PharmaSUG 2012 – Paper HO08

ESTIMATING RISK IN THE PRESENCE OF UNREPORTED ZEROS IN THE DATA

Brandon Fleming, UMBC, Baltimore, Maryland

ABSTRACT

When evaluating risk of injury (exposure) using sample data alone, it is desirable to have as the denominator the total population size. However, if the sample itself contains only information about the frequency of exposure (e.g. frequency of N=1 exposure, 2, 3, etc.), and those who have never been exposed would not be entered into the database, how is one to ascertain the unreported zeros (frequency of N=0)? This paper explores the use of PROC NLMIXED in estimating the risk using the truncated Poisson assumption. The performances of three candidate estimators are analyzed with simulated data that contains frequency information. PROC NLMIXED is used first to identify the truncated Poisson mean, and then to estimate the frequency of N=0, i.e. those without injury or exposure (unreported zeros).

Abstract Keywords:
Estimating unobserved zeroes, parameter estimation, evaluating risk, simulation models, occupational injury, Paramedic/EMT

INTRODUCTION

One of the hallmarks of statistics is to extrapolate from a sample, both precise and accurate information about a comparatively larger population. If one is constructing a study *de novo*, there are many statistical techniques available to ensure a representative sample. However, when one inherits a data set there are limitations on the kind and nature of the questions that can be explored. Sometimes the samples in these data sets are statistically not representative, or biased. Nonetheless, it is of some interest to statisticians to develop rigorous analytical tools to analyze these types of data sets.

One example in the pharmaceutical industry is the databases used for reporting adverse drug reactions (ADR). Only individuals who actually experience an ADR are going to report it. The problem with these types of databases is two-fold. The first is underreporting: individuals not reporting or not recognizing an ADR. The second is that the population of patients on a particular medication is unknown.

The impetus for this paper however is the risk of occupational injury. Estimating the risk of certain occupational injuries is of some concern to many invested parties. For example, insurance companies would like more accurate analysis to better calculate premiums. More importantly, employees and their managers would like to know which injuries people in certain occupations are prone to. This information can lead to a safer work environment.

However, the estimation of risk is problematic when the population size is unknown. This problem arises in particular, when the data is available only for incidences and no information is available regarding the part of the population which is exposed to the risk but has not produced any observations. This project will explore techniques of calculating the risks of occupational injuries from an unknown population size. The specific occupation is Emergency Medical Technicians (EMT) and Paramedics. The data set is from an unknown, multistate company database.
The first problem is estimating the frequency of unreported zeros. Several techniques from the literature will be explored:

- Homogeneous truncated Poisson
- Heterogeneous truncated Poisson
- Zelterman Estimators and Horvitz-Thompson estimates
- Computational methods using SAS (PROC NLMIXED)

The original number of observations in the data set is 6691. This research paper focuses on the injury rates of emergency workers in general and pre-hospital specifically. Through SAS filtering processes, all extraneous occupations (defined here to mean non-EMT or Paramedic positions) were eliminated (reduced from 600 to 309 distinct occupations). The final sample size used is 5262 observations.

The data we use is a specific form of capture-recapture data. In capture-recapture data captures and recaptures are at specific time points, and for each animal seen at least once there is a capture history. For example if there are five capture times a history could be 01101 if the animal is seen at captures 2, 3, and 5 and not seen at 1and 4 [Heijden 2003]. Here, we use only the total number of times someone is injured (and reports it) since the data was collected in continuous time.

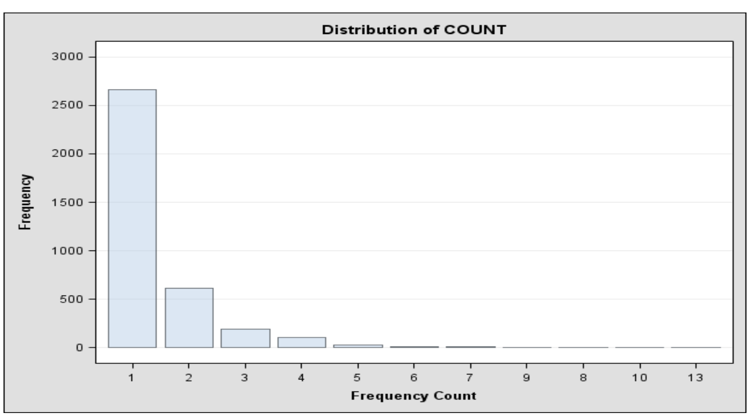
To ensure that all the cells are at least 5, N=8 thru 13 were combined in the subsequent analysis to give N ≥ 8 = 10. Note that the total number of injuries (1x2666+2x611+…) is 5250. Consequently combining count 8 thru 13 into a single group results in a negligible loss of data. As indicated on Table 1, N=0 is unobserved. An estimate for N=0 is needed to estimate the risk for the total population.

Below is a table of count data (Diagram 1) where N is the number of injuries and Frequency is the frequency of individuals who experience this number of accidents. For example 611 providers in the data set had 2 reported injuries.

Table 1: Count data for number of injuries

Frequency Count	
N	**Frequency**
1	2666
2	611
3	195
4	101
5	29
6	13
7	10
≥8	10
Total	**3635**

Figure 1: Count data of number of injuries with frequencies (to ensure that all the cells are at least 5, N=8 thru 13 were combined in the subsequent analysis to give N ≥ 8 = 10)



The origin of this paper is the question: What is a valid estimate of N=0?

ESTIMATORS OF LAMBDA FOR A HOMOGENEOUS POISSON DISTRIBUTION

Towards developing an estimate for the zero frequency, the number of injuries suffered by the 3635 distinct EMT workers are assumed follow a Poisson distribution with mean λ. The maximum likelihood estimator (MLE) for a Poisson mean is, $\hat{\lambda} = \frac{\sum x_i}{n}$. For the above count data

$\hat{\lambda} = \frac{5262}{3635} = 1.45$

However N=0 is unobserved and therefore the above estimate is likely an overestimate of the mean parameter, and therefore has to be modified. Thus we use a truncated Poisson model for the observed data, by truncating the zero probability term from the probability function. Using a SAS [see below PROC NLMIXED calculation of MLE, p 4], we obtain $\hat{\lambda}$=0.7887.Setting $\widehat{p_0}$ = exp (-$\hat{\lambda}$), an estimate for population size [Heijden 2003] is

$$\widehat{N} = \widehat{f_0} + N_{obs} \text{ where } \widehat{f_0} = \frac{\widehat{p_0}}{1- \widehat{p_0}} N_{obs}$$

An alternative to the truncated Poisson approach is using the Zelterman estimator. The Zelterman estimator is known to be robust under potential unobserved heterogeneity [Bohning and Heijden 2009]. This estimator is often considered for two reasons: for one, $\widehat{\lambda_1}$ is using frequencies in the vicinity of f_0 which is the target of prediction, and two, in many applications studies (especially this one) for estimating f_0 the majority of counts fall into f_1 and f_2. Clearly, the estimator is unaffected by changes in the data for counts larger than 2, which contributes largely to its robustness. When used to calculate the population size, N=0, the Zelterman estimator applied to the Horvitz-Thompson estimator. The Horvitz-Thompson estimator is,

$$\widehat{N} = \sum_{i=1}^{n} \frac{1}{P_i(Y>0)} = \frac{1}{1- P_i(Y=0)} = \frac{n}{1-g(\lambda)}$$

Where g (λ) =$e^{-\lambda}$, or more generally, g (λ) is the probability of a zero count for a given count distribution. Both the truncated Poisson method and Horvitz-Thompson estimator approach are useful when the distribution follows a *homogeneous* truncated Poisson distribution. A Poisson distribution has the property that, Pr(μ |X) = Pr(σ^2 | X). However when overdispersion is present, Pr(μ |X) < Pr(σ^2 | X). What happens when *overdispersion* is present? The most common source of overdispersion with count data is *unobserved heterogeneity*, i.e. variables and factors that have not been incorporated into the model. Can be detected with:

- Lagrange Multiplier test
- Pearson Residuals

The Zelterman is actually a special case of a more general family of estimators where j=k=1. One of the benefits of the Zelterman estimator is that it is better able to deal with unobserved heterogeneity. Because we are dealing here with count data, the assumption is that that P(X) is a Poisson distribution. For this type of model j and k are the number of injuries an individual person experiences. The variable N=0 is replaced with X=0 below, i.e. the number of individuals with no injuries is unobserved. To estimate X=0 we can calculate λ for the truncated Poisson distribution. A possible estimator for λ using observed values, i.e. (j, k) > 0 is:

$$\frac{P(X=j+k)}{P\ (X=j)} = \frac{e^{-\lambda}\left(\lambda^{j+k}/_{(j+k)!}\right)}{e^{-\lambda}\left(\lambda^{j}/_{j!}\right)} = \lambda^k (j!)\Big/(j+k)!$$

The above equation can be rearranged to,

$$\frac{(j+k)!}{j!}\frac{P(X=j+k)}{P(X=j)} = \lambda^k$$

The expressions for P(X=x) are replaced with F_x and solving for λ,

$$\left[\frac{(j+k)!}{j!}\ \frac{F_{j+k}}{F_j}\right]^{1/k} = \lambda \qquad \text{(GFEZE)}$$

This is considered an estimated unbiased estimator of λ. Furthermore, if j=k=1 is then

$$\left[\frac{(1+1)!}{1!}\ \frac{F_2}{F_1}\right]^{1/1} = \frac{2F_2}{F_1} = \lambda$$

is the original Zelterman estimator. The advantage of this more general family is that depending on how j and k are selected, the estimator will always exist. The restriction on the Zelterman estimator is that for high values of λ, the frequency of X=1 and/or X=2 may not be present in the data, simulation models or otherwise. The above equation will be referred to as the General Family Equations of Zelterman Estimators (GFEZE).

COMPARISON OF ESTIMATORS FOR LAMBDA IN SIMULATED DATA

The actual injury count data above has too many covariates and violated assumptions. Therefore to start, several data tables were generated under the Poisson assumption using simulation techniques. In the real world, only individuals who have actually been injured are entered into an injury database. Those without injuries are excluded. The data was meant to simulate the number of injuries that an individual could potentially experience, including those without any injuries. The program firsts generates 1,000 separate data sets with 500 numbers each using a Poisson distribution with parameter=1 (note the parameter, lambda is known). The variable X is the number of injuries one individual experience. In each of the data sets, X=0, 1, 2 are identified (i.e. the number of individuals with no injuries, 1 injury, and 2 injuries). The data sets are then sorted and the total number of X=0 (n0), X>0 (n_obs), X=1 (F1), and X=2 (X=2) is calculated for each data set (MC_num) and outputted to TruncPoiMeans. Then X=0 (unobserved) is isolated from X>0 (observed) and subsequently deleted to create truncated Poisson distributions for each data set. The frequency of X=0 (no injuries) constitute the population size and is the statistic that needs to be estimated. An example of an output is summarized as follows: (note 500-1,000 different tables were generated)

Table 1: Example of a SAS simulation table used in the analysis (1,000 tables generated for the estimates)

Injury, X	**Frequency**
1	169
2	156
3	104
4	48
5	14
6	6
7	3

The X represents the number of injuries a particular person has with the frequency of individuals with that number of injuries. For example, X=4 with 48 means there are 48 individuals that experience 4 injuries. The total frequency in Table 3.1 is n = 500. There were multiple tables generated (MC_num = 1,000). There were additional values generated but they were the X=0 values and they were place in separate tables. This was done by generating data until the total size of n=500 was reached but X=0 discarded. This was done so that all of the tables (MC_num) could be the same sample size and avoid the issue of uneven sample sizes (each data table will, however, have a separate value for the frequency of zero's present, X=0). The data is generated under a Poisson distribution with known lambda. Each estimator was compared to see how accurately it estimates the true lambda. Three of the estimators: MLE, ordinary Zelterman (binominal distribution), and Multinomial Zelterman (multinomial distribution extension) were calculating using aggregated tables where Table 3.1 is just one example.

All three estimators were obtained within PROC NLMIXED using the method of maximum likelihood. The principal SAS procedure used in this analysis was PROC NLMIXED (SAS version 9.2). The NLMIXED procedure fits nonlinear mixed models—that is, models in which both fixed and random effects enter nonlinearly. These models have a wide variety of applications, two of the most common being pharmacokinetics and overdispersed binomial data. The latter is more closely analogous to the computational problem of identifying an unknown population size, N. PROC NLMIXED enables one to specify a conditional distribution for your data (given the random effects) having either a standard form (normal, binomial, Poisson) or a general distribution that you code using SAS programming statements. In this simulation model the distribution is assumed to be a truncated Poisson.

PROC NLMIXED fits nonlinear mixed models by maximizing an approximation to the likelihood integrated over the random effects. Different integral approximations are available, the principal ones being adaptive Gaussian quadrature and a first-order Taylor series approximation. A variety of alternative optimization techniques are available to carry out the maximization; the default is a dual quasi-Newton algorithm.

Successful convergence of the optimization problem results in parameter estimates along with their approximate standard errors based on the second derivative matrix of the likelihood function. PROC NLMIXED enables you to use the estimated model to construct predictions of arbitrary functions by using empirical Bayes estimates of the random effects.

1) Using PROC NLMIXED to calculate the maximum likelihood function (MLE) for a truncated Poisson:

```
PROC NLMIXED DATA=IDx;
X = COUNT;
PARMS lambda=2;
LL   = -lambda+X*log(lambda)-lgamma(X+1) -log(1-exp(-lambda));
MODEL X ~ general(ll);
RUN;
ODS SELECT ALL;
```

The function to be optimized is the truncated Poisson log likelihood function,

LL=-lambda+X*log(lambda)-lgamma(X+1) -log(1-exp(-lambda))

The parameter to be estimated is lambda and it is initially set to 2 (PARMS lambda = 2). The parameter will be optimized vis-à-vis the inputted COUNT variable which is transformed to X when invoking PROC NLMIXED in this section of the code. The purpose of calculating lambda is to estimate the probability of getting zero,

$$\Pr(X=0) = 1\text{-}\exp^{(-\text{lambda})}$$

which will subsequently be used to estimate N=0, the unknown population size.

2) Zelterman (original using binomial): The original Zelterman estimator can be generalized with the binomial distribution with the following:

```
PROC NLMIXED DATA=IDx;
X = COUNT;
PARMS lambda=5;
p1=exp(-lambda)*lambda/FACT(1);
p2=exp(-lambda)*lambda**2/FACT(2);
DEN=p1+p2;
ll=log( (p1/(p1+p2))**(X=1)) + log( (p2/(p1+p2))**(X=2) );
MODEL X ~ general(ll);
RUN;
```

3) Modified Zelterman (multinomial extension – includes an array):

```
PROC NLMIXED DATA=IDx;
X = COUNT;
PARMS lambda=2;
lik=exp(-lambda)*(lambda**X)/FACT(X)/den;
ll=log(lik);
MODEL X ~ general(ll);
RUN;
```

The maximum likelihood estimator (MLE) was calculated using PROC NLMIXED using the log likelihood expression:

LL = -lambda+X*log(lambda)-lgamma(X+1) -log(1-exp(-lambda))

With the parameter lambda initially set to 2. The variable lambda is maximized in regard to X from the data.

Comparison between just the MLE and Zelterman (Ordinary) Estimator

A comparison between MLE and Zelterman (method 1 and 2, respectively) may potentially shed some light on which is the more robust statistic and under what conditions. The multinomial extension of the Zelterman (method 3) was excluded here because it yielded values much less accurate than the first two methods, and therefore introduces unnecessary tables. We are further exploring how we can improve this estimator.

To do the comparison a simulation model was used. The data was generated under the assumption of a truncated Poisson distribution with no unobserved heterogeneity. The hypothesis is that the Zelterman estimator will performed better, or at best, reasonable for λ= 1 or 2. The MLE is predicted to perform better for greater λ. The λ: = {1, 1.5, 2, 3, 4, 7, 10}. The results showed that MLE performed much better than the Zelterman. This is not surprising since the initial data was homogeneous, i.e. without unobserved heterogeneity and the Zelterman estimator was designed to deal with the presence of overdispersion. During the presentation I will discuss what happens when covariates are introduced.

Table 2: Comparison of maximum likelihood (MLE) and Zelterman

Lamda	zelt_exists	lambda_mle	lambda_tilde	n0_mle_bias	n0_tilde_bias	n0_mle_mse	n0_tilde_mse
1	1	0.993	1.008	2.221	4.771	241.642	653.609
1.5	1	1.494	1.531	0.852	2.721	65.336	250.186
2	1	1.992	2.091	0.390	1.701	28.043	131.330
3	1	2.996	3.229	0.068	1.805	6.425	60.630
4	1	3.993	4.729	0.028	2.490	1.998	49.426
7	0.426	6.990	4.150	**-0.010**	12.121	0.105	609.382
10	0.008	9.985	2.500	0.002	12.205	0.003	184.604

The column *Lambda* in table 2 is the true mean used in the truncated Poisson model. *Zelt_exists* states the percentages of Zeltermans that actually exist. For a Zelterman estimator to exists both the frequency of 1 and 2 must exist. *Lambda_mle* and *lambda_tilde* are the MLE estimate and Zelterman estimates, respectively of the mean. The last four statistics are the bias and mean square error (MSE) of the estimators. As lambda, λ increases the difference in error measurements (MSE and bias) become more obvious between MLE and Zelterman; also note past λ = 4 there are more data sets which we are unable to calculate a Zelterman estimator, e.g. at λ = 10 there are only 8 Zelterman estimators available. Consequently, the resulting Zelterman estimates are grossly inaccurate. The problem with the Zelterman estimator is that at high values of λ, the estimator becomes ineffective.

CONCLUSION

The purpose of this paper is to discuss various methods of calculating population size using PROC NLMIXED. PROC NLMIXED is used to determine the mean of the truncated Poisson distribution and through various equations the population size can be estimated. Three estimators of λ were introduced here; two additional variations of the Zelterman estimator will be introduced later. It is important to note here that the data is simulated with a known distribution and parameters. More importantly, no covariates are assumed. Nonetheless, the usefulness of all these techniques will be demonstrated in their application to real world problems, where covariates are often present.

REFERENCES

- The NLMIXED Procedure. SAS Institute Inc. 2011. SAS/STAT® 9.3 User's Guide. Cary, NC: SAS Institute Inc.

- Bohning, Dankmar, and Van Der Heijden, Peter G. M. "A COVARIATE ADJUSTMENTS FOR ZERO-TRUNCATED APPROACHES TO ESTIMATING THE SIZE OF HIDDEN AND ELUSIVE POPULATIONS." The Annals of Applied Statistics 3.2 (2009): 595-610. Print.

- Zelterman, Daniel. "ROBUST ESTIMATION IN TRUNCATED DISCRETE DISTRIBUTIONS WITH APPLICATION TO CAPTURE-RECAPTURE EXPERIMENTS." Journal of Statistical Planning and Inference.18 (1988): 225-37. Print.

ACKNOWLEDGMENTS

I would like to thank Dr. Nagaraj K. Neerchal in the Mathematics and Statistics Department at the University of Maryland, Baltimore County (UMBC) for providing much needed criticism, an abundance of analytical insight, many hours of SAS programming, and more importantly for his unwavering support of this project. All of the formulas, derivations, and SAS codes present here in this paper would not have been possible without his assistance.

I would also like to thank my other supervisor, Dr. Brian Maguire of the Department of Emergency Health Services at the University of Maryland, Baltimore County (UMBC) for providing me with the data and leading me to this interesting statistical problem.

CONTACT INFORMATION

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

Name:	Brandon Earl Fleming
Enterprise:	University of Maryland, School of Pharmacy
Address:	
City, State ZIP:	Baltimore, Maryland 21201
Work Phone:	(443) 740-5565
Fax:	(410) 435-9738
E-mail:	brandon.fleming@umaryland.edu