

Answers to Commonly Asked Statistics Questions - June 1996 Edition

Mike Stockstill, Technical Support Statistician

SAS Institute Inc., Cary, NC

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 particularly useful to site representatives who face a lot of questions about statistics.

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, the information is constantly updated, and we use this information to help resolve as many calls as possible on initial contact.

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 paper is to store 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. TS-507 (see below for comments on TS-###) is an ASCII version of this paper, and you may obtain it by phoning SAS Technical Support and requesting TS-507, or via the Institute home page, SIBBS, or anonymous ftp.

Frequently Asked Questions for all of the areas in Technical Support are available on the Technical Support WWW page. These are updated weekly and are searchable. To access, connect to:

<http://www.sas.com/service/techsup/FAQ.html>

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. Most of them are available online, and you can access them yourself as described below.

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

STATFAQS (STATISTICAL/MATHEMATICAL FREQUENTLY-ASKED-QUESTIONS)

Q: How do I access the TSNotes (TSdocs) from the Internet?

A: Those available online can be downloaded via ftp from the directory: techsup/download/technote, or from the Web in the directory:
<http://www.unx.sas.com/techsup/download/technote/> Note that some files are flat, 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: Yes. Some procedures implement these methods: MULTTEST can use bootstrap- or permutation-resampling to adjust the p-values of the t-test, Cochran-Armitage trend test or Fisher exact test. DISCRIM uses crossvalidation to obtain nearly unbiased estimates of the classification error rates in a discriminant analysis. Also, there are macros available online to perform jackknife and bootstrap analysis for simple random samples, computing approximate standard errors, bias-corrected estimates, and confidence intervals assuming a normal sampling distribution. In order to use these macros, you need to know enough about the SAS macro language to write simple macros yourself.

- ☐ User can obtain via WWW by connecting to:
<http://www.sas.com/techsup/download/stat/> and getting jackboot.sas.
- ☐ User can get these files via anonymous ftp: ftp ftp.sas.com get techsup/download/stat/jackboot.sas
- ☐ User can get via SIBBS: download jackboot.sas from STAT area.

Q: Has someone at SAS written a special program to do <insert statistic name here>?

A: If the answer is no, you can try to find out if someone in the user community has done so by posting your inquiry 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 causes "ERROR: Unable to access message ###.###"?

A: Messages like this can occur with ANY procedure. It means that the procedure was unable to find its 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: Is there a way to do a power analysis / determine the sample size needed ?

A: The %POWER macro will. It uses GLM's OUTSTAT data set 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 data set as the input data set, 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):

- ☐ User can obtain via WWW by connecting to:
<http://www.sas.com/techsup/download/stat/power.sas>
- ☐ User can obtain via anonymous ftp:
 - ☐ ftp ftp.sas.com
 - ☐ get techsup/download/stat/power.sas

- User can get via SIBBS: download power.sas from the STAT area.
To use, see documentation in header. Or see "%Power: A Simple Macro for Power and Sample Size Calculations" by Latour, SUGI 17, 1992, pp. 1173-1177 (TS-272).

Q: Is there a test of multivariate normality?

A: Yes. A macro to test multivariate normality is available online. It performs Mardia's tests based on multivariate skewness and kurtosis. It also provides a chi-square quantile-quantile plot that allows for visual assessment of multivariate normality. It requires Version 6.06 or later of SAS/IML Software or Version 6.09 TS450 or later (except Version 6.10) of SAS/ETS Software. If neither is found, only univariate normality tests can be provided (and, optionally, the Q-Q plot).

- User can obtain via WWW by connecting to: <http://www.sas.com/techsup/download/stat/> and getting multnorm.sas.
- User can get these files via anonymous ftp: ftp ftp.sas.com get techsup/download/stat/multnorm.sas
- User can get via SIBBS: download multnorm.sas from STAT area. To use, see documentation in header.

Q: How can I get a matrix of polychoric (or tetrachoric) correlations?

A: The PLCORR option in PROC FREQ computes the polychoric (or tetrachoric) correlation between two ordinal (or binary) variables. However, in one execution it cannot compute many such correlations and output a correlation matrix. The POLYCHOR macro creates a SAS data set containing a correlation matrix of polychoric correlations or a distance matrix based on polychoric correlations. The macro is available online:

- User can obtain via WWW by connecting to: <http://www.sas.com/techsup/download/stat/> and getting polychor.sas.
- User can get these files via anonymous ftp: ftp ftp.sas.com get techsup/download/stat/polychor.sas
- User can get via SIBBS: download polychor.sas from STAT area. To use, see documentation in header.

Q: How can I estimate parameters or test the fit or generate values from Johnson distributions?

A: The SAS/QC sample library programs, JOHNSYS1.SAS through JOHNSYS7.SAS, generate values from each of the Johnson Su, Sb and Sl systems of distributions, estimate the parameters of the distribution from the generated data, compute the Cpk index, and draw a high-resolution histogram overlaid with the fitted Johnson distribution curve. The file JOHNSYS4.SAS contains a SAS/IML based macro that chooses the proper system and estimates the parameters of the distribution. However, no test of the model fit is done. These files appear in the SAS/QC sample library in release 6.08 or later. SAS/IML and SAS/QC are required (PROC CAPABILITY is used).

Q: What machines and what procedures support vector processing?

A: IBM 3090 and DEC VAX 6000 and 9000 series machines are capable of vector processing (multiple processors used simultaneously). SAS Software uses the vector processing capabilities of these machines. See SAS Note V6-SYS.SYS-0719 and V6-SYS.SYS-3412 and TS-161.

Q: Is there an interface between SAS Software and StatXact(r)?

A: SAS Institute provides no interface with StatXact. StatXact [StatExact] is a software product from Cytel Software Corp. They have created a limited interface that allows StatXact to read SAS

transport files. There may be additional capabilities. For information, contact Cytel Software Corp. at (617) 661-2011.

Q: How can I compute points along the curve (density, mass or distribution) of a theoretical distribution (such as Normal or Gamma)?

A: Beginning in 6.11, the new PDF, CDF, and LOGPDF functions are available to calculate points on the probability density (or mass) function or cumulative distribution functions and the log of the density for many distributions. Distributions available are the Bernoulli, Beta, Binomial, Cauchy, Chi-square, Exponential, F, Gamma, Geometric, Hypergeometric, Inverse Gaussian (Wald), Laplace, Logistic, Lognormal, Negative Binomial, Normal, Pareto, Poisson, T, Uniform, and Weibull. Documentation is available in TS-506.

Q: How can I get the data sets and programs from a statistical book published by SAS Institute?

A: Following is information for online access to the data sets and SAS code from our statistical books. Access is via the WWW, anonymous ftp, or SIBBS. For WWW, the URL and filename are given. For anonymous ftp, connect to ftp.sas.com, change directory to the directory shown and download the filename shown. For SIBBS, go to the STAT area of the Download menu and download the filename shown. In some cases, compressed files are also available. The filenames of compressed files end in ".Z" or ".zip".
SAS System for Linear Models, 3rd Edition (56140)
 WWW: <http://www.sas.com/samples/> file: A56140 or A56140.Z
 ftp: pub/publications/ file: A56140 or A56140.Z SIBBS: linmod.sas or linmod.zip
SAS System for Regression, 2nd Edition (56141)
 WWW: <http://www.sas.com/samples/> file: A56141 or A56141.Z
 ftp: pub/publications/ file: A56141 or A56141.Z SIBBS: reg.sas or reg.zip
Categorical Data Analysis Using the SAS System (22320)
 WWW: <http://www.sas.com/samples/> file: A55320 or A55320.Z
 ftp: pub/publications/ file: A55320 or A55320.Z SIBBS: catdat.sas
Logistic Regression Examples Using the SAS System, Version 6, 1st Edition (55201)
 WWW: <http://www.sas.com/samples/> file: A55201 or A55201.Z
 ftp: pub/publications/ file: A55201 or A55201.Z SIBBS: logistex.sas
Survival Analysis Using the SAS System: A Practical Guide (55233)
 WWW: <http://www.sas.com/samples/> file: A55233
 ftp: pub/publications/ file: A55233 SIBBS: survival.sas
SAS System for Statistical Graphics, 1st Edition (56143)
 WWW: <http://www.sas.com/techsup/download/stat/> program file: statgraf.sas or statgraf.zip data file: statgdat.sas or statgdat.zip
 ftp: techsup/download/stat/ program file: statgraf.sas or statgraf.zip data file: statgdat.sas or statgdat.zip SIBBS: program file: statgraf.sas or statgraf.zip data file: statgdat.sas or statgdat.zip
SAS System for Forecasting Time Series, 1986 Edition
 WWW: <http://www.sas.com/techsup/download/stat/> file: forecast.sas or forecast.zip
 ftp: techsup/download/stat/ file: forecast.sas or forecast.zip SIBBS: forecast.sas or forecast.zip

ADX MACROS (QC)

1st available: 6.03. doc.: *SAS/QC Software, Usage & Reference, Version 6, 1st Ed., Volumes 1 and 2*

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

A: The old QC Guide (*SAS/QC Software, Reference, Version 6, 1st Ed.*) doesn't show examples. However, the new QC Guide (*SAS/QC Software, Usage & Reference, Version 6, 1st Ed., Volumes 1 and 2*) does. Also, examples of all QC procedures are in Technical Report P-188.

ADX MENU SYSTEM (QC)

1st available: 6.03 (no doc.) Different interface starting 6.06 (doc.: ADX Menu System for Design of Experiments)

Q: Where is the documentation?

A: The book, "ADX Menu System for Design of Experiments", is available from SAS Institute Book Sales.

Q: What is required?

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 I already have a design, how can I 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 and filling in the resulting window.

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

A: Yes. How you do this depends on whether you have saved the design or not. If you are creating a design, then you can see the aliasing structure by clicking on Examine: Parameters: Examine Aliasing in NOPROMPT mode (click 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, click on the MODEL button to display the "Change Model" window and click on "Second order". Click on File: End to close the "Change Model" window. Now enter PROMPT mode by clicking on Help: Prompt and then click on Examine Effect Estimates. This shows any aliasing involving main effects and 2-way interactions. Note that the response data are not used, but it must exist. Consequently, random data will do. You don't need ADX to get the aliasing structure of a design. Simply run it through 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 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 REG does not accept MODEL terms like A*B, it is easier to use 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|x5q3 / aliasing e; run;
```

PROC ANOVA (STAT)

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

A: Yes. In 6.11 maintenance, the MEANS statement in ANOVA and GLM includes options for testing homogeneity of variances for one-way ANOVA models and will perform Welch's test for differences between group means when the group variances are not assumed to be equal. The Bartlett, Brown-Forsythe, Levene, and O'Brien tests are provided. Example:

```
PROC GLM; CLASS TRT; MODEL Y=TRT;
MEANS TRT/HOVTTEST=LEVENE WELCH;
```

Prior to 6.11 maintenance, the %HOMOVAR macro performs the O'Brien, Brown-Forsythe, Levene, Bartlett, and Welch Anova F-tests for homogeneity of variance in one-way ANOVA settings.

The macro requires base SAS and STAT (V6.06 or later). It is available via anonymous ftp, SIBBS or the Worldwide Web (WWW):

- ☐ User can obtain via WWW by connecting to: <http://www.sas.com/techsup/download/stat/> and getting homovar.sas.
- ☐ User can obtain via anonymous ftp: ftp ftp.sas.com get techsup/download/stat/homovar.sas
- ☐ User can get via SIBBS: download homovar.sas from the STAT area.

To use, see documentation in header or see SUGI17 (1992) pp. 1178-1182.

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

A: See TS-282, "Strategies for Performing Multiple Comparisons on Means" by Kendall, SUGI 18, 1993, pp 1283-1289.

PROC ARIMA (ETS)

Q: Is there any additional documentation?

A: The *SAS/ETS Software: Applications Guide 1* 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 the 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 that 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 a way to output the parameter estimates to 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. They contain the same information, however, the data sets are structured differently, so you can choose the one most appropriate for your application needs. The options are OUTEST= and OUTMODEL=. These are documented in either the second edition of the SAS/ETS User's Guide or in Technical Report P-231.

Q: Is there a way to output the statistics of fit to an output data set?

A: As of release 6.07.03, the OUTSTAT= option on the ESTIMATE statement allows you to output the statistics of fit such as AIC, SBC, SSE, and the log likelihood. This is documented in either the second edition of the SAS/ETS User's Guide or in Technical Report P-231.

PROC AUTOREG (ETS)

Q: Can I 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 the second edition of the SAS/ETS User's Guide.

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 6.11). (Suggestions for alternative distributions are welcome.)

Q: Is it possible to fit a Cochrane-Orcutt model?

A: No. The Cochrane-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 (STAT)

Q: Is multigroup analysis supported?

A: No. It is planned for V7.

PROC CANDISC (STAT)

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

A: CANDISC (or 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 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 (QC)

doc: *SAS/QC Software, Usage & Reference, Version 6, 1st Ed., Volumes 1 and 2*

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

A: The old QC Guide (*SAS/QC Software, Reference, Version 6, 1st Ed.*) doesn't show examples. However, the new QC Guide (*SAS/QC Software, Usage & Reference, Version 6, 1st Ed., Volumes 1 and 2*) does. Also, examples of all QC procedures are in Technical Report P-188.

Q: Is there any additional documentation?

A: Yes. TS-291: "Process Capability Analysis Using SAS/QC Software, Release 6.08" (the orange booklet) contains documentation for new 6.08 features including the `INSET` and `COMPHISTOGRAM` statements.

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. Or you can use SAS/INSIGHT to overlay multiple curves.

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 (STAT)

N: Additional documentation from SAS Institute: "Categorical Data Analysis Using the SAS System" discusses the CATMOD, LOGISTIC and FREQ procedures at length and is available from Book Sales. Also, the Categorical Data Analysis course notes are available from Book Sales. These notes were not designed to be used by themselves, but they are done pretty well and may be useful on their own.

N: Recommended references: Agresti (1990) "Categorical Data Analysis", published by Wiley is 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.

N: Out of memory or taking a lot of time: 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: Backward parameter estimate signs — why? How can I fix it?

A: 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, 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: Which Parameter estimate goes with which independent var level? Why aren't there as many parms as levels of my independent var?

A: See Example 1 in PROC CATMOD. 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 parm 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 $(-(\text{parm1}+\text{parm2}+\text{parm3}))$.

Q: Infinite parameters or redundant effects message — What do I do?

A: These happen when some parameters of the model are detected to be infinite (see middle of page 470 in Vol. 1). 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 parms commonly happen when there are zeros in the table that 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: This 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 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 that 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 (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, Fourth Edition*. 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 the

- ☐ 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 thinks that all raw data are continuous. If you want to cluster binary data, create a distance matrix (you may be able to use the %DISTANCE macro to do this) as you think it should be defined, and 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 FASTCLUS, cluster these preliminary clusters with CLUSTER, and then 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 FASTCLUS; the other uses DISCRIM.

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.

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 CLUSTER. See example 5 in the DISCRIM chapter.

PROC CORRESP (STAT)

Q: How to fix the error message: "DIMENS=2 is too large for an 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 documentation.

Q: Do we 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 can produce either printer plots or hi-res plots with SAS/GRAPH. There are two versions of %PLOTIT: one for release 6.08 and an enhanced one for 6.10 and later. %EQUATE and %PLOTIT are both available via anonymous ftp, SIBBS or the Web:

- ☐ User can obtain from 6.08 (or later) STAT sample library, members equate.sas (for %EQUATE) or crsplot.sas (for %PLOTIT).
- ☐ User can obtain via WWW by connecting to: <http://www.sas.com/techsup/download/stat/> and getting equate.sas (for %EQUATE), crsp608.sas or crsplot.sas (for 6.08 and 6.10 and later versions of %PLOTIT, respectively).
- ☐ User can get these files via anonymous ftp:
ftp ftp.sas.com
get techsup/download/stat/equate.sas
get techsup/download/stat/crsp608.sas
get techsup/download/stat/crsplot.sas
- ☐ User can get via SIBBS: download equate.sas, crsp608.sas or crsplot.sas from the STAT area.

To use, see documentation in header. For %PLOTIT, send TS-259 which discusses and illustrates its use, or equivalently, refer to 4Q94 Observations, pg 23.

PROC CUSUM (QC)

doc: SAS/QC Software, Usage & Reference, Version 6, 1st Ed., Volumes 1 and 2

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

A: The old QC Guide (SAS/QC Software, Reference, Version 6, 1st Ed.) doesn't show examples. However, the new QC Guide (SAS/QC Software, Usage & Reference, Version 6, 1st Ed., Volumes 1 and 2) does. Also, examples of all QC procedures are in Technical Report P-188.

PROC DISCRIM (STAT)

Q: How do I display or 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' (like Vol 1 of SAS/STAT, page 762, bottom). When METHOD=NPART, no functions are used and therefore they cannot be displayed or output. With METHOD=NPART, 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 682-3 of Vol 1 SAS/STAT.

Q: Why are there 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 they insist on getting 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.

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 the user does not have to do this himself unless he doesn't want to use SAS. PROC DISCRIM Example 5 on pages 760-761 (example 2, page 330 in V5 book) shows how to use the TESTDATA= option to let the procedure do the work. The general method is also discussed on page 698.

Q: But why do 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 best discriminators)?

A: Use STEPDISC to help select the best discriminating variables. Once a set of variables is selected, CANDISC (or DISCRIM with the CANONICAL option) can tell you how important each variable is via the structure coefficients. The discriminant function coefficients from 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 CATMOD or LOGISTIC, or if discriminant analysis is required, try the nonparametric methods available with METHOD=NPART. Then use either K= for the k-nearest neighbor method, or R= for the kernel density method.

DISTANCE macro

There are currently three versions of %DISTANCE macro:
DISTANCE.SAS: 6.09 and 6.10 SAS/STAT sample libraries; doc:
internal DISTNEW.SAS: 6.11 SAS/STAT sample library; doc:
internal TS-404 DIST611.SAS: available by download; doc:
internal & TS-404

The DISTANCE macros found in DISTNEW.SAS and DIST611.SAS are newer and recommended. Basic syntax information is in the header, but not enough to really take advantage of the macro. The additional documentation (TS-404) must be sent by mail or retrieved via SIBBS or FTP and printed on a postscript printer.

Q: What does the macro do?

A: The macro computes 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 with DISTNEW.SAS and DIST611.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 does it 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 it?

A: DIST611.SAS is newest and preferred. It is documented in the file DIST611.PS which requires a postscript printer or viewer. Older versions appear in the 6.09, 6.10 and 6.11 SAS/STAT sample libraries called DISTANCE.SAS or DISTNEW.SAS. The XMACRO.SAS file is required to run the version in the 6.10 and 6.11 sample libraries.

- User can get from 6.11 (or later) STAT sample library, members dist611.sas and xmacro.sas (optionally stdize.sas).
- User can obtain via WWW by connecting to:
<http://www.sas.com/techsup/download/stat/> and getting dist611.ps, dist611.sas and xmacro.sas (optionally stdize.sas)
- User can get these files via anonymous ftp:
ftp ftp.sas.com
get techsup/download/stat/dist611.ps
get techsup/download/stat/dist611.sas
get techsup/download/stat/xmacro.sas
(and optionally, if want to standardize variables):
get techsup/download/stat/stdize.sas
- User can get via SIBBS: download dist611.ps, dist611.sas and xmacro.sas (optionally stdize.sas) from the STAT area.

To use %DISTANCE, see the header for basic syntax or the documentation in the postscript file, dist611.ps. Or see TS-404.

PROC EXPAND (ETS)

Q: I need to compute a moving average of a variable in 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 second edition of the SAS/ETS User's Guide. 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 the publication, *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 (QC)

doc: SAS/QC Software, Usage & Reference, Version 6, 1st Ed., Volumes 1 and 2

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

A: The old QC Guide (*SAS/QC Software, Reference, Version 6, 1st Ed.*) doesn't show examples. However, the new QC Guide (*SAS/QC Software, Usage & Reference, Version 6, 1st Ed., Volumes 1 and 2*) does. Also, examples of all QC procedures are in Technical Report P-188.

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, like this: data outplan; set outplan; random=ranuni(123); run; proc sort; by random; run;

Q: FACTEX didn't find the "best" (in terms of resolution or aberration). Why not and how can I get it to find the "best" design?

A: See SAS Note V6-FACTEX-A650.

PROC FACTOR (STAT)

Q: Is there a good "how to" reference for factor analysis in general?

A: See "The Application and Misapplication of Factor Analysis in Marketing Research" by David W. Stewart from the *Journal of Marketing Research* Vol. 18 (February 1981), pp 51-62. This is an excellent article explaining when to use factor analysis as well as when NOT to use factor analysis, rotations, assumptions, and interpretations. The article is very useful and not difficult to read. See also the book "A Step-By-Step Approach to Using the SAS System for Factor analysis and Structural Equation Modeling" by Larry Hatcher, available from SAS Institute Book Sales.

Q: How are the standardized scoring coefficients computed and how are they used?

A: These are simply standardized regression coefficients for predicting the factors from the variables, and the computation is just the usual formula for computing standardized regression coefficients from correlation matrices. The standardized scoring coefficients are computed by multiplying the inverse of the variable correlation matrix (even if the COV option is used) by the structure matrix if a non-orthogonal rotation is specified, or else by the pattern matrix. PROC FACTOR will do all of this automatically if you specify the OUT= and N= options on the PROC statement. Unstandardized scoring coefficients are not available with PROC FACTOR.

Q: How are the estimated factor scores in the OUT= data set computed? How can I compute them myself, or on a second data set?

A: For every observation used in the analysis, PROC FACTOR will multiply the standardized data vector by the standardized scoring coefficients. If you want to do any of this manually, or score a new data set based on the original factors, then just use PROC SCORE DATA= the new data set, SCORE= the OUTSTAT data set from FACTOR, type='SCORE', and OUT= a data set to contain the estimated factor scores. Be sure to use a VAR statement if appropriate.

Q: Can I do a Q-mode factor analysis?

A: All you have to do for a Q mode principal component analysis is to transpose the data set. If any standardization is required to make the variables comparable, it should be done with PROC STANDARD before transposing, and you may want to use the COV and/or NOINT options in FACTOR/PRINCOMP depending on the purpose of the analysis.

PROC FASTCLUS (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). So, 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. Use PROC STANDARD with the S= option to do this.

FORECASTING MENU SYSTEM (ETS)

1st available: exp'tal: 6.07.03. production 6.11 TS020. doc.: SAS/ETS Software: Time Series Forecasting System, Version 6, 1st Edition *** New interface starting with 6.10 ***

Q: (6.07-6.09) How do I start it up?

A: 1) Under Assist, select Planning tools, then Forecasting, or 2) type FORECAST on the command line, or 3) via pmenus: Globals->Invoke Application->Forecasting

Q: (6.07-6.09) Where is it 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: (6.07-6.09) Why do I get the message, NOTE: Check ID variable, when I try to specify my input data set?

A: It could be due to several possible reasons:

1. the ID variable is not a SAS date variable.
2. 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.)
3. 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/STAT)

N: Additional documentation from SAS Institute *'Categorical Data Analysis Using the SAS System'* discusses the FREQ, CATMOD and LOGISTIC procedures at length and is available from Book Sales.

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 FREQ chapter in the SAS/STAT guide.

Q: How can I get a confidence interval for $p = \Pr(\text{success})$ in a single binomial sample?

A: Approximate, large-sample confidence intervals and an exact confidence interval for p are computed by EXACTPCI.SAS, available via SIBBS, anonymous ftp or WWW:

- User can obtain via WWW by connecting to:
http://www.sas.com/techsup/download/stat/ and getting exactpci.sas
- User can get it via anonymous ftp:
ftp ftp.sas.com get
techsup/download/stat/exactpci.sas
- User can get it via SIBBS: download file exactpci.sas from STAT area.

Q: How do I test the equality of proportions from two independent samples? (for dependent samples, see below)

A: The Pearson chi-square test, available using the CHISQ option on the TABLES statement of PROC FREQ will do this. For example, say 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	100
FEMALE	40	60	100

and run in 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 compute the power or sample size for a test of two independent proportions, or for the Pearson or Likelihood Ratio Chi-square statistics printed by FREQ?

A: Agresti (*'Categorical Data Analysis'*, Wiley, 1990, pg 239-243) shows how to compute approximate power for a test of two independent proportions, and more generally for the Pearson and Likelihood Ratio statistics. This information can be used before data collection to select a sample size that yields sufficiently low type I and type II error rates or after data collection and analysis to tell you what the power was for the test. Three programs are available:

- POWER2X2.SAS computes power of a test comparing proportions in two independent, equal-size samples for a range of total sample sizes
- PWR2X2UN.SAS computes power of a test comparing two independent proportions for given, possibly unequal, sample sizes

- POWERRXC.SAS contains a macro that computes power of FREQ's Pearson and Likelihood Ratio chi-square tests for a range of total-table sample sizes.

The programs are available via SIBBS, anonymous ftp or WWW:

- User can obtain via WWW by connecting to:
http://www.sas.com/techsup/download/stat/ and getting power2x2.sas, pwr2x2un.sas or powerrxs.sas
- User can get via anonymous ftp:
ftp ftp.sas.com
get techsup/download/stat/power2x2.sas
get techsup/download/stat/pwr2x2un.sas
get techsup/download/stat/powerrxs.sas
- User can get via SIBBS: download file(s) power2x2.sas, pwr2x2un.sas or powerrxs.sas from STAT area.

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

A: Two common situations are:

1. 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.)
2. 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: FREQ does not do any tests on one-way tables. To test that the cell frequencies in a one-way table are equal or follow some distribution that can be specified, see TS-176. It is a paper on how to compute the goodness of fit chi-square test. It 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 6.10 or later, the AGREE option on the TABLES statement computes all of these statistics. In earlier releases, page 129 of the STAT Guide 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 6.10, kappa (and weighted kappa) statistics can be computed with CATMOD, but it is not at all easy. A complex RESPONSE statement must be written for each situation. A SUGI paper (in 1987 proceedings, page 1149, Robert Terry is author) discusses how this is done. The paper is also TS-118.

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 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 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. Because multiple tests are being done when performing multiple comparisons in this way, adjusting for the multiplicity is important. Consider using the MULTTEST procedure to perform multiplicity-adjusted Fisher or Cochran-Armitage tests. Note that 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 row1 than in row2. 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 column1 than in column2. 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 row1 than in row2. 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 column1 than in column2.

FUNCTIONS (BASE - Probability, Random number, Quantile)

Q: Is there a PI function?

A: There is no PI() function, but you can get it's value with either of the following: $\arccos(-1)$ or $\gamma(1/2)^{**2}$. However, the precision is not unlimited. On HP700, 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, $\text{GAMMA}(X)$ is $(X-1)!$. Thus, to find 6!, in a DATA step you would 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 as follows:

```
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 (as discussed above regarding factorials). But this will only work when n is relatively small since the computations can overflow rather quickly. The exact limit before the GAMMA function overflows is machine dependent (for example, it is greater than about 57 (56!) on an IBM mainframe; smaller on a VAX). A more general approach, suitable when the factorials involved are large, is to use the EXP and LGAMMA functions. In general:

$$\frac{A! B!}{C! D!} = \text{EXP}(\text{LGAMMA}(A+1) + \text{LGAMMA}(B+1) - \text{LGAMMA}(C+1) - \text{LGAMMA}(D+1))$$

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:

$$\text{GAMMA}(6) / (\text{GAMMA}(3) * \text{GAMMA}(4)) \text{ or } \text{EXP}(\text{LGAMMA}(6) - \text{LGAMMA}(3) - \text{LGAMMA}(4))$$

Q: How can I get a list of the combinations (or permutations) of n objects taken r at a time?

A: See TS-498 or SAS Communications, 2nd quarter 1992, page 36.

Q: How many random numbers can you generate before it repeats? I.e., what is the period of the random number generators? **A:** $(2^{**31})-1$. See page 1059 of the 1987 SUGI Proceedings for an article on these functions.

GAGE application (QC)

1st avail: 6.10 sample library doc.: SAS/QC Usage & Reference, Vol. 2, pg. 1563

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

A: Yes. The GAGE application is part of the 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, 1994 SUGI Proceedings

Q: What is required?

A: SAS/QC is required. SAS/AF is required if you want to modify the application, it is not needed to run it. SAS/STAT is required for the Variance Components Method.

Q: How do I invoke the application?

A: Issue a LIBNAME statement pointing to the location of the the SAS/QC Sample Library (this is where the gage catalog is stored). For example, if you are using SAS under Windows:

```
libname gage '!sasrootqcsample';
```

Then invoke the application with the following command from any display manager window:

```
af c=gage.gage.gage.frame
```

PROC GENMOD (STAT)

1st available: exp'tal in 6.08 & 6.09 (prod'n in 6.08TS415, 6.09TS039) doc.: P-243

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

Q: Can GENMOD fit 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 TRANSREG and GENMOD if you want to do something like GAMs.

Q: Can GENMOD use GEEs (Generalized Estimating Equations)?

A: No, but someone has written a 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 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 GENMOD fit data using the negative binomial distribution?

A: Yes. There is a user-contributed program that defines a macro to fit a model to negative binomial data using GENMOD. The methodology is discussed in the SUGI19 paper by Hilbe. This program is supported by the author, not by SAS Institute. The author's name and email address are included in the file which is available by anonymous ftp, WWW or SIBBS:

- User can get from 6.10 (*NOT* 6.11) STAT sample library, member genmodx4.sas.
- User can obtain via WWW by connecting to: <http://www.sas.com/techsup/download/stat/> and getting negbinom.sas .
- User can get via anonymous ftp: <ftp://ftp.sas.com>
get techsup/download/stat/negbinom.sas
- User can get via SIBBS: download negbinom.sas from the STAT area.

To use, see documentation in header. An example appears at the end of the file.

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.

GLIMMIX macro

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

Q: What does the macro do?

A: The macro uses iteratively reweighted likelihoods to fit the generalized linear mixed 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 it require?

A: 6.08 SAS/STAT. SAS/IML is optional — the macro can run without IML, but to get certain options to work (like confidence limits on the parameter estimates) IML must be available.

Q: How do I get it?

A: The file GLIMMIX.SAS contains the macro definition; GLIMMIXE.SAS contains examples of its use.

- User can get from 6.10 (or later) STAT sample library, member glimmix.sas.
- User can obtain via WWW by connecting to: <http://www.sas.com/techsup/download/stat/> and getting glimmix.sas and glimmixe.sas .
- User can get via anonymous ftp: <ftp://ftp.sas.com>
get techsup/download/stat/glimmix.sas
get techsup/download/stat/glimmixe.sas
- User can get via SIBBS: download glimmix.sas and glimmixe.sas from the STAT area.

To use, see documentation in header. Examples of use are in glimmixe.sas.

PROC GLM (STAT)

N: Additional references and documentation "SAS System for Linear Models". Also the DETAILS section and the introductory chapters in the SAS/STAT User's Guide.

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=parmests;
```

Prior to 6.07.03, the options you have are:

1. PROC PRINTTO; or
2. Coding your own dummy variables and using 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.

Note: This is not the end of this paper. You may request a complete version of this paper by calling Technical Support at SAS Institute and asking for TS-507. You may also access a complete version of this paper through the Institute's home page, SIBBS, or anonymous FTP.

SAS, SAS/ETS, SAS/INSIGHT, SAS/IML, SAS/LAB, SAS/OR, SAS/QC SAS/STAT, and JMP Software are registered trademarks or trademarks of SAS Institute Inc in the USA and other countries. IBM, OS/2, and 3090 are registered trademarks or trademarks of International Business Machines Corporation. ® indicates USA registration.

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

DATA ENTRY MADE EASY WITH SAS/FSP

PANEL MEMBERS

Priscilla Van Grevenhof
Randy Vrabel
Rita Black
Mary Lichty

Mayo Foundation, Rochester, MN

SAS/FSP has easy to set up and easy to use procedures for entering and editing data. This panel has almost 30 years experience in developing hundreds of data entry systems using PROC FSEDIT and PROC FSVIEW. They will present a technique for use in the development process, describe an simple to develop data entry system, demonstrate some easy to implement advanced features, and share lessons learned from the many projects developed. Time will be available at the end of presentation for discussion.

Outline

1. So you want to enter data into a SAS dataset
 - a. What do you need to know before you begin
 - b. Things to consider
2. The "PDQ" data entry system
 - a. How to build it (demonstration)
 - b. Looks like a custom screen
 - c. Full-screen data entry in rows and columns
3. Now let's add some more features
 - a. Customizing a screen
 - b. Error-checking
 - c. Setting parameters
4. Online data entry from a user perspective
 - a. Helpful hints
 - b. Benefits of user designed system