# Analysis of dichotomous outcomes in small datasets: assessment of statistical methods in PROC LOGISTIC

Charles W. Goss

Washington University School of Medicine, Division of Biostatistics, 660 S. Euclid Ave, St. Louis, 63110

## Background

In biomedical research, logistic regression is commonly employed to analyze dichotomous outcomes. Although analysis and interpretation using logistic regression can be fairly straightforward with large data sets, in small datasets the asymptotic normality assumption of maximum likelihood parameter estimates (MLE) becomes untenable leading to biased parameter estimates and problems with model convergence due to complete or quasi-complete separation. In order to analyze these small datasets, exact methods and Firth's penalized maximum likelihood estimates (FMLE) are often employed.

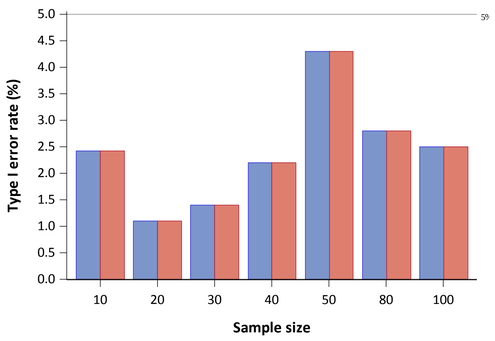
In SAS, perhaps the most commonly used procedure for the analysis of dichotomous outcomes is PROC LOGISTIC. PROC LOGISTIC provides several *P*-value estimates associated with exact methods as well as confidence interval estimates based on the conditional maximum likelihood estimation. In Simulation 1, we report results from Monte Carlo simulations aimed at assessing the consequences of 2 of these different *P*-value estimates for statistical inference. In addition to exact methods, PROC LOGISTIC uses FMLE which has been proposed as an alternative to exact methods for the analysis of small datasets. Even though it has been argued that these small-sample approaches should be used to analyze dichotomous outcomes regardless of sample size, the benefits and drawbacks of these methods are not widely known (but see Heinze 2006). In Simulation 2, we use a simple 1 dummy variable model and a more complex 3 dummy variable model to illustrate how changing sample size and model complexity influence the type I error rates and power for exact, Firth, and "standard" maximum likelihood estimation methods.

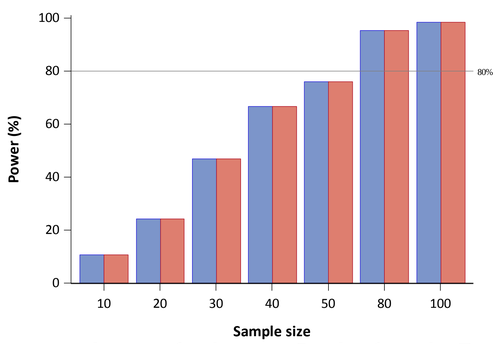
## Simulation methods

In the 1 dummy variable models for Simulations 1 and 2 we assessed 2 different *P*-value estimates from PROC LOGISTIC exact methods by generating data from a model with the linear predictors of 0.05 (intercept) and a coefficient of 2 for the dummy variable. For the 3-dummy variable model (Simulation 2), data were generated from the linear predictors: -1 for the intercept, and 2, 1.5, and -1 for the dummy variable coefficients. Linear predictions on the log-odds scale were back-transformed to the probability scale and then used to generate random samples from the Bernoulli distribution. Power was estimated by analyzing the data generated by these models and calculating the percentage of simulations that correctly rejected the false null hypothesis ($H_o$: all regression coefficients = 0; reject $H_o$ if $P \leq 0.05$). We assessed type I error rates by generating a single population outcome with a linear predictor equal to 0.4 (probability of "success" = 0.60) and analyzed the number of times the analysis falsely rejected the null hypothesis for 1 and 3 dummy variable models. For all simulations we used 1,000 replications of 7 different sample sizes (10, 20, 30, 40, 50, 80, 100) and estimated *P*-values from PROC LOGISTIC using results from the *Exact Conditional Tests* (score test *P*-value) and *Exact Odds Ratio* (two-sided *P*-value) tables and the *Global* table *P*-values for the Firth (FMLE) and default maximum likelihood methods (MLE).

## Simulation 1: exact logistic regression and *P*-value estimates

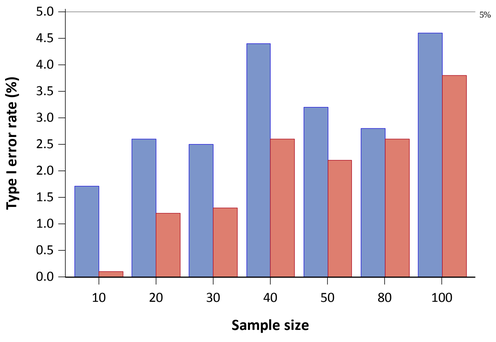

### Balanced data

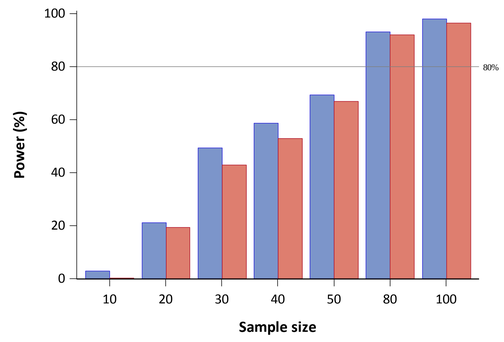



For the balanced-data simulations, both power and type I error rate were the same for the score *P*-value from the *Exact Conditional Tests* table and the two-sided *P*-value from the *Exact Odds Ratio* table.

### Unbalanced data

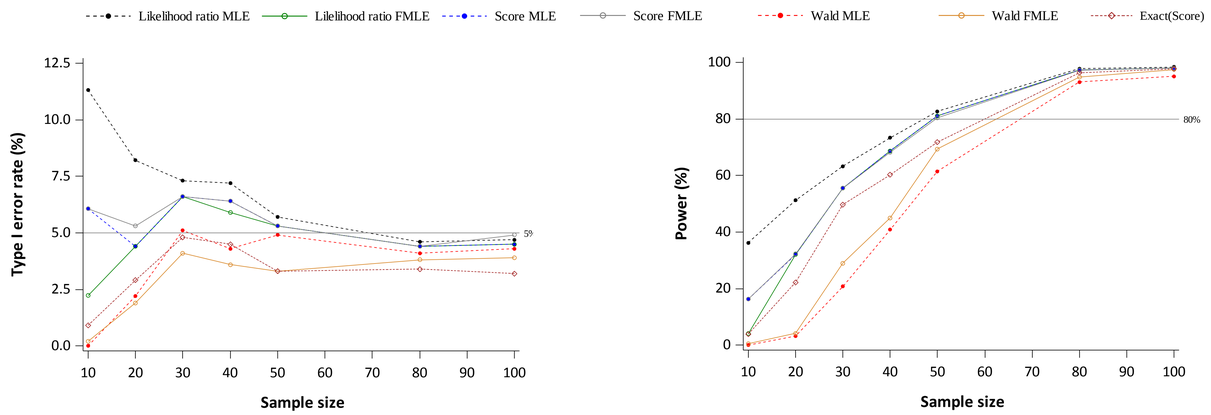



When we compared the test statistics using unbalanced data, the type I error rate for revealed that the *P*-value from the *Exact Odds Ratio* table was more conservative than the score *P*-value from the *Exact Conditional Tests* table (0.2-1.8% difference) whereas the power from the latter had as much as 6 % greater power than the former.

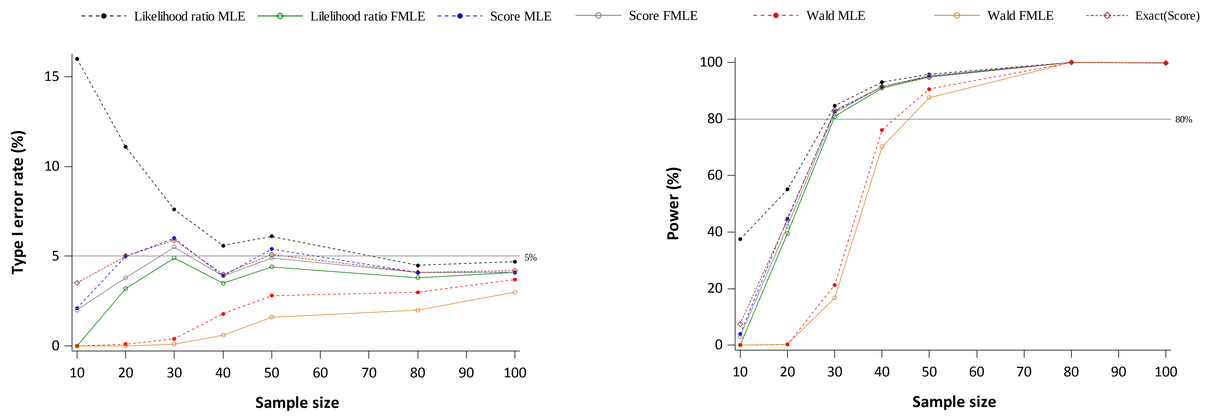
## Simulation 2: exact, Firth, and maximum likelihood estimation

### 1 dummy variable model




For the 1 dummy variable model, our results indicate that the exact and Wald test statistics had the lowest type I error rate (≤ 5% for all sample sizes) whereas the likelihood ratio MLE had the greatest type I error rate approaching 12%. The exact and Wald statistics had lower power than the other MLE and FMLE statistics across all sample sizes, reaching a maximum difference of 30 and 50% lower power for the exact and Wald statistics, respectively.

### 3 dummy variable model




For the 3 dummy variable model, our results revealed that all of the statistics except for the likelihood ratio MLE (which had a max. of > 15%) were at or below the nominal 5% type I error rate across all sample sizes. The power for the likelihood ratio test was greater than all other statistics, but the FMLE and exact test statistics converged with the likelihood ratio MLE at sample sizes > 20. The power of the Wald statistics was markedly lower than all other statistics at sample sizes less than 80.

## Conclusions

The results of this simulation study indicate the importance of thoroughly understanding the assumptions, strengths, and weaknesses of statistical procedures in SAS prior to applying them. In Simulation 1, the subtle difference in calculations of exact methods *P*-values is revealed by variation in data structure (balanced vs. unbalanced) resulting in the exact score statistic (provided in the *Exact Conditional Tests* table) having up to 6% more power to detect effects than the two-sided *P*-value reported from the *Exact Odds Ratio* table. Given balanced data, the user might erroneously assume that these statistics are equivalent for all situations (as are the Wald statistics for parameter estimates and *Global* statistics tables). In Simulation 2, the exact and FMLE type I error rates were near or below the nominal type I error rate (5%), whereas the likelihood ratio MLE typically had the highest type I error rate; however, the power of the exact test statistics (esp. for the 1 dummy variable model) were often much lower than the MLE and FMLE statistics with a maximum of 32% lower power than the MLE likelihood ratio test. At sample sizes < 80, the Wald statistic performed poorly with a mean maximum difference of about 30% lower power than the other statistics.

Our simulations were limited in that we only investigated PROC LOGISTIC *P*-value estimates for specific, simple scenarios. Nevertheless, our results were suggestive of more general patterns including elevated type I error rates for the MLE likelihood ratio statistic at small sample sizes that converge on the nominal rate at intermediate sample sizes while still retaining greater power than other statistics. Our results also showed that the FMLE score statistic had perhaps the best overall performance by minimizing the type I error rate across a range of sample sizes while retaining a power ≥ the exact score statistic. Future studies should incorporate additional statistics and more complex models to more comprehensively assess the influence of small sample sizes and estimation methods on statistical inference.

**Acknowledgements**: Thanks to members of the Schechtman lab for fruitful discussions involving logistic regression and small datasets that inspired this project. Also, thanks to SAS customer support for clarifying the discrepancies between *P*-value estimates in PROC LOGISTIC.

**References**

Allison, P.D., 2012. Logistic Regression using SAS®: Theory and Application. SAS® Institute Inc., Cary, NC.

Heinze, G., 2006. A comparative investigation of methods for logistic regression with separated or nearly separated data. Statistics in Medicine 25, 4216-4226.

Stokes, M.E., Davis, C.S., Koch, G.G., 2012. Categorical Data Analysis using SAS®. SAS® Institute, Cary, NC.