Paper 346-2011

Up To Speed With Categorical Data Analysis

Maura Stokes, SAS Institute, Inc.
Gary Koch, University of North Carolina, Chapel Hill, NC

Abstract

Categorical data analysis remains a substantial tool in the practice of statistics, and its techniques continue to evolve with time. This paper reviews some of the basic tenets of categorical data analysis today and describes the newer techniques that have become common practice. The use of exact methods has expanded, including additional assessments of statistical hypotheses, conditional logistic regression, and Poisson regression. Bayesian methods are now available for logistic regression and Poisson regression, and graphics are a regular component of many analyses.

This paper describes recent techniques and illustrates them with examples that use SAS/STAT® software.

Introduction

The practice of statistics, like any discipline, evolves over time, and categorical data analysis is no different. Methodological and computing advances give rise to new and better ways of analyzing data in which the responses are dichotomous or counts. Sometimes the statistical approach is the same, such as in examining the difference of binomial proportions and its confidence interval, but newer methodology makes the estimates much improved. Other times, the approach itself is quite different, such as the adoption of Bayesian techniques in place of the more traditional frequentist approach in statistical modeling.

This paper is the answer to the question: What are some of the differences in how categorical data analysis is performed now compared to 10 or 15 years ago? Five fairly different areas are covered: the difference of binomial proportions, Poisson regression and incidence density estimates, conditional logistic regression, nonparametric analysis of covariance, and Bayesian methods.

To review, categorical responses include dichotomous outcomes and discrete counts as well as ordinal outcomes. For example:

- dichotomous: favorable or unfavorable response to a treatment in a clinical trial
- discrete: number of time periods with colds
- ordinal: good, fair, or poor response to a medication for rheumatoid arthritis

When the explanatory variables are discrete, the data are often displayed in a contingency table. The simplest case would be a 2 × 2 display as illustrated in Table 1:

Table 1 Respiratory Treatment Outcomes

Treatment	Favorable	Unfavorable	Total
Test	10	2	12
Control	2	4	6
Total	12	6	18

Categorical data analysis strategies generally take one of two forms:

- hypothesis testing with randomization methods to assess association
- statistical modeling techniques that seek to describe the nature of the association, using maximum likelihood, Bayesian, or weighted least squares modeling methods

As an illustration of the first strategy, consider a multi-center study in which the null hypothesis is whether there is no association between treatment and outcome, controlling for the effect of center. The null hypothesis implies the hypergeometric distribution, and that serves as the basis for exact tests or expected values and covariance structure for approximate tests. As an illustration of the second strategy, consider performing logistic regression via maximum likelihood to assess the influence of that treatment on the outcome, adjusting for center as well as other covariates.

The techniques described in the following sections cover both types of strategies.

Difference of Binomial Proportions

While we are bombarded with news about the contemporary data flood, with megabytes and gigabytes and terabytes, oh my, the simple 2 × 2 table remains a stalwart of categorical data analysis, although the methods for analyzing it have been refined over time. Analyzing the data presented in Table 1 addresses the question of whether patients receiving a treatment are more likely to have a favorable outcome than those who do not. The rows in the table might represent simple random samples from two groups that produce two independent binomial distributions for a binary response. Or, they might represent randomization of patients to two equivalent treatments, resulting in the hypergeometric distribution.

With adequate sample size, say a total count of greater than 20, and individual cell counts of at least 5 in each cell, asymptotically based association statistics, such as the continuity-corrected Pearson chi-square, are appropriate statistics to assess the association at the 0.05 significance level. For the data in Table 1, Fisher's exact test is more suitable since the overall sample size is 18 and 3 cell counts are less than 5. The p-value for this test is 0.0573.

You might also be interested in describing the association in the table by comparing the proportion of favorable response in each treatment group. Consider the following representation of this table:

Table 2 2×2 Contingency Table

	Favorable	Unfavorable	Total	Proportion Favorable
Test	n_{11}	n_{12}	n_{1+}	$p_1 = n_{11}/n_{1+}$
Control	n_{21}	n_{22}	n_{2+}	$p_2 = n_{21}/n_{2+}$
Total	n_{+1}	n_{+2}	n	

The proportion p_i of favorable events in the ith row is defined as n_{i1}/n_{i+}. When the groups represent simple random samples, and the difference $d = p_1 - p_2$ for the proportions p_1 and p_2,

$$E\{p_1 - p_2\} = \pi_1 - \pi_2$$

The variance is

$$V\{p_1 - p_2\} = \frac{\pi_1(1-\pi_1)}{n_{1+}} + \frac{\pi_2(1-\pi_2)}{n_{2+}}$$

for which a consistent estimate is

$$v_d = \frac{p_1(1-p_1)}{n_{1+}} + \frac{p_2(1-p_2)}{n_{2+}}$$

A $100(1-\alpha)$% confidence interval for $(\pi_1 - \pi_2)$ with continuity correction is written as

$$d \pm \left\{ z_{\alpha/2}\sqrt{v_d} + \frac{1}{2}\left\{ \frac{1}{n_{1+}} + \frac{1}{n_{2+}} \right\} \right\}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution; this confidence interval is based on Fleiss, Levin, and Paik (2003). These confidence limits include a continuity adjustment to the Wald asymptotic confidence limits that adjust for the difference between the normal approximation and the discrete binomial distribution.

The default Wald asymptotic confidence limits (without the continuity correction) in the FREQ procedure are appropriate for larger sample sizes, say cell counts of at least 12. The continuity-corrected Wald confidence limits are more appropriate for moderate sample sizes, say cell counts of at least 8.

But the data in Table 1 do not meet these criteria, much as they do not meet the criteria for asymptopic tests that assess association. So, are there exact methods for determining the confidence interval for the difference of proportions? Yes. However, methods for computing the exact confidence interval have limitations: for one, they are unconditional exact tests, not conditional exact tests, like Fisher's exact test, due to the presence of a nuisance parameter. Unconditional methods proceed by maximizing the p-value over the possible values of the nuisance parameter in order to eliminate it. In addition, the coverage of the exact confidence interval is not exactly $100 * (1-\alpha)$ but is at least $100 * (1-\alpha)$. See Agresti (1992) for more information.

You request the exact unconditional interval based on the raw risk difference with the RISKDIFF option in the EXACT statement in the FREQ procedure. The interval is produced by inverting two separate one-sided tests where the size of each test is at most $(\alpha/2)$, resulting in a confidence coefficient of at least $(1-\alpha)$. In addition, beginning with SAS/STAT 9.3, you can request the exact unconditional confidence interval by inverting two separate one-sided tests based on the score statistic proposed by Farrington and Manning; this has better coverage properties (Chang and Zhang 1999).

However, you can often do better than these exact methods with the Newcombe method (Newcombe 1998). The Newcombe method uses the quadratic equation method for the confidence interval for each proportion separately and integrates them to compute the confidence interval for the difference of proportions. You plug in the quadratic solutions P_{U1}, P_{L1} and P_{U2}, P_{L2} to obtain the lower (L) and upper (U) bounds for the confidence interval for the proportion difference.

$$L = (p_1 - p_2) - \sqrt{(p_1 - P_{L1})^2 + (P_{U2} - p_2)^2} \text{ and } U = (p_1 - p_2) + \sqrt{(P_{U1} - p_1)^2 + (p_2 - P_{L2})^2}$$

A continuity-corrected Newcombe's method also exists, and it should be considered if the total number of row counts is less than 10. The continuity correction is made in the computations of the confidence intervals for the binomial proportions. The Newcombe method can be somewhat conservative. The Appendix contains SAS/IML code that performs these computations with the continuity correction. In SAS/STAT 9.3, you can request these computations with new options for the RISKDIFF option in the FREQ procedure. In addition to the Newcombe method, you can also request the Farrington-Manning method (Farrington and Manning 1990). The Farrington-Manning method generally does well power-wise but the confidence interval it produces may be too narrow when applied to situations where you have less than a count of 5 per group.

The following SAS statements create the data set SMALL with the frequencies from Table 1.

```
data small;
   input treat $ outcome $ count @@ ;
   datalines;
Test Favorable 10 Test Unfavorable 2
Placebo Favorable 2 Placebo Unfavorable 4
;
```

The following PROC FREQ statements request the Wald, Newcombe, and Farrington-Manning confidence intervals for the difference of the favorable proportion for Test and Placebo. The unconditional exact interval based on the Farrington-Manning score statistic is requested with the RISKDIFF option in the EXACT statement. The CORRECT option specifies that the continuity correction be applied where possible, and the NORISK option suppresses the rest of the relative risk difference results.

```
proc freq order=data;
   weight count;
   tables treat*outcome / riskdiff(cl=(wald newcombe fm exact) correct norisks);
   exact riskdiff(fmscore);
run;
```

Figure 1 displays the confidence intervals.

Figure 1 Various Confidence Interval for Difference of Propotions

```
                          The FREQ Procedure

                 Statistics for Table of treat by outcome

           Confidence Limits for the Proportion (Risk) Difference
                       Column 1 (outcome = Favorabl)
                      Proportion Difference = 0.5000

            Type                         95% Confidence Limits

            Exact (FM Score)             -0.0000        0.8433
            Farrington-Manning            0.0380        0.9620
            Newcombe Score (Corrected)   -0.0352        0.8059
            Wald (Corrected)             -0.0571        1.0000
```

Note that the corrected Wald-based confidence interval is the widest interval, and it might have boundary issues with the upper limit of 1. The scored-based exact (unconditional) confidence interval has a relatively wide interval, (–0.0296, 0.8813) as does the interval based on Farrington-Manning scores, (0.0380, 0.9620), which also excludes the value zero. The score-based exact confidence interval has a narrower interval of (0.000, 0.8433) but comes with a zero boundary. Not shown here, the default exact unconditional interval is (−0.0296, 0.8813), which is fairly wide. The corrected Newcombe interval is the narrowest at (−0.0352, 0.8059), and it might be the most suitable for these data as it does not exclude zero and thus is in harmony with the Fisher exact test result.

Incidence Density and Poisson Regression

The 2 × 2 table can also represent incidence densities, in which you have counts of subjects who responded with an event versus extent of exposure for that event. These counts often follow the Poisson distribution. Some examples are:

- colony counts for bacteria or viruses
- accidents or equipment failure
- incidences for disease

Table 3 is based on the United States participants in an international study of the effect of a rotavirus vaccine in nearly 70,000 infants on health care events (hospitalizations and emergency room visits). Subjects received three doses of the vaccine or a placebo at 4 to 12 week intervals. They were followed for one year (Vesikari 2007). Table 3 displays the failures (hospitalizations or emergency room visits) and the person-years of exposure since not all subjects were in the study for an entire year.

Table 3 Vaccine Study Results for US

Treatment	Events	Person Years
Vaccine	3	7500
Placebo	58	7250

It is of interest to compute the ratio of incidence densities (IDR) of events for vaccine compared to placebo. Consider Table 4.

Table 4 Vaccine Study

Treatment	Events	Person Years
Vaccine	n_v	N_v
Placebo	n_p	N_p

You can assume that n_v and n_p have independent Possion distributions with expected values $\lambda_p N_p$ and $\lambda_v N_v$, respectively. You are interested in whether the incidence densities λ_v and λ_p are the same, which can be addressed by seeing whether the IDR is equal to 1; you determine this by evaluating its confidence limits. When the counts are small, such as the count of 3 in Table 3, then the exact confidence limits are appropriate.

Because n_v given $(n_v + n_p)$ has the conditional binomial distribution (under the assumption that n_v and n_p are independent Poisson);

$$\text{Bin}\left(n = n_v + n_p, P = \frac{\lambda_v N_v}{\lambda_p N_p + \lambda_v N_v}\right)$$

then

$$P = \frac{\frac{\lambda_v}{\lambda_p}\left(\frac{N_v}{N_p}\right)}{\frac{\lambda_v}{\lambda_p}\left(\frac{N_v}{N_p}\right) + 1} = \frac{RC}{RC + 1}$$

You then compute $p = n_v/(n_v + n_p)$ to produce a $100(1 - \alpha)$ confidence interval (P_L, P_U) for P. If you want the exact confidence interval for the incidence density ratios, you compute the exact confidence interval for the binomial proportion P. Use

$$\frac{N_p P}{(1-P)N_v}$$

as an estimator for R, and then

$$\left\{\frac{P_L}{(1-P_L)C}, \frac{P_U}{(1-P_U)C}\right\}$$

serves as a $100(1-\alpha)$% confidence interval for IDR, the incidence density ratio.

The following SAS statements compute the exact binomial confidence intervals for the binomial proportion P in the table and output them to a SAS data set.

```
      data vaccine2;
        input Outcome $ Count @@;
      datalines;
      fail 3 success 58
      ;

   ods select BinomialCLs;
   proc freq;
      weight count;
      tables Outcome / binomial (exact);
   ods output BinomialCLs=BinomialCLs;
   run;
```

The following SAS/IML® statements then produce the exact confidence interval for the IDR:

```
   proc iml ;
      Use BinomialCLs var{LowerCL UpperCL};
      read all into CL;
      print CL;
      q = { 3, 7500, 58 , 7250 };
      C= q[2,]/q[4];
      P= q[1]/ (q[1] + q[3]);
      R= P/ ((1-P) *C) ;
      CI= CL[1]/((1-CL[1])*C) || CL[2]/((1-CL[2])*C) ;
      print r CI;
   quit;
```

The IDR that compares vaccine failure to placebo is 0.05 with an exact confidence interval of (0.01002, 0.15355). Obviously the vaccine is much more effective than the placebo. The percent rate reduction in failures, a quantity often used in the assessment of vaccines, is computed as 100(1–IDR)=95%.

You might also be interested in determining the relationship of the rate or incidence to a set of explanatory variables. Events can be occurrences of rare diseases for populations with different sizes in epidemiological studies where the explanatory variables could be risk factors. In clinical trials, the events could be occurrences of rare disorders with possibly different levels of exposure risk; the explanatory variables are treatment and background covariates. In vaccine studies, the rare disorders are the diseases to be prevented.

Table 5 displays the vaccine data for both the United States and Latin America. There is interest in evaluating the effects of the vaccine with adjustment for the different regions.

Table 5 Medical Events for Rotavirus Vaccine Study

	Vaccine		Placebo	
Region	Events	Person Years	Events	Person Years
United States	3	7500	58	7250
Latin America	1	1250	10	1250

Since the events are considered to come from a Poisson distribution, you can perform Poisson regression to address the influence of the explanatory variables on the rates. Suppose that Y represents the counts and has expected value and variance both equal to μ. If the exposure time is denoted as N, then the rate can be expressed as

$$\log\left\{\frac{\mu}{N}\right\} = \alpha + x\beta$$

and this equation can be arranged as

$$\log(\mu) = \alpha + x\beta + \log(N)$$

The term $\log(N)$ is called an *offset* and must be accounted for in the estimation process.

If you have s independent groups referenced by $i = 1, 2, \ldots, s$, each with a vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{it})$ of t explanatory variables, you can write a likelihood function for the data as

$$\Phi(\mathbf{n}|\boldsymbol{\mu}) = \prod_{i=1}^{s} \mu_i^{n_i}\{\exp(-\mu_i)\}/n_i!$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_s)'$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_s)'$.

The loglinear Poisson model is often written as

$$\log\{n_i\} = \log\{N_i\} + \mathbf{x}_i'\boldsymbol{\beta}$$

in the generalized linear models framework, where the quantity $\log\{N_i\}$ is the offset. Thus, you can fit Poisson regression models with the GENMOD procedure, using the Poisson distribution, the log link function, and $\log\{N_i\}$ as the offset.

Exact Poisson regression is a useful strategy when you have small numbers of events because it does not depend on asymptotic results. Instead, it relies on estimation of the model parameters by using the conditional distributions of the sufficient statistics of the parameters. Since the event counts are small (1, 3) for the vaccine data in Table 5, then exact Poisson regression might also be indicated. This analysis is available through the GENMOD procedure, beginning with SAS/STAT 9.22.

The following SAS statements create the SAS data set ROTAVIRUS:

```
data rotavirus;
   input region $ treatment $ counts years_risk @@ ;
   log_risk=log(years_risk);
   datalines;
 US     Vaccine  3 7500 US     Placebo 58 7250
 LA     Vaccine  1 1250 LA     Placebo 10 1250
 ;
 run;
```

The following PROC GENMOD statements specify Poisson regression and exact Poisson regression. The ESTIMATE statement is used to produce the estimated IDR for the standard analysis, comparing Vaccine and Placebo, and the ESTIMATE=ODDS and CLTYPE=EXACT options are used to produce the same quantities in the exact analysis (actually CLTYPE=EXACT is the default).

```
proc genmod;
   class region  treatment/ param=ref;
   model counts = treatment region / dist=poisson offset= log_risk type3;
   estimate 'treatment' treatment 1 /exp;
   exact treatment / estimate=odds cltype=exact;
run;
```

Figure 2 contains the criteria for evaluating goodness of fit. With p-values of 0.2979 for the Deviance and 0.3431 for the Pearson chi-square (both 1 df), the model fit appears appropriate, altough the number of events is too small for formal use of these statistics.

Figure 2 Model Fit

```
                              The GENMOD Procedure

                     Criteria For Assessing Goodness Of Fit

         Criterion                     DF           Value        Value/DF

         Deviance                       1          0.2979          0.2979
         Scaled Deviance                1          0.2979          0.2979
         Pearson Chi-Square             1          0.3431          0.3431
         Scaled Pearson X2              1          0.3431          0.3431
         Log Likelihood                          189.6784
         Full Log Likelihood                      -7.6740
         AIC (smaller is better)                  21.3481
         AICC (smaller is better)                       .
         BIC (smaller is better)                  19.5069
```

Figure 3 contains the maximum likelihood parameter estimates.

Figure 3 Parameter Estimates

```
               Analysis Of Maximum Likelihood Parameter Estimates

                                     Standard       Wald 95%            Wald
    Parameter           DF  Estimate     Error   Confidence Limits  Chi-Square

    Intercept            1   -4.7886    0.3028   -5.3820   -4.1951      250.11
    treatment  Vaccine   1   -2.8620    0.5145   -3.8704   -1.8536       30.94
    region     US        1   -0.0467    0.3276   -0.6888    0.5953        0.02
    Scale                0    1.0000    0.0000    1.0000    1.0000

                         Analysis Of Maximum Likelihood
                              Parameter Estimates

                         Parameter           Pr > ChiSq

                         Intercept               <.0001
                         treatment  Vaccine      <.0001
                         region     US           0.8865
                         Scale

 NOTE: The scale parameter was held fixed.
```

Region is clearly not an influential effect, but it is considered a study design factor that should be maintained regardless of its p-value.

Figure 4 contains the estimate of the incidence density ratio for vaccine compared to placebo. The IDR takes the value of 0.057, which is $e^{-2.862}$ with a confidence interval of (0.0209, 0.1567) based on the Wald chi-square statistic.

Figure 4 Estimate Statement Results

```
                           Contrast Estimate Results

                    Mean            Mean            L'Beta   Standard
  Label         Estimate     Confidence Limits    Estimate      Error    Alpha

  treatment       0.0572     0.0209      0.1567    -2.8620     0.5145     0.05
  Exp(treatment)                                    0.0572     0.0294     0.05

                           Contrast Estimate Results

                             L'Beta             Chi-
          Label          Confidence Limits    Square    Pr > ChiSq

          treatment      -3.8704     -1.8536   30.94        <.0001
          Exp(treatment)  0.0209      0.1567
```

Figure 5 contains the exact analysis results.

Figure 5 Exact Odds Ratio and Confidence Intervals

```
                              Exact Odds Ratios

                                            95% Confidence      Two-sided
    Parameter              Estimate             Limits            p-Value    Type

    treatment Vaccine         0.057        0.015        0.153      <.0001    Exact
```

Here, the estimate of IDR is 0.057, with an exact confidence interval of (0.015, 0.153), which is fairly similar to the asymptotic results. The percent rate reduction of events due to the vaccine is 94%.

Randomization-Based Nonparametric Analysis of Covariance

Frequently, statisticians use analysis of covariance (ANCOVA) techniques to take the effects of covariates into account when estimating an effect of interest, for example, treatment effect in the analysis of clinical trials. When these covariates are associated with the outcome, adjusting for the covariates often results in a more precise estimate of the treatment effect. In randomized trials, covariate adjustment can help to offset the random imbalances among groups that still occur.

Often, the parametric procedures for performing analysis of covariance, such as linear regression for continuous outcomes, logistic regression for dichotomous outcomes, or proportional hazards regression for time-to-event outcomes are satisfactory, but often they are not. They require large samples for the maximum likelihood methods to apply, and when there are numerous categorical or time-to-event variables, this requirement is not met. Including interaction terms in the model makes this even worse.

Nonparametric methods, such as rank analysis of covariance is one solution. See p. 174 in Stokes et al. (2000) for an example of using ranks in the framework of extended Mantel-Haenszel statistics to conduct nonparametric comparisons of treatment groups, adjusted for the effects of covariates (Koch et al. 1982, 1990).

Another solution is to use randomization-based methods. These method uses weighted least squares analysis of the differences in groups between outcome variables and covariates simultaneously with the assumption that the differences in the covariates among groups are zero. With typical randomized studies, this expected difference would indeed be zero. You would use these methods for the types of situations in which, except for the covariates, you would use the CMH (Cochran-Mantel-Haenszel) test for dichotomous outcomes or the Wilcoxon Rank Sum test for ordinal outcomes.

Suppose that $\mathbf{f}_h = (d_h, \mathbf{u}_h{}')$ where $d_h = (\bar{y}_{h1} - \bar{y}_{h2})$ and $u_h = (\bar{\mathbf{x}}_{h1} - \bar{\mathbf{x}}_{h2})$. $\bar{y}_{hi}$ represents the mean responses and $\mathbf{x}_{\bar{h}i}$ represents the mean covariates. Suppose $\mathbf{S}_h$ is the estimated variance matrix for f_h relative to the randomization hypothesis under the null hypothesis for the statistical test or relative to an assumed sampling process for confidence interval estimation.

Suppose that w_h are weights, usually $w_h = n_{h1}n_{h2}/(n_{h1} + n_{h2})$, which are known as Mantel-Haenszel weights.

Form

$$\bar{\mathbf{f}} = \left(\frac{\sum_{h=1} w_h \mathbf{f}_h}{\sum_h w_h}\right) \text{ and } \mathbf{S} = \frac{(\sum_{h=1} w_h^2 \mathbf{S}_h)}{(\sum_h w_h)^2}$$

See Lavange, Durham, and Koch (2005) for the specific forms of **S**.

Then you fit weighted least squares regression relative to **S** (where **S** is $\mathbf{S}_0$ for testing the null hypothesis of no group effect and **S** is $\mathbf{S}_A$ for confidence interval estimation) on **f** in the following manner:

$$b = (\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{S}^{-1}\mathbf{f}$$

The estimated variance can be expressed as:

$$v_b = (\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Z})^{-1}$$

$\mathbf{Z} = [1, \underline{0}']'$ and b represents group difference and $b \pm 1.96\sqrt{v_b}$ is a 0.95 confidence interval for the group comparison. $Q{=}b^2/v_b$ is a chi-square (df=1) test statistic for the hypothesis of equal group effect.

Effectively, this approach averages group differences for the means of the outcome variables as well as the differences of the means of the covariates across strata using the appropriate weights to control for the effects of the strata. In a clinical trial setting, the group would represent treatment and the strata would represent clinics or centers. Covariate adjusted group differences are then computed with the weighted least squares regression step.

If the stratum sample sizes are sufficiently large, then these steps can be reversed. The covariate adjustments are performed on a within-stratum basis, and then the adjusted group difference estimates are averaged across the strata using the appropriate weights. The model above is applied to the f_h and S_h for each strata to obtain b_h and v_{bh}, and then the weights are applied as above to form the linear combinations $\bar{b}$ and $v_{\bar{b}}$. If you want to evaluate group by subgroup interactions, you can apply these operations to each subgroup and then compare the bs across the subgroups.

These methods are applicable when the outcomes are continuous, ordinal, or time-to-event variables. They are also applicable to the evaluation of differences of binomial proportions that result when the outcome is dichotomous. For more information, see LaVange, Durham, and Koch (2005), and Koch, Tangen et al. (1998).

Respiratory Data Example

A multi-center, multi-visit clinical trial compared two treatments for patients with a respiratory disorder (Stokes et al. 2000). Responses were recorded on a 1–5 scale, for terrible, poor, fair, good, or excellent response. In addition, researchers recorded age, gender, and baseline. One hundred eleven patients were seen at two centers at each of four visits.

Table 6 Partial Listing of Center 1 Respiratory Patients

Patient	Drug	Sex	Age	Baseline	Visit 1	Visit 2	Visit 3	Visit 4
53	A	F	32	1	2	3	4	2
18	A	F	47	2	2	3	4	4
43	A	M	11	4	4	4	4	2
12	A	M	14	2	3	3	3	2
51	A	M	15	0	2	3	3	3
20	A	M	20	3	3	2	3	1
16	A	M	22	2	1	3	3	3
.	.	.	.	.	.	.	.	.

Table 7 displays p-values for the association of covariables with both treatment and response. These p-values should be viewed as descriptive guides only.

Table 7 Descriptive p-Values for Association

	Age	Gender	Baseline
Visit 1	0.53	0.34	<0.001
Visit 2	0.22	0.17	<0.001
Visit 3	0.13	0.47	<0.001
Visit 4	0.03	0.54	<0.001
Treatment	0.73	0.013	0.98

Baseline is strongly associated with the response (Visits) and exhibits no imbalance for treatment. Gender has essentially no association with the response, but it does demonstrate a random imbalance. Age is associated with the response for the last visit, and there is no imbalance for treatment with respect to age.

You can apply the randomization-based nonparametric ANCOVA methods to the analysis of these data with the NparCov2 SAS macro (Zink and Koch 2001). The following analysis focuses on the Visit 1 results and is directed at the dichotomous response variable that compares treatments on good or excellent response. The d_h are the differences in the proportion of (good or excellent response) for each stratum. Four analyses are performed:

- unadjusted
- adjusted for baseline
- adjusted for baseline, stratified by center under H_0
- adjusted for baseline, stratified by center under H_A

The NparCov2 macro requires that the SAS data set be structured so that all of a subject's data are on one line (you can analyse multiple responses) and all variables need to be numeric. The following SAS statements create data set RESP and compute a few dummy variables from character variables.

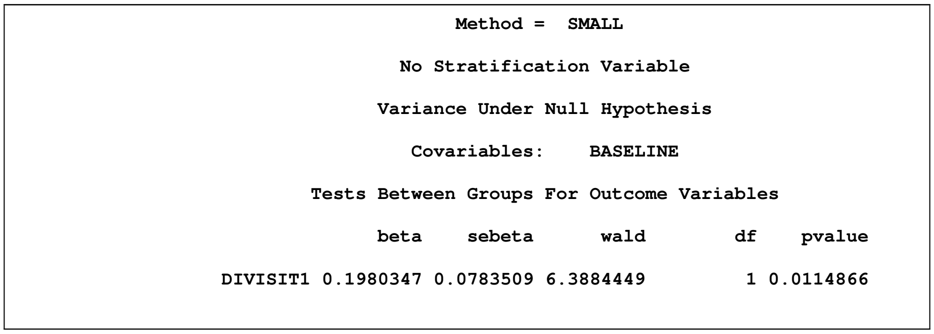
```
%include 'NParCov2.macro.sas';
   data resp;
      input center id treatment $ sex $ age baseline
      visit1-visit4 @@;
   datalines;
   1  53 A F  32 1  2 2 4 2  2  30 A F  37 1  3 4 4 4
   1  18 A F  47 2  2 3 4 4  2  52 A F  39 2  3 4 4 4
   1  54 A M  11 4  4 4 4 2  2  23 A F  60 4  4 3 3 4
   1  12 A M  14 2  3 3 3 2  2  54 A F  63 4  4 4 4 4
   1  51 A M  15 0  2 3 3 3  2  12 A M  13 4  4 4 4 4
   1  20 A M  20 3  3 2 3 1  2  10 A M  14 1  4 4 4 4
   1  16 A M  22 1  2 2 2 3  2  27 A M  19 3  3 2 3 3
   1  50 A M  22 2  1 3 4 4  2  16 A M  20 2  4 4 4 3
   .....
   data resp2; set resp;
      id=center*1000 + id;
      Itreat=(treatment='A');
      Icenter=(center=1);
      Isex=(sex='M');
      divisit1=(visit1=3 or visit1=4);
      di_base = (baseline=3 or baseline=4);
   run;
```

The following NPARCOV2 statements specify an analysis with an adjustment just for BASELINE under the null hypothesis. C=1 specifies Mantel-Haenszel weights, and METHOD=SMALL specifies that the weighted averages of treatment group differences are taken across the strata and then the covariance adjustment is performed. This is appropriate when the strata sizes are relatively small. See Zink and Koch (2001) for more information about using the NparCov macro.

```
%let deps = divisit1;
%let covs=baseline;
%let stra=Icenter;
%let treat=Itreat;
%NParCov(outcomes=&deps,covars=&covs, treatmnt=&treat,c=1,hypoth=0, method=small, data=resp2);
```

Figure 6 displays the results.

Figure 6 Results of Nonparametric ANCOVA for Dichotomous Outcome



```
                            Method =  SMALL

                       No Stratification Variable

                     Variance Under Null Hypothesis

                        Covariables:    BASELINE

                Tests Between Groups For Outcome Variables

                           beta     sebeta      wald        df    pvalue

         DIVISIT1 0.1980347 0.0783509 6.3884449         1 0.0114866
```

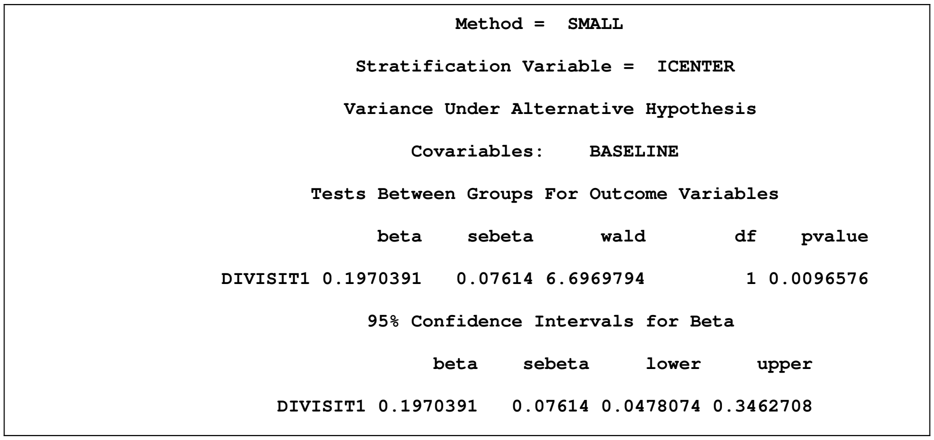
Q takes the value 6.388 with $p = 0.0115$.

The next call to NparCov requests an analysis that is stratified by center and adjusted by baseline under the alternative hypothesis. This produces estimates of the treatment difference and its confidence interval.

```
%NParCov(outcomes=&deps,covars=&covs, strat=&stra, treatmnt=&treat,c=1,hypoth=1, method=small, data=resp2);
```

Figure 7 displays the results. Q_A is 6.697 with p=0.0097. Since the alternative hypotheses was used, the treatment difference and confidence intervals are also computed.

Figure 7 Results of Nonparametric ANCOVA for Dichotomous Outcome



```
                    Method =  SMALL

            Stratification Variable =  ICENTER

            Variance Under Alternative Hypothesis

                 Covariables:    BASELINE

         Tests Between Groups For Outcome Variables

                 beta    sebeta      wald        df    pvalue

DIVISIT1 0.1970391   0.07614 6.6969794         1 0.0096576

               95% Confidence Intervals for Beta

                     beta    sebeta     lower     upper

     DIVISIT1 0.1970391   0.07614 0.0478074 0.3462708
```

Table 8 summarizes the results for these analyses. The unadjusted results and the results adjusted for baseline and center under the null hypothesis are also displayed (code not shown).

Table 8 Descriptive p-Values for Association

Model	Chi-Square (1df)	p-Value	Treatment Difference	95% CI
Unadjusted	4.260	0.0390	0.194	(0.015,0.373)
Baseline adjusted (H_0)	6.388	0.0115		NA
Baseline adjusted and center stratified (H_0)	6.337	0.0118		NA
Baseline adjusted and center stratified (H_A)	6.697	0.0097	.197	(0.048,0.346)

Adjusting for baseline produces stronger results than those produced by the unadjusted analysis. However, also stratifying on center does not change the picture. The treatment difference is roughly 0.20 for patients receiving treatment compared to those receiving the placebo with a confidence interval of (0.048, 0.346). The treatment difference is similar to that seen in the unadjusted analysis (0.194), but the confidence interval for the former is narrower than that of the latter (0.015, 0.373). See Lavange and Koch (2008) for a more detailed analysis of these data with these methods.

Note that the H_0 unadjusted nonparametric ANCOVA results without stratification are the same as would be obtained if you fit a linear regression model to the response variable with baseline as the predictor, took the residuals from this model, and then analyzed them for their association with treatment using the extended MH chi-square test to compare mean scores.

These methods can also be used for ordinal and continuous outcomes, so, for example, you could use the ordinal scores for the responses in this data set. In addition, these methods have been extended to compare the logits in a logistic regression and have also been extended to time-to-event data (Tangen and Koch 1999 and Saville et al. 2010, 2011).

Conditional Logistic Regression

Conditional logistic regression has been used as a technique for analyzing retrospective studies in epidemiological work for many years. In these studies, you match cases with an event of interest with controls who didn't have the event. You determine whether the case and control have the risk factors under investigation, and, by using a conditional likelihood, predict the event given the explanatory variables. You set up the probabilities for having the exposure given the event and then use Bayes' theorem to determine the relevant condtional likelihood.

More recently, conditional logistic regression is used frequently in the analysis of highly stratified data, and it is also used for the analysis of crossover studies. With highly stratified data, common in many clinical trials settings, you may have a small number of events per stratum relative to the number of parameters you want to estimate. Thus, the sample size requirements for the standard maximum likelihood approach to unconditional logistic regresson are not applicable. Crossover designs, in which subjects act as their own controls, provide a similar conumdrum.

Stokes et al. (2000) include an example of a clinical trial in which researchers studied the effect of a treatment for a skin condition. A pair of patients participated in each of 79 clinics; one patient received the treatment, and the other patient received the placebo. Age, sex, and initial grade for skin condition were also recorded. The response was whether the patient's skin improved.

Since there are only two observations per center, you cannot estimate all of these parameters without bias. However, you can fit a model based on conditional probabilities that condition away the center effects, which results in a model that contains substantially fewer parameters.

Suppose y_{ij} represents the patient from the *i*th clinic and the *j*th treatment (*i*=1 through 79 and *j*=1, 2). Suppose y_{ij} =1 if improvement occurs and y_{ij}=0 otherwise. Suppose $x_{ij} = 1$ for treatment and $x_{ij} = 0$ for placebo. Suppose **z**=$z_{ij1}, z_{ij2}, ..., z_{ijt})'$ represents the t explanatory variables.

The usual likelihood for $\{y_{ij}\}$ is written

$$\text{Pr}\{y_{ij}\} = \pi_{ij} = \frac{\exp\{\alpha_i + \beta x_{ij} + \boldsymbol{\gamma}'\mathbf{z}_{ij}\}}{1 + \exp\{\alpha_i + \beta x_{ij} + \boldsymbol{\gamma}'\mathbf{z}_{ij}\}}$$

where α_i is the effect of the ith clinic, β is the treatment parameter, and $\boldsymbol{\gamma}' = (\gamma_1, \gamma_2, \ldots, \gamma_t)$ is the parameter vector for the covariates **z**. The α_i are known as *nuisance parameters*.

You can write a conditional probability for $\{y_{ij}\}$ as the ratio of the joint probability of a pair's treatment patient improving and the pair's placebo patient not improving to the joint probability that either the treatment patient or the placebo patient improved.

$$\text{Pr}\left\{y_{i1}{=}1, y_{i2}{=}0 \middle| y_{i1}{=}1, y_{i2}{=}0 \text{ or } y_{i1}{=}0, y_{i2}{=}1\right\} = \frac{\text{Pr}\{y_{i1}{=}1\}\,\text{Pr}\{y_{i2}{=}0\}}{\text{Pr}\{y_{i1}{=}1\}\,\text{Pr}\{y_{i2}{=}0\} + \text{Pr}\{y_{i1}{=}0\}\,\text{Pr}\{y_{i2}{=}1\}}$$

You write the probabilities in terms of the logistic model,

$$\text{Pr}\{y_{i1}{=}1\}\,\text{Pr}\{y_{i2}{=}0\} = \frac{\exp\{\alpha_i + \beta + \boldsymbol{\gamma}'\mathbf{z}_{i1}\}}{1 + \exp\{\alpha_i + \beta + \boldsymbol{\gamma}'\mathbf{z}_{i1}\}} \cdot \frac{1}{1 + \exp\{\alpha_i + \boldsymbol{\gamma}'\mathbf{z}_{i2}\}}$$

and

$$\text{Pr}\{y_{i1}{=}1\}\,\text{Pr}\{y_{i2}{=}0\} + \text{Pr}\{y_{i1}{=}0\}\,\text{Pr}\{y_{i2}{=}1\} =$$

$$\frac{\exp\{\alpha_i + \beta + \boldsymbol{\gamma}'\mathbf{z}_{i1}\}}{1 + \exp\{\alpha_i + \beta + \boldsymbol{\gamma}'\mathbf{z}_{i1}\}} \cdot \frac{1}{1 + \exp\{\alpha_i + \boldsymbol{\gamma}'\mathbf{z}_{i2}\}}$$

$$+ \frac{1}{1 + \exp\{\alpha_i + \beta + \boldsymbol{\gamma}'\mathbf{z}_{i1}\}} \cdot \frac{\exp\{\alpha_i + \boldsymbol{\gamma}'\mathbf{z}_{i2}\}}{1 + \exp\{\alpha_i + \boldsymbol{\gamma}'\mathbf{z}_{i2}\}}$$

If you form their ratio and cancel like terms, this expression reduces to:

$$\frac{\exp\{\beta + \boldsymbol{\gamma}'(\mathbf{z}_{i1} - \mathbf{z}_{i2})\}}{1 + \exp\{\beta + \boldsymbol{\gamma}'(\mathbf{z}_{i1} - \mathbf{z}_{i2})\}}$$

Focusing on modeling a meaningful conditional probability allows you to develop a model that has a reduced number of parameters that can be estimated without bias.

You can actually fit this model with a standard logistic regression model by using the pairs that have either a (1,0) or (0,1) response pattern and using the $Z_{i1} - Z_{i2}$ as the explanatory variables. There are 34 centers with the pattern (1,0) and 20 centers with the pattern (0,1).

The LOGISTIC procedure provides conditional logistic regression with the inclusion of a STRATA statement to denote the conditioning variable.

The following PROC LOGISTIC statements produce this analysis for the skin condition data (SAS data set TRIAL2 assumed). The variable CENTER is placed in the STRATA statement. Explanatory variables in the MODEL statement include the main effects and pairwise interactions. By specifying the SELECTION=FORWARD option in conjunction with the INCLUDE=4 option, you ensure that initial skin condition, age, sex, and treatment main effects are included in the model and a residual score statistics is produced.

```
ods graphics on;
proc logistic plots(only)=oddsratio (logbase=e);
   class center treat sex / param=ref ref= first;
   strata center;
   model improve = initial age sex treat sex*age sex*initial age*initial
     sex*treat treat*initial treat*age / selection=forward include=4 details clodds=wald;
   units age=10 /default=1;
run;
ods graphics off;
```

Note that the UNITS statement specifies units per changes for the odds ratios for continuous variables; you need to specify the DEFAULT= option if you still want odds ratios produced for the unspecified continuous variables in the model.

Figure 8 displays the strata information. Note the only the observations with Response Profile 2 play a role in the analysis, as discussed previously.

Figure 8 Strata Summary

```
                       The LOGISTIC Procedure

                        Conditional Analysis

                           Strata Summary

                       improve
          Response     -------     Number of
           Pattern     1     2        Strata    Frequency

                 1     0     2             7           14
                 2     1     1            54          108
                 3     2     0            18           36
```

Using SELECTION=FORWARD and the INCLUDE= option produces the residual chi-square to assess the fit of the main effects model since it assesses the joint contribution of the interaction terms in the MODEL statement.

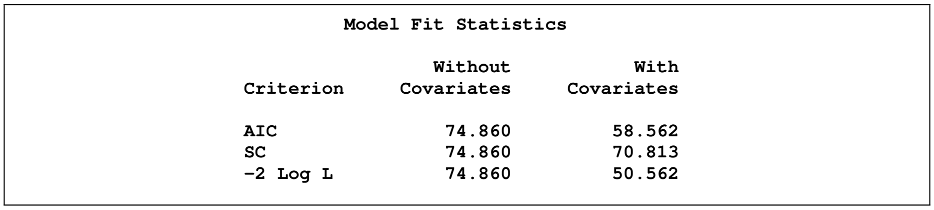
Figure 9 Residual Chi-Square

```
                      Residual Chi-Square Test

               Chi-Square       DF     Pr > ChiSq

                   4.7214        6         0.5800
```

Since 20 observations have the less prevalent response, this model can only support about 20/5=4 terms, so the appropriateness of the residual chi-square is questionable. However, since its p-value is quite high—greater than 0.5, and the individual tests of the interaction terms generally have reasonably high p-values as well (not shown here), there is good confidence that the main effects model fits reasonably well. (Note that if you specify only TREAT as the explanatory variable, the residual score test for TREAT is comparable to McNemar's test. That model has an estimated odds ratio that is the same as the matched pairs odds ratio.)

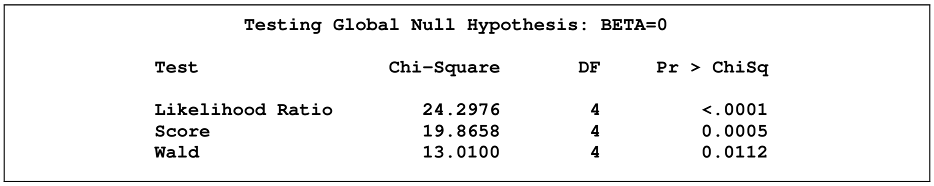
Model adequacy is also supported by the model fit statistics and the global tests, as shown in Table 10 and Table 11.

Figure 10 Fit Statistics



```
                     Model Fit Statistics

                              Without           With
             Criterion     Covariates     Covariates

             AIC               74.860         58.562
             SC                74.860         70.813
             -2 Log L          74.860         50.562
```

Figure 11 Global Tests

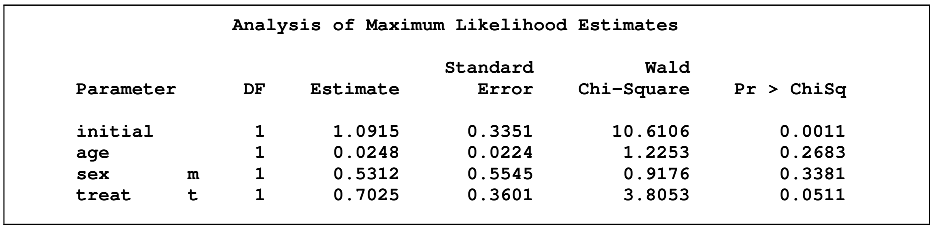


```
             Testing Global Null Hypothesis: BETA=0

     Test                 Chi-Square       DF     Pr > ChiSq

     Likelihood Ratio        24.2976        4         <.0001
     Score                   19.8658        4         0.0005
     Wald                    13.0100        4         0.0112
```

However, note that the disagreement among the statistics in Figure 11 suggests the need to consider an exact analysis as well.

Figure 12 displays the parameter estimates for the main effects model. Initial grade has a strong influence, and treatment is marginally influential (p=0.0511).

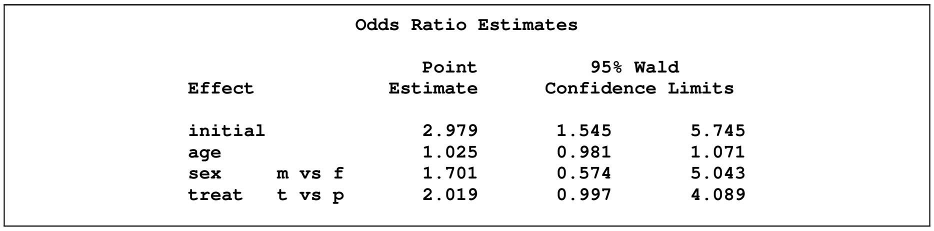
Figure 12 Parameter Estimates



```
                Analysis of Maximum Likelihood Estimates

                                         Standard          Wald
  Parameter      DF     Estimate        Error    Chi-Square    Pr > ChiSq

  initial         1       1.0915       0.3351       10.6106        0.0011
  age             1       0.0248       0.0224        1.2253        0.2683
  sex       m     1       0.5312       0.5545        0.9176        0.3381
  treat     t     1       0.7025       0.3601        3.8053        0.0511
```

Figure 13 displays the odds ratios for this model.

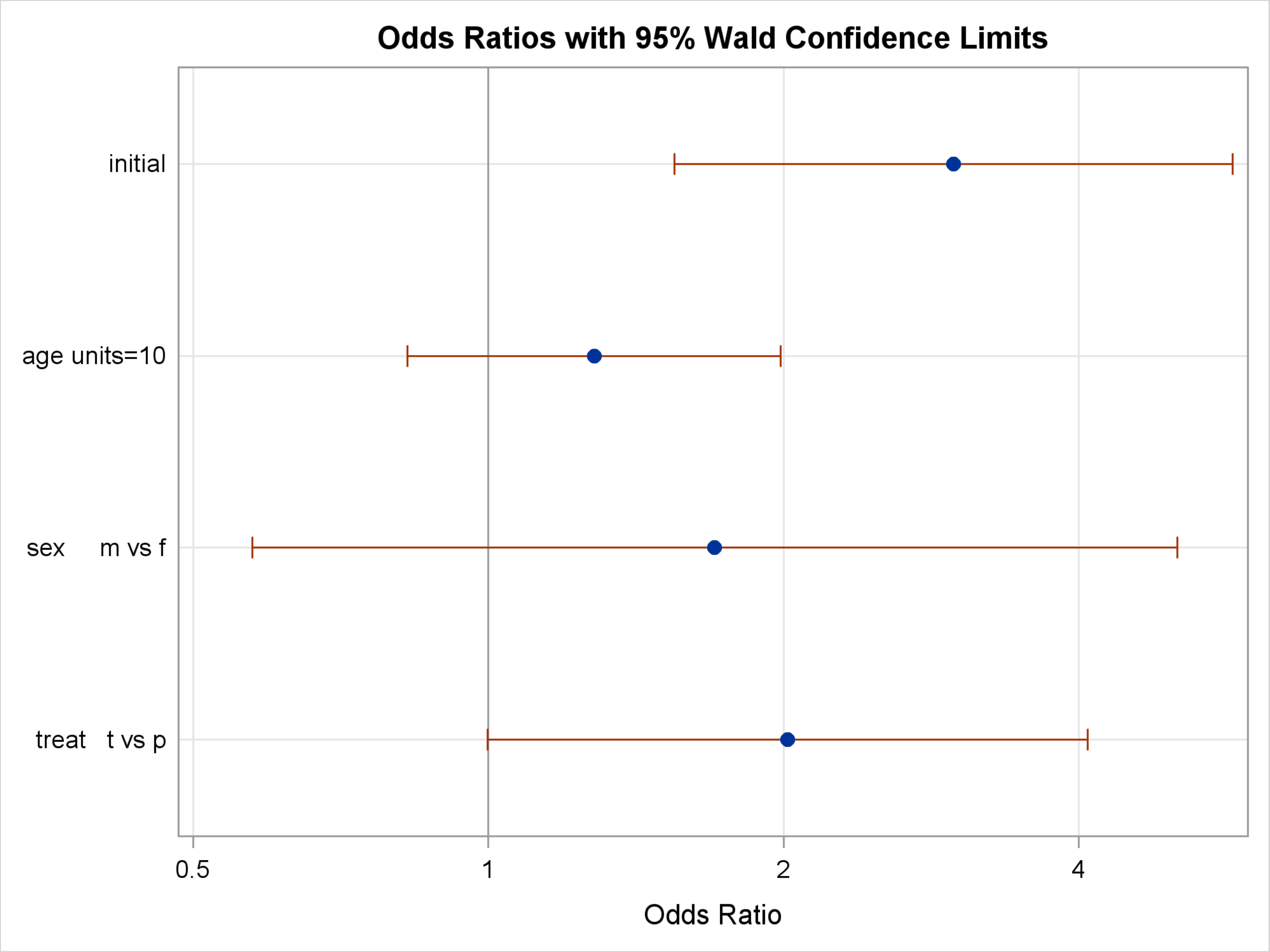
Figure 13 Odds Ratios



```
                      Odds Ratio Estimates

                             Point          95% Wald
      Effect              Estimate      Confidence Limits

      initial                2.979       1.545       5.745
      age                    1.025       0.981       1.071
      sex      m vs f        1.701       0.574       5.043
      treat    t vs p        2.019       0.997       4.089
```

The odds of improvement for those on treatment is $e^{0.7025}$=2.019 times higher than for those on placebo, adjusted for initial grade, age, and sex. The odds ratios for odds of improvement and related confidence intervals for all of the explanatory variables are displayed graphically in Figure 14.

Figure 14 Odds Ratios from the Main Effects Model



You can also obtain the exact odds ratio estimates by fitting the same model with the EXACT statement. You need to remove the SELECTION=FORWARD option. Specify TREAT in the EXACT statement and specify the ESTIMATE=ODDS option:

```
proc logistic;
  class center treat sex / param=ref ref= first;
  strata center;
  model improve = initial age sex treat / clodds=wald;
  exact treat /  estimate=odds;
run;
```

Figure 15 displays these results.

Figure 15 Exact Odds Ratio for Treatment Adjusted for Other Variables

```
                          The LOGISTIC Procedure

                        Exact Conditional Analysis

                            Exact Odds Ratios

                                 95% Confidence     Two-sided
      Parameter      Estimate        Limits          p-Value    Type

      treat     t       1.943    0.950      4.281     0.0715    Exact
```

The exact conditional analysis odds ratio estimate is 1.943 for treatment compared to placebo, and its 95% confidence interval is (0.950, 4.281). The exact p-value is 0.0715.

Consider repeating the exact analysis with treatment as the only explanatory variable. You would obtain the same exact odds ratio estimate as you do when you perform a stratified analysis by center of treatment and improvement with the FREQ procedure and request the exact odds ratio (COMOR option), which is only available for the analysis of stratified 2×2 tables. See Table 16 for the results.

Figure 16 Exact Odds Ratio for Treatment

```
                          The LOGISTIC Procedure

                        Exact Conditional Analysis

                            Exact Odds Ratios

                                   95% Confidence     Two-sided
     Parameter      Estimate           Limits           p-Value    Type

     treat     t       1.700      0.951       3.117       0.0759    Exact
```

Note that the$p$-value of 0.0759 is also identical to what you would get for the exact McNemar's test for these data for the table of treatment by outcome with counts representing the pairs of subjects from the centers.

```
proc freq;
   tables center*treat*improve / noprint;
   exact comor;
run;
```

Figure 17 Exact Odds Ratio for Treatment with PROC FREQ

```
                            The FREQ Procedure

                  Summary Statistics for treat by fimprove
                          Controlling for center

                            Common Odds Ratio
                 ------------------------------------
                 Mantel-Haenszel Estimate      1.7000

                 Asymptotic Conf Limits
                 95% Lower Conf Limit          0.9785
                 95% Upper Conf Limit          2.9534

                 Exact Conf Limits
                 95% Lower Conf Limit          0.9509
                 95% Upper Conf Limit          3.1168
```

The PROC FREQ results are the same as those for PROC LOGISTIC. The value of obtaining this estimate with the LOGISTIC procedure is that you are not limited to the case of a single, dichotomous explanatory variable.

Bayesian Methods

Bayesian methods have become widely used in the past 15 years. These methods treat parameters as random variables and define probabilities in terms of 'degrees of belief'; you can make probability statements about them. Bayesian methods allow you to incorporate prior information into your data analysis, such as historical knowledge or expert opinion, and they provide a framework for addressing specific scientific questions that can not be addressed with single point estimates.

Given data $\mathbf{x} = \{x_1, ..., x_n\}$, Bayesian inference is carried out in the following way:

1. You choose a *prior distribution* $\pi(\theta)$ for θ. The distribution describes your beliefs about the parameter *before* you examine the data.
2. Given the observed data $\mathbf{x}$, you select a model (density) $f(\mathbf{x}|\theta)$ to describe the distribution of $\mathbf{x}$ given θ.
3. You update your beliefs about θ by combining information from the prior distribution and the data through the calculation of the *posterior distribution* $\pi(\theta|\mathbf{x})$.

The third step is carried out by using Bayes' theorem, which enables you to combine the prior distribution and the model.

Bayesian methods provide a convenient setting for a wide range of statistical models, including hierarchical models and missing data problems. These methods provide inferences that are conditional on the data and are exact, without relying on asymptotic approximation. Small sample inference proceeds in the same way as if you had a large sample. Bayesian estimation does often come with high computational costs, since they usually depend on simulation algorithms to draw samples from target posterior distributions and use these samples to form estimates, such as Monte Carlo methods. However, modern computing makes these computational requirements tractable. For more information, see “Introduction to Bayesian Analysis Procedures” in the SAS/STAT User’s Guide, which includes a reading list of Bayesian textbooks and review papers at varying levels of expertise.

Bayesian methods have been applied to a variety of categorical data analyses, mainly for statistical modeling applications. You can produce Bayesian counterparts of frequentist categorical data modeling techniques such as logistic regression and Poisson regression. The following example describes how to perform Bayesian logistic regression using the GENMOD procedure in SAS/STAT software. It assumes a knowledge of basic Bayesian methodology.

Researchers were interested in determining the influence of an environmental toxin on the number of deaths of beetles in a certain geographic location. The following SAS statements create the SAS data set BEETLES:

```
data beetles;
   input n y x @@;
   datalines;
6  0  25.7   8  2  35.9   5  2  32.9   7  7  50.4   6  0  28.3
7  2  32.3   5  1  33.2   8  3  40.9   6  0  36.5   6  1  36.5
6  6  49.6   6  3  39.8   6  4  43.6   6  1  34.1   7  1  37.4
8  2  35.2   6  6  51.3   5  3  42.5   7  0  31.3   3  2  40.6
;
```

The variable Y represents the number of deaths for total beetles N exposed to a specific level X of the environmental toxin. Thus, y_i can be considered to come from a binomial distribution (n_i, p_i) where p_i is the probability of death. Logistic regression is an appropriate modeling technique for assessing the influence of the toxin on the number of beetle deaths.

In order to perform a Bayesian analysis, you need to decide on the appropriate priors for the model parameters, an intercept α, and a coefficient β for the toxin. Since no previous study information or expert opinion is available, a diffuse normal prior is assumed for both parameters, or Normal$(0,\ 10^6)$. The following PROC GENMOD statements request a Bayesian logistic regression analysis.

```
proc genmod data=beetles;
   model y/n= x / dist=bin link=logit;
   bayes seed=1 coeffprior=normal outpost=postbeetle;
run;
```

You specify a logistic regression model with the MODEL statement, using events/trials syntax for the response in this case. Specifying DIST=BIN and LINK=LOGIT requests logistic regression. The BAYES statement requests a Bayesian analysis. The COEFFPRIOR=NORMAL requests the normal diffuse prior; otherwise, the uniform prior would be the default. Jeffreys’ prior is also available. The OUTPOST=POSTBEETLE option requests that the posterior estimates be placed in the data set POSTBEETLE. Using a specific SEED value means that you can reproduce the exact results in a subsequent execution of these statements. The adaptive rejection algorithm is used to draw posterior samples in a Gibbs scheme.

Figure 18 displays the diagnostic plots for this model. The trace, autocorrelation, and density plots all show that the Markov chain has achieved satisfactory convergence for the X parameter. The same is true for the intercept (not shown here).

Figure 18 Diagnostics for X

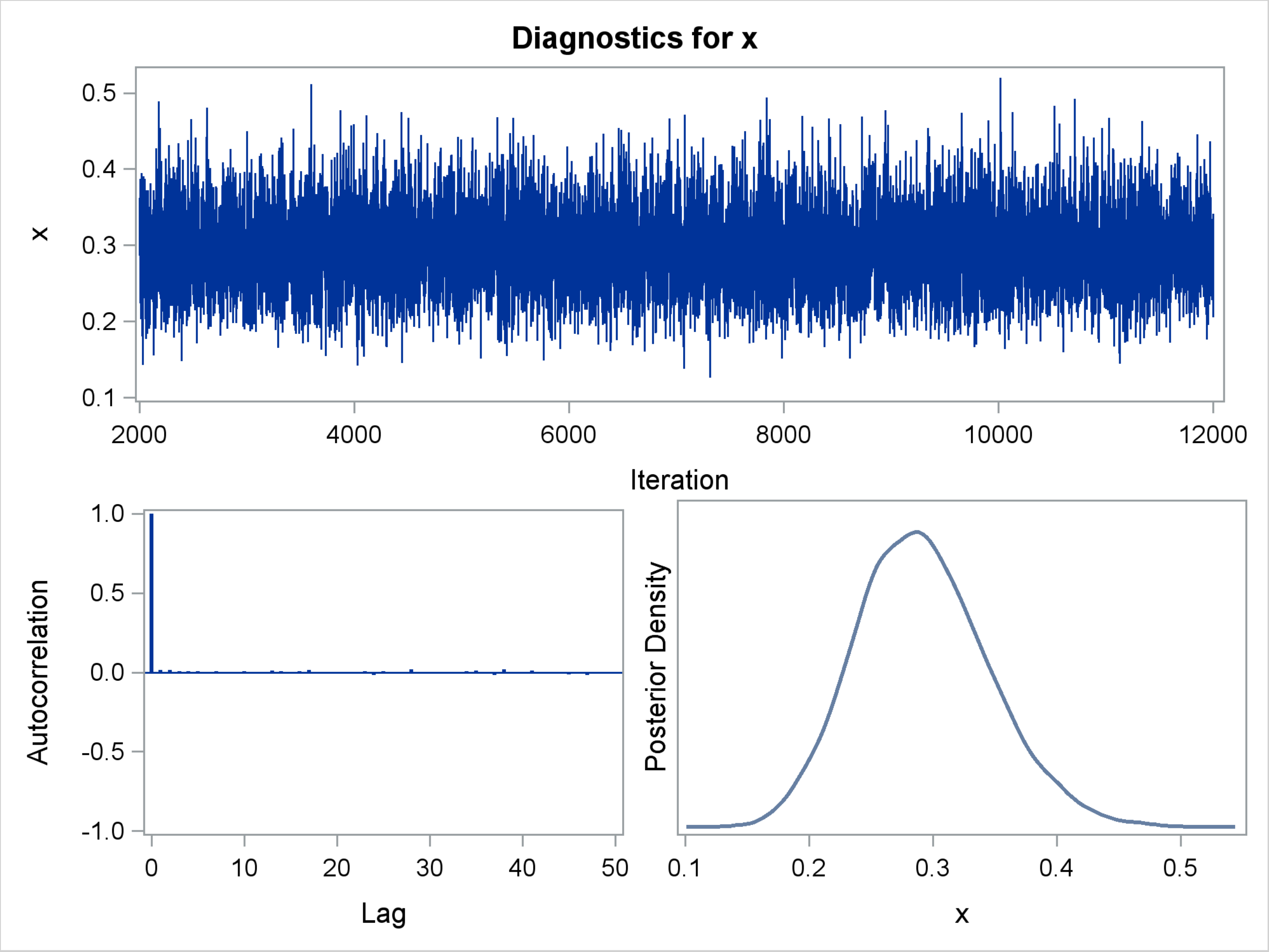
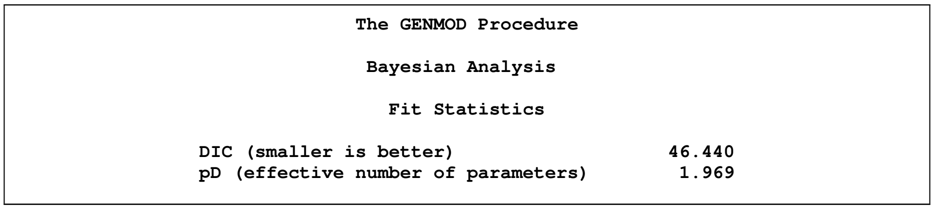


Figure 19 displays the DIC fit statistic. The DIC (Deviance Information Criterion)is defined similarly to the AIC and trades off goodness of fit and model complexity. Smaller is better, and you can use it to compare different (non-nested) models.

Figure 19 Bayesian Fit Statistics



```
                              The GENMOD Procedure

                               Bayesian Analysis

                                 Fit Statistics

                 DIC (smaller is better)                  46.440
                 pD  (effective number of parameters)      1.969
```

Figure 20 and Figure 21 display the autocorrelations and the effective sample sizes for this analysis. The posterior sample autocorrelations after 1 lag are negligible, as previously seen on the autocorrelation plot—another indication of good mixing.

Figure 20 Posterior Autocorrelations

```
                         The GENMOD Procedure

                          Bayesian Analysis

                      Posterior Autocorrelations

         Parameter       Lag 1       Lag 5      Lag 10      Lag 50

         Intercept      0.0290      0.0052      0.0039     -0.0084
         x              0.0154      0.0057      0.0048     -0.0050
```

Figure 21 Effective Sample Sizes

```
                        Effective Sample Sizes

                                       Autocorrelation
         Parameter          ESS                   Time    Efficiency

         Intercept       9208.6                 1.0859        0.9209
         x               9427.9                 1.0607        0.9428
```

The effective sample sizes for both parameters are over 9,000, reasonable since a simulation size of 10,000 samples was used.

Figure 22 contains the posterior summaries. The estimate of the mean of the marginal posterior distribution for the intercept is −11.7746, and the sample mean for the coefficient for X is 0.2922. The two 95% credible intervals for the intercept are both negative, as shown in Figure 23, and the same intervals are both positive for the X coefficient. Clearly, the environment toxin has a sizable impact on the probability of death for these beetles.

Figure 22 Posterior Summaries

```
                               The GENMOD Procedure

                                Bayesian Analysis

                               Posterior Summaries

                                            Standard                 Percentiles
Parameter              N        Mean       Deviation         25%         50%         75%

Intercept          10000    -11.7746          2.0588    -13.0990    -11.6434    -10.3366
x                  10000      0.2922          0.0532      0.2546      0.2891      0.3263
```

Figure 23 Posterior Intervals

```
                               Posterior Intervals

     Parameter     Alpha     Equal-Tail Interval         HPD Interval

     Intercept     0.050    -16.1021     -8.0552    -16.0120     -8.0001
     x             0.050      0.1957      0.4045      0.1929      0.4005
```

Note that since noninformative priors are used, these parameter estimates are similar to those you would obtain with a maximum likelihood analysis.

An advantage of Bayesian methods, regardless of whether informative or noninformative priors are used, is that its framework that allows you to address scientific questions directly. You can determine the posterior probability that the environmental toxin increases the probability of death, or Pr$\{\beta > 0|\mathbf{y}\}$. You do this by assessing whether the simulated values of β saved to SAS data set POSTBEETLE are greater than zero. The following PROC FORMAT and PROC FREQ statements are one way to perform this analysis.

```
proc format;
   value Xfmt low-0 = 'X <= 0' 0<-high = 'X > 0';
run;
proc freq data=postbeetle;
```

```
      tables X /nocum;
      format X Xfmt.;
   run;
```

All of the simulated values for β are indeed greater than zero, as seen in Figure 24, so the sample estimate of the posterior probability that $\beta > 0$ is 100%. The evidence that the environmental toxin impacts beetle devastation is overwhelming.

Figure 24 Effective Sample Sizes

```
                    The FREQ Procedure

                 x     Frequency      Percent
            ---------------------------------
            X > 0         10000        100.00
```

Suppose you are interested in computing some of the percentiles of the tolerance distribution, or the level of the environmental toxin X at which 5%, 50%, and 95% of the beetles succumbed. These quantities are known as LD5, LD50, and LD95, for lethal dose. If p represents the probability of interest, then you solve the following equation for x_p

$$\hat{\alpha} + \hat{\beta} x_p = \log\left\{\frac{p}{1-p}\right\}$$

for the various values of p to obtain the LDs. For example, LD50=$-\hat{\alpha}/\hat{\beta}$.

With Bayesian analysis, you can readily make inferences on quantities that are transformations of the random variables. You compute the LDs with DATA step operations on the saved posterior samples data set POSTBEETLE. In addition, by producing the 0.025 and 0.975 percentiles of these quantities, you can generate equal-tailed 95% credible intervals for them.

First, LDs are created for each of the posterior samples:

```
   data LDgenerate;
      set postbeetle;
      LD05 = (log(0.05 / 0.95) - intercept) / x;
      LD50 = (log(0.50 / 0.50) - intercept) / x;
      LD95 = (log(0.95 / 0.05) - intercept) / x;
   run;
```

The UNIVARIATE procedure computes the posterior mean for the LD variables as well as their 0.025 and 0.975 percentiles.

```
   proc univariate data=LDgenerate noprint;
      var ld05 ld50 ld95;
      output out=LDinfo
      mean=LD5 LD50 LD95
      pctlpts=2.5 97.5
      pctlpre= ld05_ ld50_ ld95_;
   run;
   proc print data=LDinfo;
   run;
```

Figure 25 LD Information

```
                                                 ld05_             ld50_             ld95_
  Obs    LD5      LD50     LD95   ld05_2_5    97_5   ld50_2_5    97_5   ld95_2_5    97_5

   1   29.9404  40.3664  50.7924   25.7704  32.9615   38.6752  42.3798   46.7259  56.5187
```

Table 9 displays the estimated LDs and their 95% equal-tailed credible intervals.

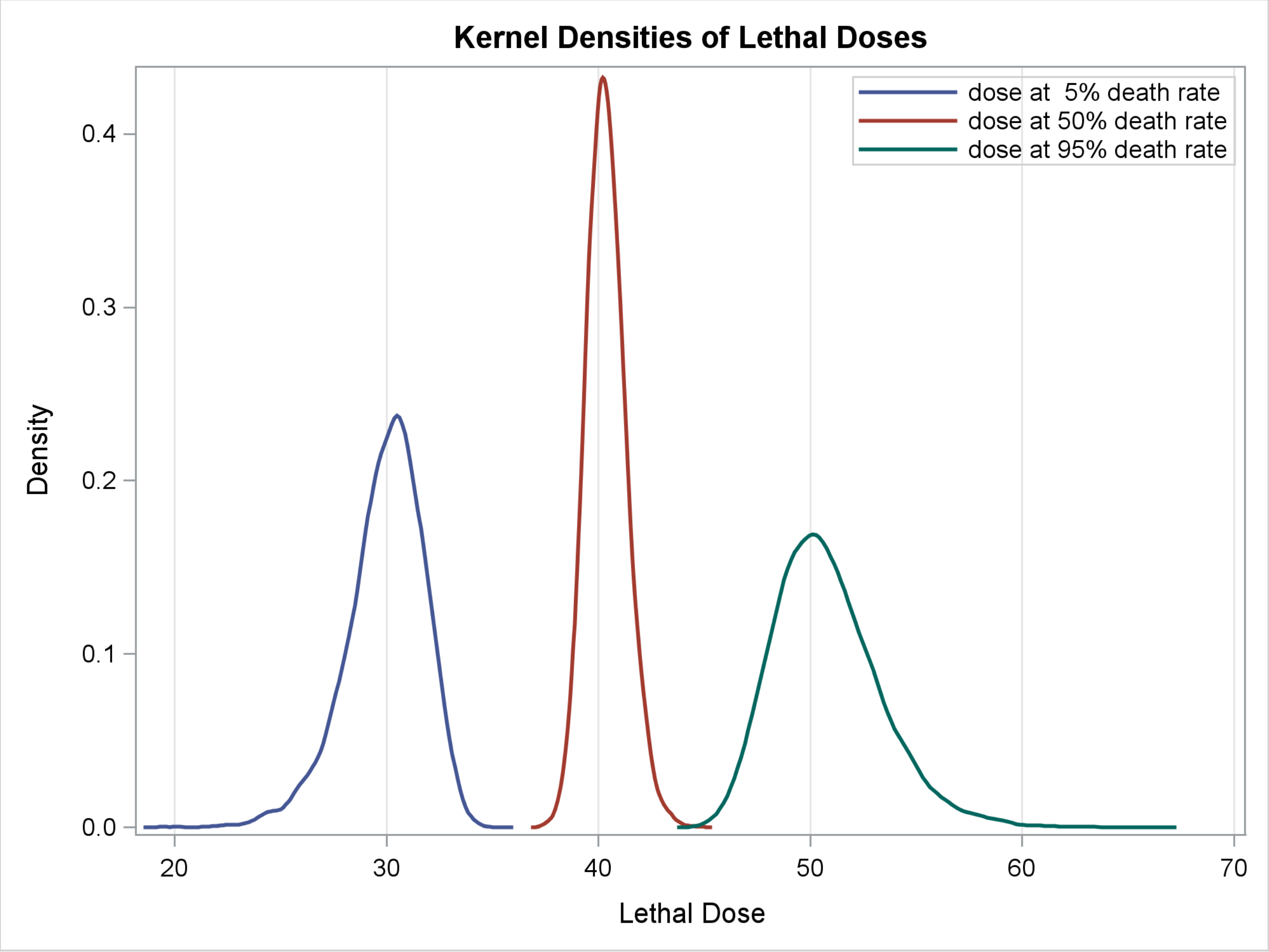
Table 9 Lethal Dose Estimates for Beetle Data

Lethal Dose	Posterior Mean	95% Equal-Tailed Interval
LD5	29.9404	(25.7704, 32.9615)
LD50	40.3664	(38.6752, 42.3798)
LD95	50.7924	(46.7259, 56.5187)

The SGPLOT procedure displays these results graphically with overlaid kernel density plots.

```
proc sgplot data=LDgenerate;
   title "Kernel Densities of Lethal Doses";
   xaxis label = "Lethal Dose" grid;
   density LD05 / type = kernel legendlabel="dose at  5% death rate";
   density LD50 / type = kernel legendlabel="dose at 50% death rate";
   density LD95 / type = kernel legendlabel="dose at 95% death rate";
   keylegend / location=inside position=topright across=1;
run;
```

Figure 26 Kernel Densities for Beetle LDs



The use of Bayesian methods for categorical data analysis is an active area of reseach and application. For example, see Chapter 11 in Agresti (2010) for a discussion of Bayesian methods applied in ordinal categorical data analysis.

Agreement Plots

Observer agreement studies have been a mainstay in many fields such as medicine, epidemiology, and psychological measurement and testing in which observer error can be an important source of measurement error. Sometimes, different observers evaluate the same subject, image, or test and reach different conclusions. Evaluating observer agreement is thus important both for understanding the sources of measurement error and as part of a protocol for evaluating new procedures.

The kappa coefficient is a statistic used to evaluate observer agreement (Cohen 1960).

Suppose π_{ij} is the probability of a subject being classified in the ith category by the first observer and the jth category by the second observer. Then

$$\Pi_o = \sum \pi_{ii}$$

is the probability that the observers agree. If the ratings are independent, then the probability of agreement is

$$\Pi_e = \sum \pi_{i+}\pi_{+i}$$

So, $\Pi_o - \Pi_e$ is the amount of agreement beyond that expected by chance. The kappa coefficient is defined as

$$\kappa = \frac{\Pi_o - \Pi_e}{1 - \Pi_e}$$

Since $\Pi_o = 1$ when there is perfect agreement (all non-diagonal elements are zero), κ equals 1 when there is perfect agreement, and κ equals 0 when the agreement equals that expected by chance. The closer the value is to 1, the more agreement there is in the table.

A graph can help to interpret the underlying agreement associated with a particular estimated kappa coefficient, and the FREQ procedure provides the agreement plot with SAS/STAT 9.3. The following example is from p. 112 in Stokes et al. 2000. Patients with multiple sclerosis were classified into four diagnostic classes by two neurologists.

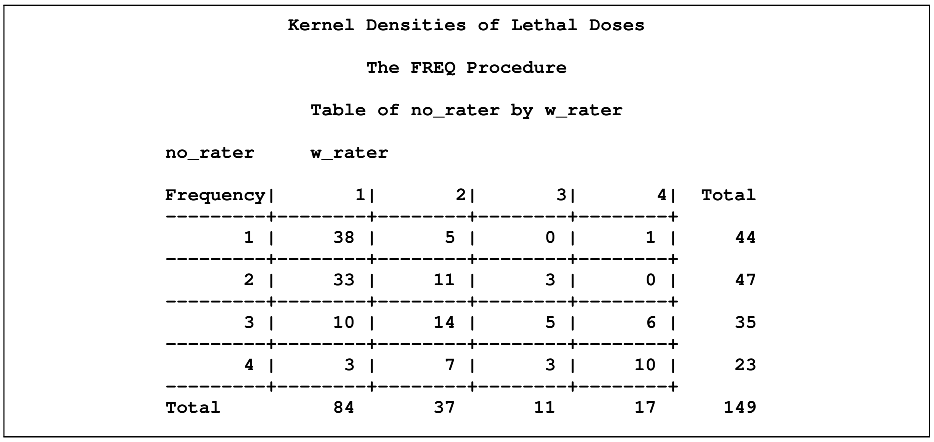
The following code creates SAS data set CLASSIFY.

```
data classify;
   input no_rater w_rater count @@;
datalines;
1 1 38 1 2  5 1 3 0 1 4  1
2 1 33 2 2 11 2 3 3 2 4  0
3 1 10 3 2 14 3 3 5 3 4  6
4 1  3 4 2  7 4 3 3 4 4 10
run;
```

PROC FREQ procedures the corresponding contingency table and the kappa coefficient. The agreement plot is produced when ODS Graphics is enabled:

```
ods graphics on;
proc freq;
   weight count;
   tables no_rater*w_rater / agree norow nocol nopct;
run;
ods graphics off;
```

Figure 27 Classified Patients



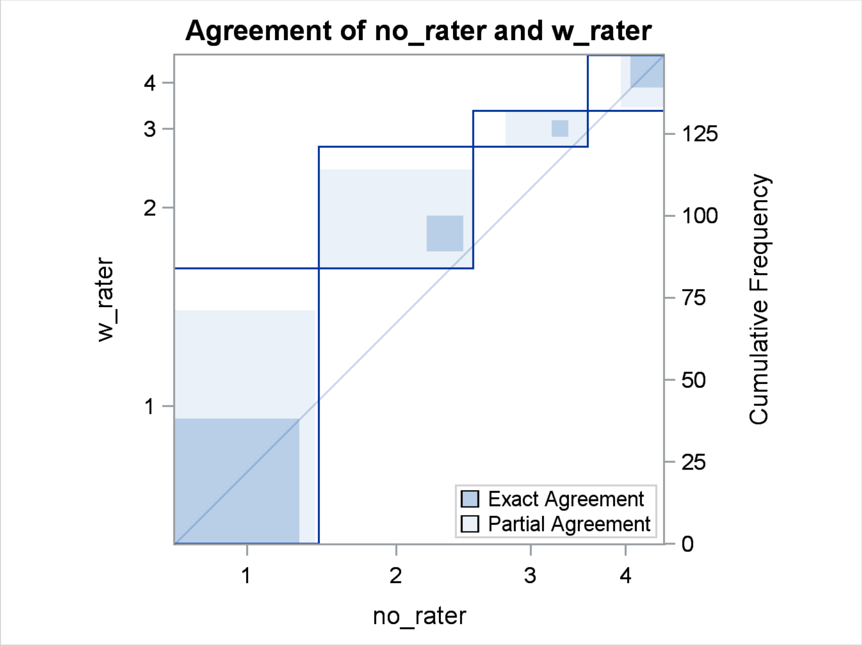
```
                    Kernel Densities of Lethal Doses

                          The FREQ Procedure

                     Table of no_rater by w_rater

          no_rater     w_rater

          Frequency|       1|       2|       3|       4|  Total
          ---------+--------+--------+--------+--------+
                 1 |     38 |      5 |      0 |      1 |     44
          ---------+--------+--------+--------+--------+
                 2 |     33 |     11 |      3 |      0 |     47
          ---------+--------+--------+--------+--------+
                 3 |     10 |     14 |      5 |      6 |     35
          ---------+--------+--------+--------+--------+
                 4 |      3 |      7 |      3 |     10 |     23
          ---------+--------+--------+--------+--------+
          Total          84       37       11       17      149
```

Figure 28 Kappa Coefficient

```
                    Kappa Statistics

Statistic          Value       ASE      95% Confidence Limits
-------------------------------------------------------------
Simple Kappa      0.2079    0.0505       0.1091       0.3068
Weighted Kappa    0.3797    0.0517       0.2785       0.4810
```

The estimated kappa coefficient has the value 0.2079, which indicates slight agreement. The agreement plot displays the nature of the agreement (Bangdiwala and Bryan 1987).

Figure 29 Agreement Plot for MS Patient Study



Conclusion

Like other areas in statistics, categorical data analysis continues to evolve with time, and thus so does the software that implements it. The strategy of assessing association in a data set often goes hand-in-hand with a modeling strategy for describing it and methods for both strategies have been refined. The computing advances that make exact methods and Bayesian methods feasible are often the driving force for advances in methodology, but so is the continuing research that provides more information about the properties of the statistical methods and allows one to determine best practices. This paper illustrates some of the newer methods that are now available and describes how to apply them to typical situations with SAS software.

References

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

Agresti, A. (1992), “A Survey of Exact Inference for Contingency Tables”, *Statistical Science*, 7, 131–177.

Agresti, A. (2010), *Analysis of Ordinal Categorical Analysis*, Second Edition, New York: John Wiley & Sons.

Agresti, A. and Coull, B. A. (1998), “Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions,” *The American Statistician*, 52, 119–126.

Agresti A. and Min, Y. (2001), “On Small-Sample Confidence Intervals for Parameters in Discrete Distributions,” *Biometrics*, 57, 963–971.

Bangdiwala, K., and Bryan, H. (1987), “Using SAS software graphical procedures for the observer agreement chart,” *Proceedings of the SAS User’s Group International Conference*, 12, 1083–1088.

Chan, I. S. F., and Zhang, Z. (1999), “Test-Based Exact Confidence Intervals for the Difference of Two Binomial Proportions,” *Biometrics,* 55, 1202-1209.

Cohen, J. (1960), A coefficient of agreement for nominal data, *Educational and Psychological Measurement*, 20, 37–46.

Derr, R. (2009), “Performing Exact Logistic Regression with the SAS System–Revised 2009,” SAS Institute, Inc. support.sas.com/rnd/app/papers/exactlogistic2009.pdf

Farrington, C. P. and Manning, G. (1990), “Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk,” *Statistics in Medicine*, 9, 1447–1454.

Fleiss, J. L., Levin, B., and Paik, M. C. (2003), *Statistical Methods for Rates and Proportions*, Third Edition, New York: John Wiley & Sons.

Koch, G. G., Amara, I. A., Davis, G. W., and Gillings, D. B. (1982), “A review of some statistical methods for covariance analysis of categorical data,” *Biometrics*, 38(3), 553–595.

Koch, G. G., Carr, G. J., Amara, I. A., Stokes, M. E., and Uryniak T. J. (1990), “Categorical Data Analysis,” in Berry, D. A. ed., *Statistical Methodology in Pharmaceutical Sciences*, New York: Marcel Dekker, 291–475.

Koch, G. G., Tangen, C. M., Jung, J-W, and Amara, I. A. (1998), “Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them,” *Statistics in Medicine*, 17, 1163–1892.

LaVange, L. M., Durham, T. A., and Koch, G. G. (2005), “Randomization-based nonparametric methods for the analysis of multicentre trials,” *Statistical Methods in Medical Research*, 14, 281–301.

LaVange, L. M. and Koch, G. G. (2008), “Randomization-Based Nonparametric (ANCOVA),” *Encyclopedia of Clinical Trials*, 31–38.

Newcombe, R. G. (1998), “Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods,” *Statistics in Medicine*, 17, 873–890.

Saville, B. R., LaVange, L. M., and Koch, G. G. (2011), “Estimating Covariate-Adjusted Incidence Density Ratios for Multiple Intervals using Nonparametric Randomization-Based ANCOVA,” *Statistics in Biopharmaceutical Research*, (in press).

Saville, B. R., Herring, A. H., and Koch, G. G. (2010), “A Robust Method for Comparing Two Treatments in a Confirmatory Clinical Trial via Multivariate Time-to-event Methods that jointly incorporate information from Longitudinal and Time-to-event Data,” *Statistics in Medicine*, 29, 75–85.

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

Tangen, C. M. and Koch, G. G. (1999), “Complementary nonparametric analysis of covariance for logistic regression in a randomized clinical trial setting,” *Journal of Biopharmaceutical Statistics*, 9, 45–66.

Tangen, C. M. and Koch, G. G. (2001), “Non-parametric analysis of covariance for confirmatory randomized clinical trials to evaluate dose-respons relationships,” *Statistics in Medicine*, 20, 2585–2607.

Vesikari, T., Itzler, R., Matson, D. O., Santosham, M., Christie, C. D. C., Coia, M., Cook, J. R., Koch, G., and Heaton, P. (2007), "Efficacy of a pentavalent rotavirus vaccine in reducing rotavirus-assocated health care utilization across three regions (11 countries)," *International Journal of Infectious Diseases*, 11, 528–534.

Zink, R. C. and Koch, G. G., (2001), *NparCov Version 2*, Chapel Hill, North Carolina: Biometric Consulting Laboratory, Department of Biostatistics, University of North Carolina.

Appendix

The following SAS/IML code produces confidence limits for the difference of proportions using Newcombe's method for the continuity corrected case.

```
proc iml;
z = 1.96;
y = 10; n = 12; p=y/n; q=1-p; p1=y/n;
l1 = (2*n*p + z**2 -1-z*sqrt( z**2 - 2-1/n+4*p*(n*q+1) )) /(2*(n+z**2));
u1 = (2*n*p + z**2 +1+z*sqrt( z**2 + 2-1/n+4*p*(n*q-1) )) /(2*(n+z**2));
y=2; n=6; p = y/n; q=1-p; p2=y/n;
l2 = (2*n*p + z**2 -1-z*sqrt( z**2 - 2-1/n+4*p*(n*q+1) )) /(2*(n+z**2));
u2 = (2*n*p + z**2 +1+z*sqrt( z**2 + 2-1/n+4*p*(n*q-1) )) /(2*(n+z**2));
Lp1p2 = (p1-p2) - sqrt( (p1-l1)**2 + (u2-p2)**2 );
16
Up1p2 = (p1-p2) + sqrt( (u1-p1)**2 + (p2-l2)**2 ) ;
p1p2 = p1-p2;
print "Single Proportion Group1" p1 l1 u1;
print "Single Proportion Group2" p2 l2 u2;
print "Difference p1-p2" p1p2 Lp1p2 Up1p2;
quit;
```

ACKNOWLEDGMENTS

The authors are very grateful to Fang Chen for his contributions to the Bayesian section. We also thank Donna Watts and Bob Derr for their comments and suggestions in their respective areas. We thank Allen McDowell and David Schlotzhauer for their review of this paper, and we thank Anne Baxter for her editorial assistance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Maura Stokes
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
maura.stokes@sas.com