2 x3 / s p;
  random state year;
run;
/* Dasilva model in TSCSREG */
proc tscsreg data=a dasilva m=5;
  model y = x1 x2 x3;
  id state year;
run;
/* DaSilva model in PROC MIXED please note that the estimates
will not be identical since TSCSREG uses Seely's method and
MIXED uses REML. But they should be close. */
proc mixed data=a;
  class state year;
  model y = x1 x2 x3 / s;
  random state year;
  repeated / type=toep(6) sub=state;
run;
```

PROC TTEST (SAS/STAT)

Q: Can I test the difference between two groups if I only have summary statistics (means and std dev's) and not the raw data?

A: No, but this is easily done in a DATA step. The necessary formulae are in the PROC TTEST documentation. You can use the PROBF and PROBT functions to compute the p-values.

PROC UNIVARIATE (Base SAS)

Q: Can I get just the box plot? Just the probability plot? Just the histogram?

A: No, you cannot select the plots separately. If you have SAS/QC, PROC CAPABILITY can be used to create probability plots and histograms, and PROC SHEWHART will produce box plots. SAS/GRAPH can also be used to get box plots with PROC GPLOT and

the INTERPOL=BOX option on a SYMBOL statement. See Technical Report A-106, *Probability Plotting*, for code to produce probability plots using only BASE SAS. TSSO discusses probability plots using SAS/GRAPH.

Q: I'm trying to test my data for normality with the NORMAL option in PROC UNIVARIATE. How do I interpret the values that I get?

A: You can always look at the probability value (labeled either PROB<W or PROB>D). Small values may indicate non-normality, and typical choices are 0.05 and 0.10. Be sure and look at the normal plot too, as a large sample size makes the test very powerful, and may cause the test to reject a sample that is very close to being normal. Many statistical tests that require the assumption of normality are robust, so 'very close' to normal may be good enough. See page 38 of the second quarter 1990 SAS Communications and page 627 of the *SAS Procedures Guide*, V6, 3rd ed.

Q: Can PROC UNIVARIATE compute a weighted median or other quantile?

A: No. The WEIGHT statement does not affect the quantiles. If the weight values are not extremely large and are integers (or can be made integers by multiplying them by a constant that doesn't make them too large), then the FREQ statement can be used. A weighted median can be computed with PROC FASTCLUS in Release 6.07.03 or later. The Cluster Median reported by FASTCLUS is a weighted median when the options LEAST=1 (see Technical Report P-229) and MAXCLUSTERS=1 are used along with a WEIGHT statement. The statistic computed by FASTCLUS is a median in the sense that it minimizes the weighted sum of the p-th powers of the absolute deviations.

Q: How does PROC MEANS/SUMMARY/UNIVARIATE compute this or that statistic?

A: The answer is probably in Chapter 1 "SAS Elementary Statistics Procedures" in the *SAS Procedures Guide*. There is also a very useful table on pages 2-3 of the procedures guide that shows which descriptive statistics are computed in which procedures.

Q: My standard deviation (or variance) computed by PROC MEANS/SUMMARY/ UNIVARIATE is much too large. What is the matter?

A: If you are using a WEIGHT statement you probably also need to use the VARDEF= option on the PROC statement.

Q: What percentile/quartile/decile is each observation in?

A: Use the RANK procedure (documented in the *Procedures Guide*) for this. Specify GROUPS=100 to get percentile ranks, GROUPS=10 for decile ranks and GROUPS=4 for quartile ranks.

Q: Why don't I get the same p-value for the Kolmogorov-Smirnov statistic as when I use a table?

A: The critical values of the K-S statistic depend on N (as with most statistics). However, Stephens (see the reference given at end of the PROC UNIVARIATE chapter in the *Procedures Guide*) came up with a transformation of the K-S statistic that incorporates N, so that you can compare the transformed statistic against a single set of critical values. That is what we do in SAS/INSIGHT and PROC UNIVARIATE.

Q: When outputting several statistics on several variables, the output data set has everything on one observation. Can I reshape it to be more useful?

A: Yes. See *SAS Communications*, 1Q90, pg 41 for a program that reshapes the output data set so that variables are the statistics and the observations are the original variables.

Q: How do I get a 95% confidence interval of the median or 1st or 3rd quartile? How can I get the standard error of the median?

A: For a variable, Y, PROC LIFETEST (Release 6.08 or later) gives the median and a CI: proc lifetest alpha=.05; time y; run; The standard error of the median can be obtained from the STDMD() function available in SAS/QC. See page 596 of the *QC Ref. Guide*.

PROC VARCOMP (SAS/STAT)

Q: My PROC VARCOMP and PROC GLM results don't agree - why?

A: They will agree - if you specify the model in such a way that they are fitting exactly the same Type I model in both procedures. The order of the effects on the MODEL statements is important, as well as appropriate specification of FIXED effects in PROC VARCOMP, and RANDOM effects in PROC GLM.

Q: How can I write the results of my PROC VARCOMP to a data set?

A: You can't. Use PROC MIXED.

Q: How does PROC VARCOMP compare to PROC MIXED?

A: PROC MIXED does not perform a Type I analysis. For certain models, PROC MIXED and PROC VARCOMP will not agree, and that is the nature of the algorithms used. PROC MIXED and PROC VARCOMP compute starting values differently. As a general rule, PROC MIXED is a more robust algorithm, but sometimes people are happier with PROC VARCOMP results. This is choice is up to you.

Statistical Keyword Reference List (test or item => PROC/module/sample program/info)

AID (Automatic Interaction Detector)—See CHAID

Alpha (Cronbach's)—PROC CORR

ANOVA using summary statistics—DATA STEP, see the Larson article in *The American Statistician*, May 1992.

ARCH (Autoregressive Conditional Heteroscedasticity)—PROC AUTOREG (Release 6.08)

Area estimation—TS233

Balanced incomplete block designs (BIBDs)—Beginning in Release 6.10 there are additional options in PROC OPTEX for creating BIBD designs. For earlier releases, see Technical Report P-188, Chapter 6, example 6 using PROC OPTEX. We can't guarantee that OPTEX will find BIBDs using this pre-6.10 method, but it often does.

Bartholomew test of ordering of proportions—Not available, but the Cochran-Armitage test of linear trend (see below), which tests a stricter alternative hypothesis, is available.

Bartlett test for variance homogeneity—SAMPLE LIBRARY: BARTLETT. Or download homovar.sas from STAT area of SIBBS. Or download techsup/download/samples/stat_and_uml/homovar.sas using anonymous ftp.

Bhapkar's test—PROC CATMOD using REPEATED statement to test marginal homogeneity as in example 7's 3 df test of SIDE. Reference is Agresti (1990), *Categorical Data Analysis*, pp359, 499. See Stuart-Maxwell.

Biserial correlation—not available, but see "point-biserial correlation" below.

Bivariate Probit model—Not available.

Bonferroni t-test—PROC GLM, ANOVA, MULTTEST

Box and whisker plots—PROC UNIVARIATE, SHEWHART, GPLOT (I=BOX... IN SYMBOL STMT) and SAS/INSIGHT (Release 6.08).

Box-Behnken Designs—SAS/QC ADX menu system gives them. There is no ADX macro for these.

Box-Cox Transformations—ADX menu system or ADXTRANS macro in SAS/QC.

Breslow-Day test (of homogeneity of odds ratios)—PROC FREQ, CMH option

CART (Classification and Regression Trees)—We don't do it currently, but development is considering writing a procedure to do it. If you are interested, call Technical Support so that we can ask some questions. See

CHAID below.

CHAID (Chi-square Automatic Interaction Detector)—A macro (%TREEDISC) to do this is available via SIBBS and anonymous ftp. Or the macros can be sent upon request.

Chi-square goodness-of-fit test for 1-way tables—Sample Library: Pearson *Observations* article (Vol 1, No 1), pg. 17; *SAS Communications* articles (90Q1 pg. 43 & 91Q3 pg. 35), TS176, SAS-Note 6062219, 6041468, 5186436.

Chi-square (2-way tables)—PROC FREQ

Chi-square(corrected)—PROC FREQ

Chow test—Currently, no SAS procedure automatically computes a Chow F test. However, it easy to do in either PROC REG or PROC SYSLIN using a TEST statement. "SAS/ETS Software: Applications Guide 2, Version 6, First Edition" provides several approaches from performing this test on pages 62-64.

Cluster analysis, categorical data—see DISTANCES

Cochran-Armitage trend test—PROC MULTTEST. This is a test for trend in proportions.

Cochran's Q—Release 6.10 PROC FREQ (AGREE option). Before Release 6.10: Set up as 3-way table and use the CMH option in the same way as for McNemar's test. See the McNemar example, pg 129 of *SAS/STAT User's Guide, Vol 1*.

Cohen's kappa—See kappa.

Concordance, Kendall's Coefficient—See Kendall's Coefficient of Concordance.

Conditional Logistic model—For 1:1 ("one to one") matching, see example 5 in PROC LOGISTIC. For M:N matching (for M>1, N>1), see example 3 in PROC PHREG. Or, see Multinomial Logit model below.

Confidence interval (on a mean)—DATA STEP, PROC MEANS

Conjoint Analysis—Technical Report R-109, *Conjoint Analysis Examples*.

Contrasts—PROC GLM, CATMOD

Control charts—PROC SHEWHART, MACONTROL, CUSUM

Correlations—PROC CORR (PEARSON, SPEARMAN, KENDALL)

CPK (process capability indices)—PROC CAPABILITY

CPM (Critical Path Method)—PROC CPM

D (Somer's)—PROC FREQ

Decision trees—PROC DTREE

Dickey-Fuller test for unit root—There are two macros for this in the autocall library when SAS/ETS is licensed. The macros are called %DFTEST and %DFPVALUE. %DFTEST automatically calls the %DFPVALUE macro. However, the %DFPVALUE may also be run on its own to compute the p-values for tests statistics that have already been computed. For more information regarding these and several other macros, see pp. 945-957 *SAS/ETS User's Guide, 2nd Ed*.

Differential equations—IML (CALL ODE)

Discrete Choice model—(See Multinomial Logit model below)

Distances—%DISTANCE MACRO in DISTANCE member of the Release 6.09 STAT sample library, or %DISTNEW, available upon request.

Duncan multiple range test—PROC GLM, ANOVA

Dunnett's test—PROC GLM, ANOVA

Durbin-Watson statistic—PROC GLM (with CLI or CLM options), REG (DW option), AUTOREG (DW= option)

DXY and DXY(Somer's)—PROC FREQ (MEASURES option)

ED50—See LD50.

Empirical distribution function—EDF option in PROC NPAR1WAY

Equality of means—PROC ANOVA, GLM, TTEST

Fishbone diagrams—PROC ISHIKAWA

Fisher's exact test—PROC FREQ (EXACT option)

Fisher's least significant difference—PROC ANOVA, GLM

Friedman's test—PROC FREQ (if only one response per treatment-block combination). See PROC FREQ example. Or PROC RANK; by block, then PROC GLM (if more than one response per combo).

Full-information maximum likelihood—PROC SYSLIN

Gage (or gauge) repeatability and reproducibility (R&R)—GAGE application (SAS/AF and FRAME) in 6.10 Sample library. Paper by LaBarr in *SUGI 19 proceedings* (1994) discusses this and the application.

GAM (Generalized Additive Models)—not available

GARCH (Generalized Autoregressive Conditional Heteroscedasticity)—PROC AUTOREG (Release 6.08 and later)

GEE (Generalized Estimating Equations)—Not available from the Institute. However, you can get a user-written macro that does it. See StatFAQs entry under PROC GENMOD above.

Geometric means—Sample Library (member: GEOMEAN)

Gini's mean difference—SAS/INSIGHT

Guttman scaling—There is a Version 5 supplemental library program called PROC GUTTMAN. This has not been converted to Version 6 largely because it is an extremely expensive algorithm and can only handle up to about 12 items. Guttman himself recommended correspondence analysis as a better alternative. (see "Measurement and Prediction", Stouffer and Guttman, Wiley 1966.). So, PROC CORRESP might be the thing to try.

Hoeffding's D—PROC CORR

Homogeneity of variance—Sample Library: BARTLETT, %HOMOVAR

Hotelling's T-square—PROC GLM and "SAS System for Linear Models," pp. 252-255

Integration—SAS/IML (CALL QUAD in Release 6.07.03); see also area estimation above

Intraclass correlation—PROC NESTED ("percent of total" column) or by using the %intracc macro available in the stat directory of SIBBS, or via anonymous FTP in the directory: /users/ftp/techsup/download/samples/stat_and_ims/. An example using a data set in *American Statistician*, Vol 47, No. 4, page 293 is available on request.

Inverse Mill's ratio—See Mill's ratio below

Ishikawa diagrams—PROC ISHIKAWA, Ishikawa menus & macros

Item analysis—SUGI PROC ITEM (IBM V5 ONLY) sample library (member: ITEM)

Jonckheere's test—PROC CORR with KENDALL OPTION (see *SUGI 14 Proceedings* (1989), 89 PP. 1337-9)

Kappa (Cohen's, weighted or not)—Release 6.10 PROC FREQ

(AGREE option). PROC CATMOD as described in TS118 is a poor second alternative.

Kendall's Coefficient of Concordance—Not available, but see the AGREE option of PROC FREQ in Release 6.10 (or later). It will test agreement between two raters.

Kendall correlation—PROC CORR

Kendall's Tau-a—PROC FREQ

Kurtosis—PROC UNIVARIATE, MEANS

LD50—PROC PROBIT INVERSECL option

Levene's test of equal variances—SAS/LAB

Likelihood ratio chi-square—PROC FREQ

Linear programming (optimization)—PROC LP

Ljung-box Q statistic—PROC ARIMA

Mantel-Haenszel chi square—PROC FREQ

McFadden's model—See Conditional Logistic model.

McNemar's test—6.10 PROC FREQ (AGREE option). Before Release 6.10: Set it up as a 3-way table and use the CMH option. See example on page 129, *SAS/STAT guide, Vol 1*.

Median—PROC UNIVARIATE, LIFETEST

Mill's ratio (or inverse)—SAS does not compute it, but if you want it in the context of a probit model, there is a user-contributed program that will do it. See the PROC PROBIT section above. It is on our list of suggestions.

Missing value replacement—PROC STANDARD (REPLACE option). Do not use PROC TRANSREG.

Moving average—PROC EXPAND (SAS/ETS) in Release 6.08 and higher supports this with the TRANSFORM= option on the CONVERT statement. See the *ETS User's Guide, 2nd Ed.* If you don't have SAS/ETS see the Sample Library (member: MAVERAGE). Also see pp. 223-228 of *SAS Language and Procedures, Usage 2*.

MLOGIT or MPROBIT procedures—These are not procedures written or supported by SAS Institute. If you want a Multinomial Logit Discrete-Choice model or Mill's ratio, see the entries for these above and below. See also version 5 SAS Note 3639 concerning these procedures.

Multinomial Logit model—PROC PHREG. See TS273

Multiple comparisons (repeated measures)—see "Performance of Multiple Comparisons in Repeated Measures Designs Under Nonsphericity" by Steve Thomson *SUGI 15 Proceedings*, (1990), pp. 1365—1370. It is also available as TS235.

Neural Networks—%TNN_NLP in NLPEX7 member of Release 6.09 SAS/OR sample library

Normality, test of—See "TEST OF NORMALITY," and PROC UNIVARIATE

Numerical integration—IML (CALL QUAD)

Odds Ratios—PROC LOGISTIC

Optimization—PROC LP, NETFLOW, NLP, TRANS, ASSIGN

Outlier detection—See "SAS System for Statistical Graphics, First Edition" Chapter 9 for methodology.

Paired comparisons t-test—PROC MEANS (See Example 2 in PROC TTEST for an example.)

Pareto charts—PROC PARETO

Partially Balanced Incomplete Block Designs (PBIBDs)—See Balanced Incomplete Block Designs. There is nothing that specifically creates these, but PROC OPTEX might do it in some cases.

Path analysis—PROC CALIS

Pearson correlation—PROC CORR

Percentiles—PROC RANK, UNIVARIATE

PERT—PROC CPM

Phi coefficient—PROC FREQ

Point biserial correlations—PROC CORR (see SAS Note 3956)

Polychoric correlation coefficient—PROC FREQ PLCORR option on the TABLES statement. See Technical Report P-222 pg 223.

Polyserial correlations—not currently available

Power—“%Power: A Simple Macro for Power and Sample Size Calculations” by Kristin Latour, *SUGI 17 Proceedings*, 1992, pp.1173—1177. It is available as TS272.

Probability plots—PROC CAPABILITY

Project management—PROC CPM, PROJMAN menu system

Queueing—see SIMULATION below

Random sampling—DATA STEP, see the *SAS Applications Guide*, 1987 Edition, or *SAS Language and Procedures, Usage 2* for suggested code.

Regression—PROC REG, GLM, PHREG, LIFEREG, and others.

Reliability coefficient (SPSS)—see ALPHA

Repeated measures multiple comparisons—see “Performance of Multiple Comparisons in Repeated Measures Designs Under Nonsphericity” by Steve Thomson, *SUGI 15 Proceedings*, (1990), pp. 1365—1370. It is also available as TS235.

ROC (Receiver Operating Characteristic) curve—PROC LOGISTIC - In Release 6.10, use the OUTROC= option. Before Release 6.10, plot sensitivity against (100-specificity) values from the output of the CTABLE option after specifying PPROB=0 to 1 by .05.

Rsquare—PROC REG, GLM, CALIS

Scheffe multiple comparisons—PROC GLM, ANOVA

Seemingly unrelated regression—PROC SYSLIN, MODEL

Skewness—PROC UNIVARIATE, MEANS

Somer's D—PROC FREQ

Somer's DXY and DYX—PROC FREQ

Spearman correlation—PROC CORR

Standardizing data—%STDIZE macro in STDIZE member of the Release 6.09 STAT sample library, PROC STANDARD

Stuart-Maxwell test—Use PROC CATMOD with the REPEATED statement to test marginal homogeneity. See Example 7 - the 3 df test of 'side'. This is actually Bhapkar's test, but it is asymptotically equivalent to the Stuart-Maxwell test. For reference see Agresti (1990), *Categorical Data Analysis*, pp 359, 499. See Bhapkar.

Student-Newman-Keuls test—PROC GLM, ANOVA

Student's t-test—PROC UNIVARIATE, MEANS, SUMMARY

Sum of squares—PROC UNIVARIATE, MEANS, SUMMARY, GLM

Summary statistics-ANOVA—See “ANOVA using summary statistics”

Symmetry, test of in 2-way table—AGREE option in PROC FREQ

(Release 6.10 and later)

Trimmed mean—SAS/INSIGHT

T-test—PROC TTEST

T-test(Bonferroni)—PROC GLM, ANOVA, MULTTEST

Tau-a(Kendall's)—PROC FREQ

Tau-c(Stuart's)—PROC FREQ

Test of homogeneity of variance—sample library (member: BARTLETT)

Test of normality—PROC UNIVARIATE, CAPABILITY. We have no test for multivariate normality.

Tetrachoric correlations—PLCORR option in PROC FREQ (see Technical Report P-229). The tetrachoric correlation is a polychoric correlation for 2x2 tables.

Three-stage least squares (3SLS)—PROC SYSLIN, MODEL

Trend test for 2 by c tables—PROC MULTTEST's Cochran-Armitage Test; PROC LOGISTIC's score test; PROC FREQ with CMH and SCORES=rank options

Tolerance intervals—PROC CAPABILITY

Tukey's range test—PROC GLM, ANOVA

Two-stage least squares (2SLS)—PROC SYSLIN, MODEL

Weighted means—PROC MEANS

Winsorized mean—SAS/INSIGHT

Yule's Q for 2x2 tables—same as Gamma statistic in PROC FREQ (CHISQ option)

How to Use Fse Anonymous FTP and SIBBS

ANONYMOUS FTP

SAS Institute's anonymous ftp service can be used to upload or download files without prior registration. To use this facility, connect to ftp.sas.com. Once you are connected, enter the following responses as prompted:

Name (ftp.sas.com:<userid>): anonymous
Password: <your E-Mail address>

If you get a message that ftp.sas.com is an unknown host, use the IP address instead; the IP address is 192.35.83.8.

SIBBS

The phone number to access SIBBS is: (919)677-8155. Communication software must be configured for no parity, 8 data bits, and 1 stop bit (8-N-1). Speed up to 14,400 baud. To register, log in as follows:

First name: first name
Last name: last name
City, State: city and state
PASSWORD: whatever password you desire

You will be prompted to enter your specific information. Accounts will be available for use immediately. You can track problems, review problems, resolve or add to them. You can request maintenance and submit SASware Ballots and also have the capability of downloading zaps and SAS Usage Notes.

SAS, SAS/ASSIST, SAS/ETS, SAS/FSP, SAS/GRAPH, SAS/IML, SAS/INSIGHT, SAS/LAB, SAS/OR, SAS/QC, SAS/STAT, and JMP are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. Observations, SAS Communications, and the SASware Ballot are published by SAS Institute Inc. IBM and OS/2 are registered trademarks or trademarks of International Business Machines Corporation. (R) indicates USA registration.

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