Chapter 5
Introduction to Categorical Data Analysis Procedures

Chapter Table of Contents

Chapter 5

Introduction to Categorical Data Analysis Procedures

Overview

Several procedures in SAS/STAT software can be used for the analysis of categorical data:

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.

CORRESP performs simple and multiple correspondence analyses, using a contingency table, Burt table, binary table, or raw categorical data as input. For more on PROC CORRESP, see Chapter 6, “Introduction to Multivariate Procedures,” and Chapter 24, “The CORRESP Procedure,”.

FREQ builds frequency tables or contingency tables and produces numerous tests and measures of association including chi-square statistics, odds ratios, correlation statistics, and Fisher’s exact test for any size two-way table. In addition, it performs stratified analysis, computing Cochran-Mantel-Haenszel statistics and estimates of the common relative risk. It performs a test of binomial proportions, computes measures of agreement such as McNemar’s test, kappa, and weighted kappa.

GENMOD fits generalized linear models with maximum-likelihood methods. This family includes logistic, probit, and complementary log-log regression models for binomial data, Poisson 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.

LOGISTIC fits linear logistic regression models for binary or ordinal response data with maximum-likelihood methods. It performs stepwise regression and provides regression diagnostics. The logit link function in the logistic regression models can be replaced by the normit function or the complementary log-log function.

PROBIT	computes maximum-likelihood estimates of regression parameters and optional threshold parameters for binary or ordinal response data.

Other procedures that perform analyses for categorical data are the TRANSREG and PRINQUAL procedures. PROC PRINQUAL is summarized in Chapter 6, "Introduction to Multivariate Procedures," and PROC TRANSREG is summarized in Chapter 3, "Introduction to Regression Procedures."

A *categorical variable* is defined as one that can assume only a limited number of discrete values. The measurement scale for such a 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 PROC CATMOD.

Sampling Frameworks and Distribution Assumptions

This section discusses the sampling frameworks and distribution assumptions for the CATMOD and FREQ 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 5.1.

Table 5.1. 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 $n_3 = n - n_1 - n_2$ where π_j is the true probability of observing the jth response level in the population.

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

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 PROC CATMOD when there is one population. PROC CATMOD uses 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. PROC CATMOD uses this result to compute appropriate test statistics for the specified statistical model.

Stratified Simple Random Sampling: Multiple Populations

Suppose you take two simple random samples, fifty men and fifty 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 5.2.

Table 5.2. 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 5.3.

Table 5.3. 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 PROC CATMOD when there is more than one population. PROC CATMOD uses the product multinomial distribution 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 PROC CATMOD uses this result 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 5.4.

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 5.4. 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 induces a probability distribution on the frequencies in the table; it is called 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.

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 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 fifty patients to each group. After a specified period of time, each patient's status (cured or uncured) is recorded. Suppose the data shown in Table 5.5 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 5.5. Two-Way Contingency Table: Treatment by Status

	Status		
Treatment	Cured	Uncured	Total
1	36	14	50
2	22	28	50
Total	58	42	100

Relaxation of Sampling Assumptions

As indicated above, the CATMOD procedure assumes that the data are from a stratified simple random sample, so it uses the product multinomial distribution. If the data are not from such a sample, then in many cases it is still possible to use PROC CATMOD 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).

Comparison of FREQ and CATMOD 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. PROC CATMOD is used to investigate the relationship among many variables, all of which are integrated into a parametric model.

When PROC CATMOD estimates the covariance matrix of the frequencies, it assumes that the frequencies were obtained by a stratified simple random sampling procedure. However, PROC CATMOD can also analyze input data that consist of a function vector and a covariance matrix. Therefore, if the sampling procedure is different, you can estimate the covariance matrix of the frequencies in the appropriate manner before submitting the data to PROC CATMOD.

For the FREQ procedure, Fisher's Exact Test and Cochran-Mantel-Haenszel 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 Cochran-Mantel-Haenszel 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 both PROC FREQ and PROC CATMOD. 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 the test of binomial proportions, the Cochran-Armitage test for trend, and the Jonckheere-Terpstra test for ordered differences among classes.

See the chapters on the FREQ and CATMOD procedures for more information. In addition, some well-known texts that deal with analyzing categorical data are listed in “References.”

Comparison of CATMOD, GENMOD, LOGISTIC, and PROBIT Procedures

The LOGISTIC, GENMOD, PROBIT, and CATMOD procedures can all be used for statistical modeling of categorical data. The CATMOD procedure provides maximum likelihood estimation for logistic regression, including the analysis of logits for dichotomous outcomes and the analysis of generalized logits for polychotomous outcomes. It provides weighted least squares estimation of many other 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. With the CATMOD procedure, by default, all explanatory (independent) variables are treated as classification variables.

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 regression. The GENMOD procedures 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, and you include classification variables in your models with a CLASS statement. PROC GENMOD can perform type I and type III tests, and it provides predicted values and residuals.

The LOGISTIC procedure is specifically designed for logistic regression. For dichotomous outcomes, it performs the usual logistic regression and for ordinal outcomes, it fits the proportional odds model. Note that any polychotomous response variable will be treated as an ordinal outcome by PROC LOGISTIC. This procedure has capabilities for a variety of model-building techniques, including stepwise, forward, and backwards selection. It produces predicted values and can create output data sets containing these values and other statistics including ROC, and it produces a number of regression diagnostics. The current version does not contain a CLASS statement, so that you have to code classification effects using indicator variables.

The PROBIT procedure is designed for quantal assay or other discrete event data. It performs logistic regression. This procedure includes a CLASS statement.

Stokes, Davis, and Koch (1995) provide substantial discussion of these procedures, particularly the use of the LOGISTIC and CATMOD procedures for statistical modeling.

Logistic Regression

Dichotomous Response

You have many options for performing logistic regression in the SAS System. For the dichotomous outcome, most of the time you would use the LOGISTIC procedure or the GENMOD procedure; you will need to code indicator variables for classification effects in PROC LOGISTIC but can use the CLASS statement in PROC GENMOD. The LOGISTIC procedure provides model-building, so you may choose to use it for that reason. (Note that a future release of PROC LOGISTIC will include a CLASS statement).

You may want to consider the CATMOD procedure for logistic regression since it handles classification variables; however it isn't efficient for this purpose when you have continuous variables with a large number of different values. For a continuous variable with a very limited number of values, PROC CATMOD may be useful. You list the continuous variables in the DIRECT statement.

The PROBIT procedure also performs logistic regression, and the LOGISTIC, GENMOD, and PROBIT procedures allow you to use events/trials input for the responses; the ratio of events to trials must be between 0 and 1.

Ordinal Response

The LOGISTIC and PROBIT procedures treat all response variables with more than two levels as ordinal responses and fit the proportional odds model. The GENMOD procedure fits this model with a link function of CLOGIT and the specification of the multinomial distribution.

Nominal Response

When the response variable is nominal, that is, there is no concept of ordering of the values, you can fit a logistic model to response functions called generalized logits. Only the CATMOD procedure presently performs a generalized logits analysis.

Parameterization

There are some differences in the way that models are parameterized, which means that you might get different parameter estimates if you were to perform logistic regression in each of these procedures.

- Parameter estimates from the procedures may 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 may 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 for a binary independent variable is coded using the values 1 and -1 . The same column produced by the CLASS statement of PROC GENMOD and PROC PROBIT is coded 1 and 0. PROC CATMOD uses fullrank parameterization using differential effects. As a result, the parameter estimate printed by PROC CATMOD is one-half of the estimate produced by the others. PROC LOGISTIC does not automatically create indicator variables for categorical independent variables. So, the parameterization

depends on how you code the indicator variables ($1, 0$ versus $-1, 1$). See the "Details" sections in the chapters on the CATMOD, GENMOD, and PROBIT procedures for more information on the generation of the design matrices used by these procedures.

- The maximum-likelihood algorithm used differs among the procedures. PROC LOGISTIC uses Fisher's scoring method while PROC PROBIT, PROC GENMOD, and PROC CATMOD use the Newton-Raphson method (the PROC PROBIT algorithm is ridge stabilized and is a modified Newton-Raphson algorithm.) 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 (handled by the PROBIT, GENMOD, and LOGISTIC procedures), the standard errors may differ. In general, tests computed using the standard errors from the Newton-Raphson method will be more conservative.
- The LOGISTIC, GENMOD, and PROBIT procedures can fit logistic regression models for ordinal response data 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 two procedures because of the different computational algorithms used.

References

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

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

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

Collett, D. (1991), *Modelling Binary Data*, London: Chapman and Hall.

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

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

Fleiss, J.L. (1981), *Statistical Methods for Rates and Proportions*, Second Edition, New York: John Wiley & Sons, Inc.

Freeman, D.H., (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. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.

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

Stokes, M.E., Davis, C.S., and Koch, G.G (1995), *Categorical Data Analysis Using the SAS System*, Cary NC: SAS Institute Inc.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc.,
SAS/STAT® User's Guide, Version 8, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

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

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.