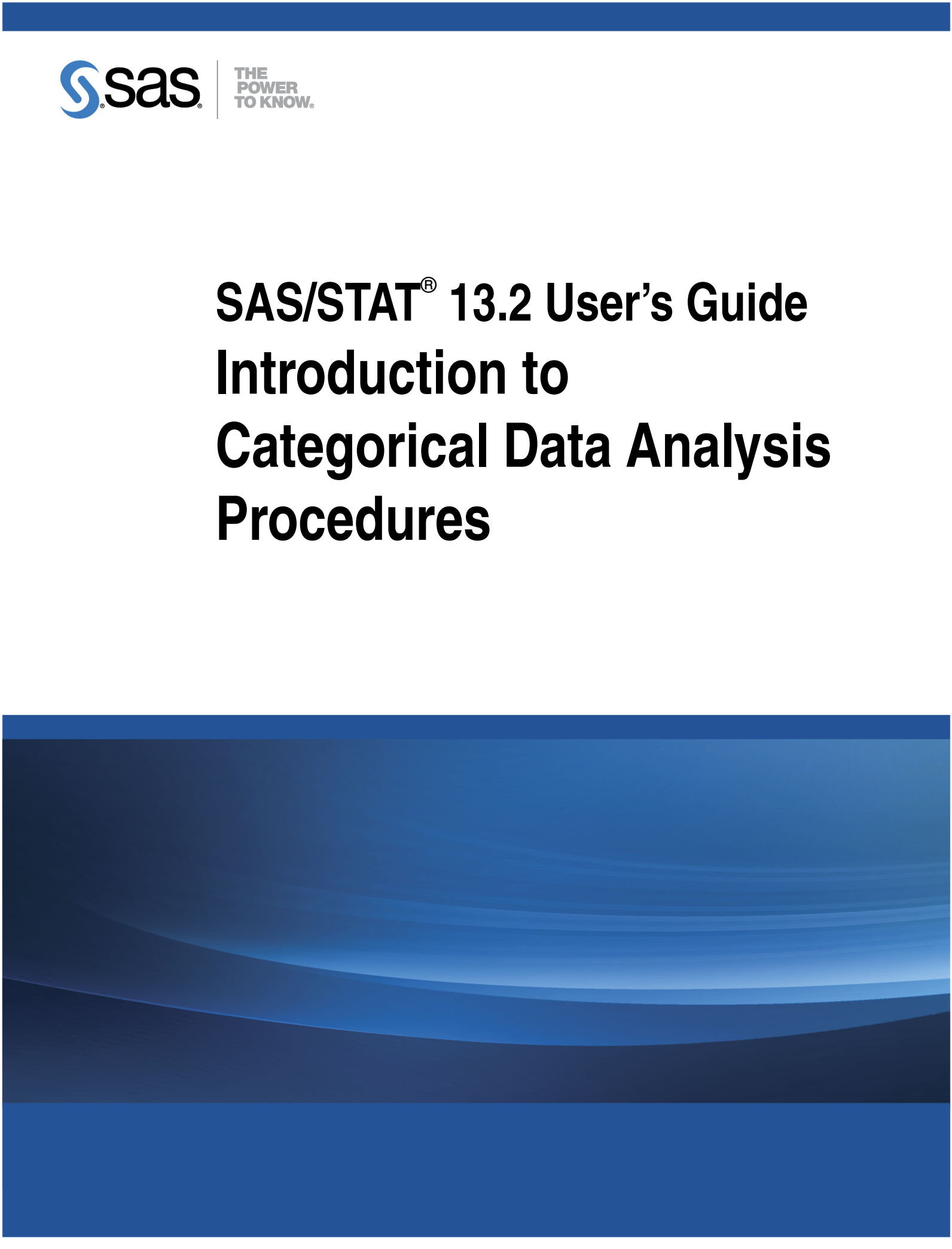


SAS/STAT® 13.2 User's Guide

Introduction to Categorical Data Analysis Procedures

This document is an individual chapter from *SAS/STAT® 13.2 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2014. *SAS/STAT® 13.2 User's Guide*. Cary, NC: SAS Institute Inc.

Copyright © 2014, SAS Institute Inc., Cary, NC, USA

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

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

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

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

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a) and DFAR 227.7202-4 and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

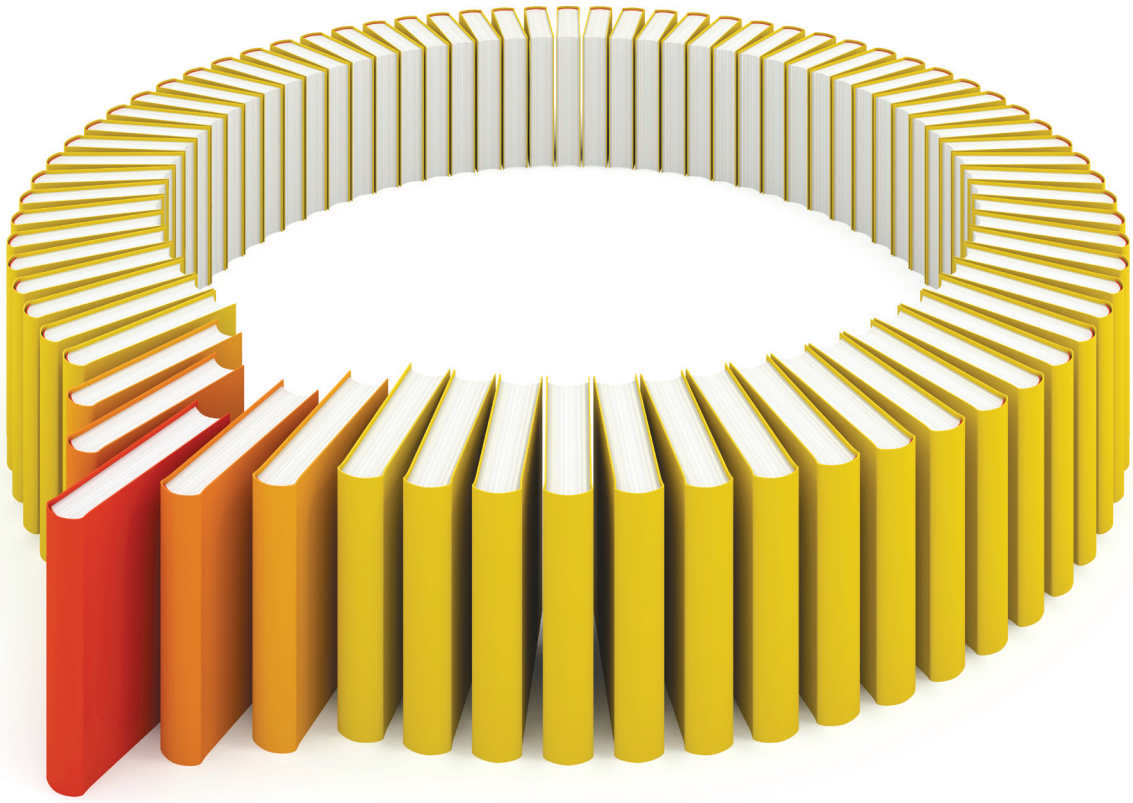
SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

August 2014

SAS provides a complete selection of books and electronic products to help customers use SAS® software to its fullest potential. For more information about our offerings, visit **support.sas.com/bookstore** or call 1-800-727-3228.

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

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



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.

sas
THE POWER TO KNOW®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0613

Chapter 8
Introduction to Categorical Data Analysis Procedures

Contents

Overview: Categorical Data Analysis Procedures **161**
Introduction **163**
Sampling Frameworks and Distribution Assumptions **164**
Simple Random Sampling: One Population 164
Stratified Simple Random Sampling: Multiple Populations 165
Observational Data: Analyzing the Entire Population 166
Randomized Experiments 167
Relaxation of Sampling Assumptions 167
Comparison of PROC FREQ and the Modeling Procedures **168**
Comparison of Modeling Procedures **169**
Logistic Regression 170
References **171**

Overview: Categorical Data Analysis Procedures

There are two approaches to performing categorical data analyses. The first computes statistics based on tables defined by *categorical variables* (variables that assume only a limited number of discrete values), performs hypothesis tests about the association between these variables, and requires the assumption of a randomized process; following Stokes, Davis, and Koch (2000), call these methods *randomization procedures*. The other approach investigates the association by modeling a categorical response variable, regardless of whether the explanatory variables are continuous or categorical; call these methods *modeling procedures*. Several procedures in SAS/STAT software can be used for the analysis of categorical data.

The randomization procedures are:

FREQ builds frequency tables or contingency tables and can produce numerous statistics. For one-way frequency tables, it can perform tests for equal proportions, specified proportions, or the binomial proportion. For contingency tables, it can compute various tests and measures of association and agreement including chi-square statistics, odds ratios, correlation statistics, Fisher's exact test for any size two-way table, kappa, and trend tests. In addition, it performs stratified analysis, computing Cochran-Mantel-Haenszel statistics and estimates of the common relative risk. Exact *p*-values and confidence intervals are available for various test statistics and measures. See Chapter 40, "The FREQ Procedure," for more information.

SURVEYFREQ incorporates complex sample designs to analyze one-way, two-way, and multiway crosstabulation tables. Estimates population totals and proportions and performs tests of goodness-of-fit and independence. See Chapter 14, "Introduction to Survey Procedures," and Chapter 97, "The SURVEYFREQ Procedure," for more information.

The modeling procedures, which require a categorical response variable, are:

CATMOD fits linear models to functions of categorical data, facilitating such analyses as regression, analysis of variance, linear modeling, log-linear modeling, logistic regression, and repeated measures analysis. Maximum likelihood estimation is used for the analysis of logits and generalized logits, and weighted least squares analysis is used for fitting models to other response functions. Iterative proportional fitting (IPF), which avoids the need for parameter estimation, is available for fitting hierarchical log-linear models when there is a single population. See Chapter 32, "The CATMOD Procedure," for more information.

GENMOD fits generalized linear models with maximum-likelihood methods. This family includes logistic, probit, and complementary log-log regression models for binomial data, Poisson and negative binomial regression models for count data, and multinomial models for ordinal response data. It performs likelihood ratio and Wald tests for Type I, Type III, and user-defined contrasts. It analyzes repeated measures data with generalized estimating equation (GEE) methods. Bayesian analysis capabilities for generalized linear models are also available. See Chapter 43, "The GENMOD Procedure," for more information.

GLIMMIX fits generalized linear mixed models with maximum-likelihood methods. If the model does not contain random effects, the GLIMMIX procedure fits generalized linear models by the method of maximum likelihood. This family includes logistic, probit, and complementary log-log regression models for binomial data, Poisson and negative binomial regression models for count data, and multinomial models for ordinal response data. See Chapter 44, "The GLIMMIX Procedure," for more information.

LOGISTIC fits linear logistic regression models for discrete response data with maximum-likelihood methods. It provides four variable selection methods, computes regression diagnostics, and compares and outputs receiver operating characteristic curves. It can also perform stratified conditional logistic regression analysis for binary response data and exact conditional regression analysis for binary and nominal response data. The logit link function in the logistic regression models can be replaced by the probit function or the complementary log-log function. See Chapter 60, "The LOGISTIC Procedure," for more information.

PROBIT fits models with probit, logit, or complementary log-log links for quantal assay or other discrete event data. It is mainly designed for dose-response analysis with a natural response rate. It computes the fiducial limits for the dose variable and provides various graphical displays for the analysis. See Chapter 81, "The PROBIT Procedure," for more information.

SURVEYLOGISTIC fits logistic models for binary and ordinal outcomes to survey data by maximum likelihood, incorporating complex survey sample designs. See Chapter 98, "The SURVEYLOGISTIC Procedure," for more information.

Also see Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software," and Chapter 4, "Introduction to Regression Procedures," for more information about all the modeling and regression procedures.

Other procedures that can be used for categorical data analysis and modeling are:

CORRESP performs simple and multiple correspondence analyses, using a contingency table, Burt table, binary table, or raw categorical data as input. See Chapter 9, "Introduction to Multivariate Procedures," and Chapter 34, "The CORRESP Procedure," for more information.

PRINQUAL performs a principal component analysis of qualitative and/or quantitative data, and multidimensional preference analysis. See Chapter 9, "Introduction to Multivariate Procedures," and Chapter 80, "The PRINQUAL Procedure," for more information.

TRANSREG fits univariate and multivariate linear models, optionally with spline and other nonlinear transformations. Models include ordinary regression and ANOVA, multiple and multivariate regression, metric and nonmetric conjoint analysis, metric and nonmetric vector and ideal point preference mapping, redundancy analysis, canonical correlation, and response surface regression. See Chapter 4, "Introduction to Regression Procedures," and Chapter 104, "The TRANSREG Procedure," for more information.

Introduction

A *categorical variable* is a variable that assumes only a limited number of discrete values. The measurement scale for a categorical variable is unrestricted. It can be *nominal*, which means that the observed levels are not ordered. It can be *ordinal*, which means that the observed levels are ordered in some way. Or it can be *interval*, which means that the observed levels are ordered and numeric and that any interval of one unit on the scale of measurement represents the same amount, regardless of its location on the scale. One example of a categorical variable is litter size; another is the number of times a subject has been married. A variable that lies on a nominal scale is sometimes called a *qualitative* or *classification variable*.

Categorical data result from observations on multiple subjects where one or more categorical variables are observed for each subject. If there is only one categorical variable, then the data are generally represented by a *frequency table*, which lists each observed value of the variable and its frequency of occurrence.

If there are two or more categorical variables, then a subject's *profile* is defined as the subject's observed values for each of the variables. Such categorical data can be represented by a frequency table that lists each observed profile and its frequency of occurrence.

If there are exactly two categorical variables, then the data are often represented by a two-dimensional *contingency table*, which has one row for each level of variable 1 and one column for each level of variable 2. The intersections of rows and columns, called *cells*, correspond to variable profiles, and each cell contains the frequency of occurrence of the corresponding profile.

If there are more than two categorical variables, then the data can be represented by a *multidimensional contingency table*. There are two commonly used methods for displaying such tables, and both require that the variables be divided into two sets.

- In the first method, one set contains a row variable and a column variable for a two-dimensional contingency table, and the second set contains all of the other variables. The variables in the second set are used to form a set of profiles. Thus, the data are represented as a series of two-dimensional contingency tables, one for each profile. This is the data representation used by PROC FREQ. For

example, if you request tables for RACE*SEX*AGE*INCOME, the FREQ procedure represents the data as a series of contingency tables: the row variable is AGE, the column variable is INCOME, and the combinations of levels of RACE and SEX form a set of profiles.

- In the second method, one set contains the independent variables, and the other set contains the dependent variables. Profiles based on the independent variables are called *population profiles*, whereas those based on the dependent variables are called *response profiles*. A two-dimensional contingency table is then formed, with one row for each population profile and one column for each response profile. Since any subject can have only one population profile and one response profile, the contingency table is uniquely defined. This is the data representation used by the modeling procedures.

NOTE: Modeling procedures for categorical data analysis only require that the response variable be categorical—the explanatory variables are allowed to be continuous or categorical. However, note that PROC CATMOD was designed to handle contingency table data, and it does not efficiently handle continuous covariates.

Sampling Frameworks and Distribution Assumptions

This section discusses the sampling frameworks and distribution assumptions for the modeling and randomization procedures.

Simple Random Sampling: One Population

Suppose you take a simple random sample of 100 people and ask each person the following question, "Of the three colors red, blue, and green, which is your favorite?" You then tabulate the results in a frequency table as shown in Table 8.1.

Table 8.1 One-Way Frequency Table

	Favorite Color			
	Red	**Blue**	**Green**	**Total**
Frequency	52	31	17	100
Proportion	0.52	0.31	0.17	1.00

In the population you are sampling, you assume there is an unknown probability that a population member, selected at random, would choose any given color. In order to estimate that probability, you use the sample proportion

$$p_j = \frac{n_j}{n}$$

where n_j is the frequency of the jth response and n is the total frequency.

Because of the random variation inherent in any random sample, the frequencies have a probability distribution representing their relative frequency of occurrence in a hypothetical series of samples. For a simple random

sample, the distribution of frequencies for a frequency table with three levels is as follows. The probability that the first frequency is n_1, the second frequency is n_2, and the third is $n_3 = n - n_1 - n_2$, is given by

$$\Pr(n_1, n_2, n_3) = \frac{n!}{n_1!n_2!n_3!}\pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3}$$

where π_j is the true probability of observing the jth response level in the population.

This distribution, called the *multinomial distribution*, can be generalized to any number of response levels. The special case of two response levels is called the *binomial distribution*.

Simple random sampling is the type of sampling required by the (non-survey) modeling procedures when there is one population. The modeling procedures use the multinomial distribution to estimate a probability vector and its covariance matrix. If the sample size is sufficiently large, then the probability vector is approximately normally distributed as a result of central limit theory. This result is used to compute appropriate test statistics for the specified statistical model.

Stratified Simple Random Sampling: Multiple Populations

Suppose you take two simple random samples, 50 men and 50 women, and ask the same question as before. You are now sampling two different populations that may have different response probabilities. The data can be tabulated as shown in Table 8.2.

Table 8.2 Two-Way Contingency Table: Sex by Color

	Favorite Color			
Sex	**Red**	**Blue**	**Green**	**Total**
Male	30	10	10	50
Female	20	10	20	50
Total	50	20	30	100

Note that the row marginal totals (50, 50) of the contingency table are fixed by the sampling design, but the column marginal totals (50, 20, 30) are random. There are six probabilities of interest for this table, and they are estimated by the sample proportions

$$p_{ij} = \frac{n_{ij}}{n_i}$$

where n_{ij} denotes the frequency for the ith population and the jth response and n_i is the total frequency for the ith population. For this contingency table, the sample proportions are shown in Table 8.3.

Table 8.3 Table of Sample Proportions by Sex

	Favorite Color			
Sex	**Red**	**Blue**	**Green**	**Total**
Male	0.60	0.20	0.20	1.00
Female	0.40	0.20	0.40	1.00

The probability distribution of the six frequencies is the *product multinomial distribution*

$$\Pr(n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}) = \frac{n_1! n_2! \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{13}^{n_{13}} \pi_{21}^{n_{21}} \pi_{22}^{n_{22}} \pi_{23}^{n_{23}}}{n_{11}! n_{12}! n_{13}! n_{21}! n_{22}! n_{23}!}$$

where π_{ij} is the true probability of observing the jth response level in the ith population. The product multinomial distribution is simply the product of two or more individual multinomial distributions since the populations are independent. This distribution can be generalized to any number of populations and response levels.

Stratified simple random sampling is the type of sampling required by the modeling procedures when there is more than one population. The product multinomial distribution is used to estimate a probability vector and its covariance matrix. If the sample sizes are sufficiently large, then the probability vector is approximately normally distributed as a result of central limit theory, and this result is used to compute appropriate test statistics for the specified statistical model. The statistics are known as Wald statistics, and they are approximately distributed as chi-square when the null hypothesis is true.

Observational Data: Analyzing the Entire Population

Sometimes the observed data do not come from a random sample but instead represent a complete set of observations on some population. For example, suppose a class of 100 students is classified according to sex and favorite color. The results are shown in Table 8.4.

In this case, you could argue that all of the frequencies are fixed since the entire population is observed; therefore, there is no sampling error. On the other hand, you could hypothesize that the observed table has only fixed marginals and that the cell frequencies represent one realization of a conceptual process of assigning color preferences to individuals. The assignment process is open to hypothesis, which means that you can hypothesize restrictions on the joint probabilities.

Table 8.4 Two-Way Contingency Table: Sex by Color

	Favorite Color			
Sex	**Red**	**Blue**	**Green**	**Total**
Male	16	21	20	57
Female	12	20	11	43
Total	28	41	31	100

The usual hypothesis (sometimes called *randomness*) is that the distribution of the column variable (Favorite Color) does not depend on the row variable (Sex). This implies that, for each row of the table, the assignment process corresponds to a simple random sample (without replacement) from the finite population represented by the column marginal totals (or by the column marginal subtotals that remain after sampling other rows). The hypothesis of randomness implies that the probability distribution on the frequencies in the table is the *hypergeometric distribution*.

If the same row and column variables are observed for each of several populations, then the probability distribution of all the frequencies can be called the *multiple hypergeometric distribution.* Each population is called a *stratum*, and an analysis that draws information from each stratum and then summarizes across

them is called a *stratified analysis* (or a *blocked analysis* or a *matched analysis*). PROC FREQ does such a stratified analysis, computing test statistics and measures of association.

In general, the populations are formed on the basis of cross-classifications of independent variables. Stratified analysis is a method of adjusting for the effect of these variables without being forced to estimate parameters for them. Note that PROC LOGISTIC can perform analyses on stratified tables as well, using the usual modeling procedure assumptions, by using conditional or exact conditional logistic regression.

The multiple hypergeometric distribution is the one used by PROC FREQ for the computation of Cochran-Mantel-Haenszel statistics. These statistics are in the class of *randomization model test statistics*, which require minimal assumptions for their validity. PROC FREQ uses the multiple hypergeometric distribution to compute the mean and the covariance matrix of a function vector in order to measure the deviation between the observed and expected frequencies with respect to a particular type of alternative hypothesis. If the cell frequencies are sufficiently large, then the function vector is approximately normally distributed as a result of central limit theory, and PROC FREQ uses this result to compute a quadratic form that has a chi-square distribution when the null hypothesis is true.

Randomized Experiments

Consider a *randomized experiment* in which patients are assigned to one of two treatment groups according to a randomization process that allocates 50 patients to each group. After a specified period of time, each patient's status (cured or not cured) is recorded. Suppose the data shown in Table 8.5 give the results of the experiment. The null hypothesis is that the two treatments are equally effective. Under this hypothesis, treatment is a randomly assigned label that has no effect on the cure rate of the patients. But this implies that each row of the table represents a simple random sample from the finite population whose cure rate is described by the column marginal totals. Therefore, the column marginals (58, 42) are fixed under the hypothesis. Since the row marginals (50, 50) are fixed by the allocation process, the hypergeometric distribution is induced on the cell frequencies. Randomized experiments can also be specified in a stratified framework, and Cochran-Mantel-Haenszel statistics can be computed relative to the corresponding multiple hypergeometric distribution.

Table 8.5 Two-Way Contingency Table: Treatment by Status

	Status		
Treatment	**Cured**	**Not Cured**	**Total**
1	36	14	50
2	22	28	50
Total	58	42	100

Relaxation of Sampling Assumptions

As indicated previously, the modeling procedures assume that the data are from a stratified simple random sample, so they use the product multinomial distribution. If the data are not from such a sample, then in many cases it is still possible to use a modeling procedure by arguing that each row of the contingency table *does* represent a simple random sample from some hypothetical population. The extent to which the inferences are

generalizable depends on the extent to which the hypothetical population is perceived to resemble the target population.

Similarly, the Cochran-Mantel-Haenszel statistics use the multiple hypergeometric distribution, which requires fixed row and column marginal totals in each contingency table. If the sampling process does not yield a table with fixed margins, then it is usually possible to fix the margins through conditioning arguments similar to the ones used by Fisher when he developed the Exact Test for 2×2 tables. In other words, if you want fixed marginal totals, you can generally make your analysis conditional on those observed totals.

For more information on sampling models for categorical data, see Bishop, Fienberg, and Holland (1975, Chapter 13) and Agresti (2002, Chapter 1.2).

Comparison of PROC FREQ and the Modeling Procedures

PROC FREQ is used primarily to investigate the relationship between two variables; any confounding variables are taken into account by stratification rather than by parameter estimation. Modeling procedures are used to investigate the relationship among many variables, all of which are integrated into a parametric model.

When a modeling procedure estimates the covariance matrix of the frequencies, it assumes that the frequencies were obtained by a stratified simple random-sampling procedure. However, some modeling procedures can handle different sampling methods. PROC CATMOD can analyze input data that consists of a function vector and a covariance matrix, so you can estimate the covariance matrix of the frequencies in the appropriate manner before modeling the data. PROC SURVEYLOGISTIC can analyze data from a completely different, but known, sampling scheme.

For the FREQ procedure, Fisher's Exact Test and Cochran-Mantel-Haenszel (CMH) statistics are based on the hypergeometric distribution, which corresponds to fixed marginal totals. However, by conditioning arguments, these tests are generally applicable to a wide range of sampling procedures. Similarly, the Pearson and likelihood-ratio chi-square statistics can be derived under a variety of sampling situations.

PROC FREQ can do some traditional nonparametric analysis (such as the Kruskal-Wallis test and Spearman's correlation) since it can generate rank scores internally. Fisher's Exact Test and the CMH statistics are also inherently nonparametric. However, the main vehicle for nonparametric analyses in the SAS System is the NPAR1WAY procedure.

A large sample size is required for the validity of the chi-square distributions, the standard errors, and the covariance matrices for PROC FREQ and the modeling procedures. If sample size is a problem, then PROC FREQ has the advantage with its CMH statistics because it does not use any degrees of freedom to estimate parameters for confounding variables. In addition, PROC FREQ can compute exact *p*-values for any two-way table, provided that the sample size is sufficiently small in relation to the size of the table. It can also produce exact *p*-values for many tests, including the test of binomial proportions, the Cochran-Armitage test for trend, and the Jonckheere-Terpstra test for ordered differences among classes. PROC LOGISTIC can perform exact conditional logistic regression and Firth's penalized-likelihood regression to compensate for small sample sizes.

See the procedure chapters for more information. In addition, some well-known texts that deal with analyzing categorical data are listed in the "References" section of this chapter.

Comparison of Modeling Procedures

The CATMOD, GENMOD, GLIMMIX, LOGISTIC, PROBIT, and SURVEYLOGISTIC procedures can all be used for statistical modeling of categorical data.

The CATMOD procedure treats all explanatory (independent) variables as classification variables by default, and you specify continuous covariates in the DIRECT statement. The other procedures treat covariates as continuous by default, and you specify the classification variables in the CLASS statement.

The CATMOD procedure provides weighted least squares estimation of many response functions, such as means, cumulative logits, and proportions, and you can also compute and analyze other response functions that can be formed from the proportions corresponding to the rows of a contingency table. In addition, a user can input and analyze a set of response functions and user-supplied covariance matrix with weighted least squares. PROC CATMOD also provides maximum likelihood estimation for binary and polytomous logistic regression.

The GENMOD procedure is also a general statistical modeling tool which fits generalized linear models to data; it fits several useful models to categorical data including logistic regression, the proportional odds model, and Poisson and negative binomial regression for count data. The GENMOD procedure also provides a facility for fitting generalized estimating equations to correlated response data that are categorical, such as repeated dichotomous outcomes. The GENMOD procedure fits models using maximum likelihood estimation. PROC GENMOD can perform Type I and Type III tests, and it provides predicted values and residuals. Bayesian analysis capabilities for generalized linear models are also available.

The GLIMMIX procedure fits many of the same models as the GENMOD procedure but also allows the inclusion of random effects. The GLIMMIX procedure fits models using maximum likelihood estimation.

The LOGISTIC procedure is specifically designed for logistic regression. It performs the usual logistic regression analysis for dichotomous outcomes and it fits the proportional odds model and the generalized logit model for ordinal and nominal outcomes, respectively, by the method of maximum likelihood. This procedure has capabilities for a variety of model-building techniques, including stepwise, forward, and backward selection. It computes predicted values, the receiver operating characteristics (ROC) curve and the area beneath the curve, and a number of regression diagnostics. It can create output data sets containing these values and other statistics. PROC LOGISTIC can perform a conditional logistic regression analysis (matched-set and case-controlled) for binary response data. For small data sets, PROC LOGISTIC can perform exact conditional logistic regression. Firth's bias-reducing penalized-likelihood method can also be used in place of conditional and exact conditional logistic regression.

The PROBIT procedure is designed for quantal assay or other discrete event data. In additional to performing the logistic regression analysis, it can estimate the threshold response rate. PROC PROBIT can also estimate the values of independent variables that yield a desired response.

The SURVEYLOGISTIC procedure performs logistic regression for binary, ordinal, and nominal responses under a specified complex sampling scheme, instead of the usual stratified simple random sampling.

Stokes, Davis, and Koch (2012) provide substantial discussion of these procedures, particularly the use of the FREQ, LOGISTIC, GENMOD, and CATMOD procedures for statistical modeling.

Logistic Regression

Dichotomous Response

You have many choices of performing logistic regression in the SAS System. The CATMOD, GENMOD, GLIMMIX, LOGISTIC, PROBIT, and SURVEYLOGISTIC procedures fit the usual logistic regression model.

PROC CATMOD might not be efficient when there are continuous independent variables with large numbers of different values. For a continuous variable with a very limited number of values, PROC CATMOD might still be useful.

PROC GLIMMIX enables you to specify random effects in the models; in particular, you can fit a random-intercept logistic regression model.

PROC LOGISTIC provides the capability of model-building and performs conditional and exact conditional logistic regression. It can also use Firth's bias-reducing penalized likelihood method.

PROC PROBIT enables you to estimate the natural response rate and compute fiducial limits for the dose variable.

The LOGISTIC, GENMOD, GLIMMIX, PROBIT, and SURVEYLOGISTIC procedures can analyze summarized data by enabling you to input the numbers of events and trials; the ratio of events to trials must be between 0 and 1.

Ordinal Response

PROC LOGISTIC fits the proportional odds model to the ordinal response data by default, PROC PROBIT fits this model if you specify the logistic distribution, and PROC GENMOD and PROC GLIMMIX fit this model if you specify the CLOGIT link and the multinomial distribution. PROC CATMOD fits the cumulative logit or adjacent-category logit response functions.

Nominal Response

When the response variable is nominal, there is no concept of ordering of the response values. Response functions called *generalized logits* can be fit by the CATMOD, GLIMMIX, and LOGISTIC procedures. PROC CATMOD fits this model by default; PROC GLIMMIX and PROC LOGISTIC require you to specify the GLOGIT link.

Numerical Differences

Differences in the way the models are parameterized and fit might result in different parameter estimates if you perform logistic regression in each of these procedures.

- Parameter estimates from the procedures can differ in sign depending on the ordering of response levels, which you can change if you want.
- The parameter estimates associated with a categorical independent variable might differ among the procedures, since the estimates depend on the coding of the indicator variables in the design matrix. By default, the design matrix column produced by PROC CATMOD and PROC LOGISTIC for a binary independent variable is coded using the values 1 and –1 (deviation from the mean coding,

which is a full-rank parameterization). The same column produced by the CLASS statement of PROC GENMOD, PROC GLIMMIX, and PROC PROBIT is coded using 1 and 0 (GLM coding, which is less-than-full-rank parameterization). As a result, the parameter estimate printed by PROC LOGISTIC is one-half of the estimate produced by PROC GENMOD. Both PROC GENMOD and PROC LOGISTIC allow you to select either a full-rank parameterization or the less-than-full-rank parameterization. The GLIMMIX and PROBIT procedures allow only the less-than-full-rank parameterization for the CLASS variables. The CATMOD procedure allows only full-rank parameterizations. See the "Details" sections in the chapters on the CATMOD, GENMOD, GLIMMIX, LOGISTIC, and PROBIT procedures for more information on the generation of the design matrices used by these procedures. See Chapter 19, "Shared Concepts and Topics," for a general discussion of the various parameterizations.

- The maximum-likelihood algorithm used differs among the procedures. PROC LOGISTIC uses the Fisher's scoring method by default, while PROC PROBIT, PROC GENMOD, PROC GLIMMIX, and PROC CATMOD use the Newton-Raphson method. The parameter estimates should be the same for all three procedures, and the standard errors should be the same for the logistic model. For the normal and extreme-value (Gompertz) distributions in PROC PROBIT, which correspond to the probit and cloglog links, respectively, in PROC GENMOD and PROC LOGISTIC, the standard errors might differ. In general, tests computed using the standard errors from the Newton-Raphson method are more conservative.
- The LOGISTIC, GENMOD, GLIMMIX, and PROBIT procedures can fit a cumulative regression model for ordinal response data by using maximum-likelihood estimation. PROC LOGISTIC and PROC GENMOD use a different parameterization from that of PROC PROBIT, which results in different intercept parameters. Estimates of the slope parameters, however, should be the same for both procedures. The estimated standard errors of the slope estimates are slightly different between the procedures because of the different computational algorithms used as default.

References

Agresti, A. (1984), *Analysis of Ordinal Categorical Data*, New York: John Wiley & Sons.

Agresti, A. (2002), *Categorical Data Analysis*, 2nd Edition, New York: John Wiley & Sons.

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, 2nd Edition, New York: John Wiley & Sons.

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

Collett, D. (2003), *Modelling Binary Data*, 2nd Edition, London: Chapman & Hall.

Cox, D. R. and Snell, E. J. (1989), *The Analysis of Binary Data*, 2nd Edition, London: Chapman & Hall.

Dobson, A. (1990), *An Introduction to Generalized Linear Models*, London: Chapman & Hall.

Fleiss, J. L., Levin, B., and Paik, M. C. (2003), *Statistical Methods for Rates and Proportions*, 3rd Edition, Hoboken, NJ: John Wiley & Sons.

Freeman, D. H., Jr. (1987), *Applied Categorical Data Analysis*, New York: Marcel Dekker.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969), "Analysis of Categorical Data by Linear Models," *Biometrics*, 25, 489–504.

Hosmer, D. W., Jr. and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd Edition, New York: John Wiley & Sons.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd Edition, London: Chapman & Hall.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2000), *Categorical Data Analysis Using the SAS System*, 2nd Edition, Cary, NC: SAS Institute Inc.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2012), *Categorical Data Analysis Using SAS*, 3rd Edition, Cary, NC: SAS Institute Inc.

Index

categorical variable, 163
cell of a contingency table, 163
classification variables, 163
contingency tables, 163

interval variable, 163

nominal variables, 163

ordinal variable, 163

population profile, 164
profile, population and response, 163, 164

qualitative variables, 163

response profile, 164