PharmaSUG China 2019 - Paper PO-44

Comparison of SAS Procedures for Categorical Analyses in Binominal Test

Wang, Yueqing, Li, Daoqing, Pfizer (China) Research & Development Co. Ltd

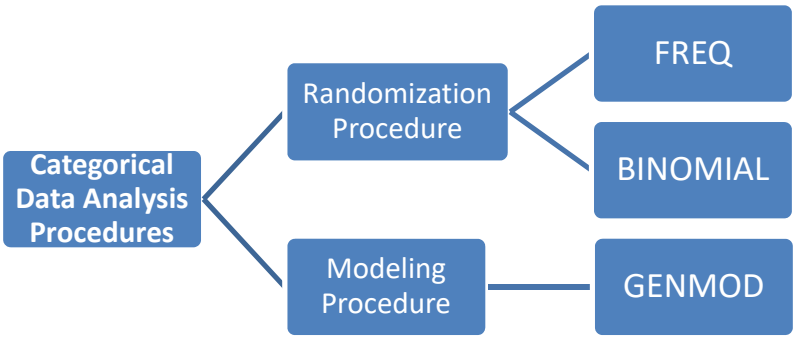
ABSTRACT

Categorical data analysis is one of the most commonly used analyses in clinical trials. Three approaches are mainly adopted in SAS to compute categorical results. While in the situation of absence of required response level, the approach of obtain CIs and p values is quite different. 1) PROC FREQ is the most well-known procedure to calculate CIs for Binomial proportions. 2) PROC GENMOD is another approach utilizing generalized linear model for categorical variables. 3) PROC BINOMIAL is a retired but still workable procedure for binary data. In this poster, these three SAS procedures will be thoroughly evaluated and compared, especially in the application of exact test and Monte Carlo simulation in then binomial test.

INTRODUCTION

There are two approaches to performing categorical data analyses. The first computes statistics based on tables defined by categorical variables (variables that assume only a limited number of discrete values), performs hypothesis tests about the association between these variables, and requires the assumption of a randomized process; call these methods randomization procedures. The other approach investigates the association by modeling a categorical response variable, regardless of whether the explanatory variables are continuous or categorical; call these methods modeling procedures.

The poster focuses on FREQ, GENMOD and BINOMIAL. FREQ and BINOMIAL belong to randomization procedure, while GENMOD belongs to modeling procedure.



FREQ

The procedure builds frequency tables or contingency tables and can produce numerous statistics.

For one-way frequency tables, it provides goodness-of-fit tests for equal proportions or specified null proportions. For one-way tables, PROC FREQ also provides confidence limits and tests for binomial proportions, including tests for noninferiority and equivalence.

For contingency tables (any size two-way table), it can compute various statistics to examine the relationships between two classification variables. For some pairs of variables, you might want to examine the existence or strength of any association between the variables. To determine if an association exist, PROC FREQ computes chi-square test. To estimate the strength of an association, PROC FREQ computes measures of association that tend to be close to zero when there is no association and close to the maximum (or minimum) value when there is perfect association. The statistics for contingency tables include the following:

- Chi-square tests and measures

- Measures of association
- Risk (binomial proportions) and risk differences for 2×2 tables
- Odds ratios and relative risks for 2×2 tables
- Tests for trend
- Tests and measures of agreement
- Cochran-Mantel-Haenszel statistics

SAS syntax for FREQ procedure in Binominal test:

```
PROC FREQ DATA=<DATASET>;
    TABLES <OUTCOME VARIABLE>/BIMOMIAL<(BIMOMIAL OPTIONS)> ALPHA=<value>;
    WEIGHT <WEIGHT VARIABLE>;
RUN;
```

GENMOD

The GENMOD procedure fits a generalized model which extends the traditional linear model and it therefore applicable to a wider range of data analysis problems. A generalized linear model consists of the following components:

- The linear component is defined just as it is for traditional linear models:

  $\eta_i = \mathcal{X}_i'\beta$

- A monotonic differentiable link function g describes how the expected value of y_i is related to the linear predictor η_i:

  $g(\mu_i) = \mathcal{X}_i'\beta$

- The response variables y_i are independent for $i = 1,2, \ldots$ and have a probability distribution from an exponential family. This implies that the variance of the response depends on the mean μ through a variance function $\mathcal{V}$:

  $\mathrm{Var}(y_i) = \frac{\emptyset V(\mu_i)}{w_i}$

Where $\emptyset$ is a constant and w_i is known weight for each observation. The dispersion parameter $\emptyset$ is either know (for example, for the binomial or Poisson distribution, $\emptyset$=1) or must be estimated.

GENMOD procedure fits a generalized linear model to the data by maximum likelihood estimation of the parameter vector β. There is, in general, no closed form solution for the maximum likelihood estimates of the parameters. The GENMOD procedure estimates the parameters of the model numerically through an iterative fitting process. The dispersion parameter $\emptyset$ is also estimated by maximum likelihood or, optionally, by the residual deviance or by Pearson's chi-square divided by the degrees of freedom. Covariances, standard errors, and p-values are computed for the estimated parameters based on the asymptotic normality of maximum likelihood estimators. A number of popular link functions and probability distributions are available in the GENMOD procedure.

SAS syntax for GENMOD procedure in Binominal test:

```
PROC GENMOD DATA=<DATASET> DESCENDING;
    MODEL <VAR1>=<VAR2>/LINK=<LINK OPTIONS> DIST=BINOMIAL;
RUN;
```

Link options specify the link function to use in the model. The key work for link options can be LOG, LOGIT, IDENTITY, PROBIT…

DIST options can also be other distribution keyword according to the models for data with correlated responses, e.g. NOMAL, GAMMA, POISSON….

BINOMIAL

StatXact is a software product from Cytel Inc.. They developed a set of SAS procedure that implement a range of exact methods, including BINOMIAL procedure. The procedure is an effective way to compute confidence interval of a binomial proportion and confidence interval for risk difference.

SAS syntax for Confidence Interval of a Binomial Proportion (Blyth-Still-Casella):

```
PROC BINOMAIL DATA=<DATASET> ALPHA=<value>;
    BI/BS;
    OU <RESPONSE VARIABLE>;
RUN;
```

SAS syntax for Confidence Interval for Risk Difference using Chan and Zhang:

```
    PROC BINOMAIL DATA=<DATASET> GAMMA=0 ALPHA=<value>;
    PD/EX ONE STD;
    PO <POPULATION VARIABLE>;
    OU <OUTCOME VARIABLE>;
RUN;
```

CONCLUSION

A categorical variable is a variable that assumes only a limited number of discrete values. The measurement scale for a categorical variable is unrestricted. It can be binomial, which means that the observed data are in two level. The FREQ, GENMOD and BINOMIAL procedures can compute statistics in binomial test.

PROC FREQ is similar with PROC BINOMIAL, they are used primarily to investigate the relationship between two variables; any confounding variables are taken into account by stratification rather than by parameter estimation. The two procedures compute statistics 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. And due to the hypergeometric distribution, statistics in binomial test from FREQ and BINOMIAL are exact test.

PROC GENMOD is a modeling procedure that estimates the covariance matrix of the frequencies, it assumes that the frequencies were obtained by a stratified simple random-sampling procedure. The procedure requires that the response variable be categorical-the explanatory variables are allowed to be continuous or categorical.

REFERENCES

Agresti, A. 2013. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Agresti, A., & Gottard, A. (2007). Nonconservative exact small-sample inference for discrete data. *Computational statistics & Data analysis*, *51*(12), 6447-6458.

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

Barnard, G. A. (1947). Significance tests for 2× 2 tables. *Biometrika*, *34*(1/2), 123-138.

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

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

Hilbe, J. M. 2007. *Negative Binomial Regression*. New York: Cambridge University Press.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, *15*(3), 209-225.