

**Constructing two-sided simultaneous confidence intervals  
for multinomial proportions for small counts in a large number of cells**

Warren L. May and William D. Johnson

Department of Preventive Medicine  
University of Mississippi Medical Center  
2500 North State Street  
Jackson, Mississippi 39216-4505

## **Constructing two-sided simultaneous confidence intervals for multinomial proportions for small counts in a large number of cells**

### **Abstract**

Confidence intervals for multinomial proportions are often constructed using large-sample methods that rely on expected cell counts of 5 or greater. In situations that give rise to a large number of categories, the cell counts may not be of adequate size to ensure the appropriate overall coverage probability and alternative methods of construction have been proposed. Sison and Glaz (1995) developed a method of constructing two-sided confidence intervals for multinomial proportions that is based on the doubly truncated Poisson distribution and their method performs well when the cell counts are fairly equally dispersed over a large number of categories. In fact, the Sison and Glaz (1995) intervals appear to outperform other methods of simultaneous construction in terms of coverage probabilities and interval length in these situations. To make the method available to researchers, we have developed a SAS macro to construct the intervals proposed by Sison and Glaz (1995).

*Keywords: Confidence limits, truncated Poisson distribution, multinomial distribution*

## 1. Introduction

Confidence intervals for multinomial proportions can be constructed by relying on the usual large-sample methods when the cell counts are adequate ( $\geq 5$  per cell) so that coverage probabilities are at or near the nominal confidence level. The authors [1] presented a SAS<sup>®</sup> macro for simultaneous confidence interval construction for multinomial proportions using several of the large-sample methods. The original macro was written in the interactive matrix language IML [2] and is easy to implement using a standard SAS macro call.

In some instances, however, the researcher may find that the cell counts in several cells are too small to rely on the large-sample methods. If the cell counts are fairly evenly dispersed over a large number of categories, methods proposed by Sison and Glaz [3] appear to maintain an adequate coverage probability where other methods perform poorly as evidenced by those authors' simulation comparisons. In the present paper, we develop a macro that constructs simultaneous intervals based on the methods proposed by Sison and Glaz [3]. We describe various methods of construction in Section 2 to differentiate the usual large-sample approaches from those of Sison and Glaz [3]. We discuss the SAS macro in Section 3 and illustrate the method with an example in Section 4. We discuss caveats in using the macro in Section 5. We give a slightly more detailed discussion of the computing aspects of the truncated Poisson probabilities in Appendix A. The macro is listed in Appendix B, an example of the calling routine in Appendix C. Sample output is listed in Appendix D.

## 2. Methods of constructing two-sided confidence

### intervals for multinomial proportions

Let  $\mathbf{n} = (n_1, \dots, n_k)^T$  represent the vector of observed cell counts from a  $k \times 1$  classification table where  $n = n_1 + \dots + n_k$  is the total sample size. Thus,  $n_i$  ( $i = 1, \dots, k$ ) is the number of observations and  $p_i = n_i / n$  is the proportion observed in the  $i^{\text{th}}$  cell of the  $k \times 1$  table ( $i = 1, \dots, k$ ). Assuming the total sample size  $n$  is fixed, the vector  $\mathbf{n}$  is an observation from a multinomial distribution with parameters  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)^T$  where  $\pi_i$  is the population proportion for the  $i^{\text{th}}$  cell. The vector  $\mathbf{p} = (p_1, \dots, p_k)^T$  is the maximum likelihood estimator of  $\boldsymbol{\pi}$  and is unbiased. The variance of  $p_i$  is  $\pi_i(1 - \pi_i)/n$  and is usually estimated by  $p_i(1 - p_i)/n$ . The covariance matrix is  $\boldsymbol{\Sigma} = (\boldsymbol{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}^T)/n$  with variances along the diagonal and is estimated by  $\mathbf{S} = (\mathbf{p} - \mathbf{p}\mathbf{p}^T)/n$ .  $\mathbf{S}$  converges to  $\boldsymbol{\Sigma}$  as  $n$  grows large.

Recently, Agresti and Coull [4] published an informative article elucidating the coverage properties of confidence interval construction for a binomial parameter. They compared score intervals attributed to Wilson [5] with Wald [6] intervals and exact intervals based on binomial probabilities. The Wald intervals are found by solving

$$n(p_i - \pi_i)^2 \leq \chi^2 p_i(1 - p_i) \quad (i = 1, 2)$$

for  $\pi_i$  where  $\chi^2 = \chi^2(\alpha, 1)$  is the upper  $100(1 - \alpha)$  percentage point of the chi-square distribution giving endpoints  $\pm \sqrt{\chi^2 p_i(1 - p_i)/n}$ . The Wald intervals are sometimes given in introductory texts and reported by some computer packages, but they do not perform well with respect to coverage probability or interval length. The Wilson intervals are found by solving

$$n(p_i - \pi_i)^2 \leq \chi^2 \pi_i(1 - \pi_i) \quad (i = 1, 2)$$

for  $\pi_i$  and do not require estimating the variance from the sample. Both types are related to Pearson's [7] chi-square statistic and can be developed using multivariate normal theory [1, 8].

Quesenberry and Hurst [9] adapted the methods of Wilson [5] for simultaneous construction of multinomial parameters using  $\chi^2 = \chi^2(\alpha, k-1)$  and Goodman [10], invoking a Bonferonni argument, suggested  $\chi^2 = \chi^2(\alpha/k, 1)$ . Thus, the Wilson intervals are a special case of the Quesenberry and Hurst [9] and Goodman [10] intervals. Agresti and Coull [4] suggested using the Wilson intervals for the binomial proportion. If  $k > 2$ , the Goodman [10] intervals are preferred.

The authors [1, 11] discussed various methods of constructing simultaneous intervals for multinomial proportions as adaptations of the Quesenberry and Hurst [9] or the Wald intervals. Adaptations include variance correction factors, continuity corrections and Bonferroni adjustments. The method proposed by Sison and Glaz [3], however, differs from others and provides an alternative that performs well in instances that are not amenable to normal large-sample theory.

Following a presentation by Levin [12], Sison and Glaz [3] used the relationship between the Poisson, truncated Poisson and multinomial distributions to construct simultaneous confidence intervals for the multinomial proportions. They presented two procedures, but recommended the procedure that was least intensive computationally. We outline that procedure here and give more detail concerning computing the truncated Poisson probabilities in Appendix A.

Assume  $Z_i$  ( $i = 1, \dots, k$ ) are independent Poisson random variables such that

$$\lambda_i = E(Z_i) = n\pi_i$$

represents the mean and, thus, the variance of the distribution. If  $A_i = \{Z_i \mid Z_i \in [b_i, a_i]\}$  is

the set of events such that  $b_i \leq Z_i \leq a_i$ , then by Bayes' theorem

$$\begin{aligned} P\left(A_1, \dots, A_k \mid \sum_{i=1}^k Z_i = n\right) &= \frac{P(A_1, \dots, A_k)}{P\left(\sum_{i=1}^k Z_i = n\right)} P\left(\sum_{i=1}^k Z_i = n \mid A_1, \dots, A_k\right) \\ &= \frac{\prod_{i=1}^k P(b_i \leq Z_i \leq a_i)}{\frac{n^n e^{-n}}{n!}} P(W = n) \end{aligned}$$

where  $W$  is the sum of  $k$  independent observations from truncated Poisson distributions with endpoints  $[b_i, a_i]$ . Assuming  $N_i$  are random variables representing the cell counts from a multinomial distribution, it follows from the work of Levin [12] that

$$P(b_i \leq N_i \leq a_i ; i = 1, \dots, k) \approx \frac{n!}{n^n e^{-n}} \left\{ \prod_{i=1}^k P(b_i \leq Z_i \leq a_i) \right\} P(W = n)$$

where  $Z_i, i = 1, \dots, k$ , are independent Poisson random variables with  $\lambda_i = n\pi_i$  and  $W$  is the sum of  $k$  independent random variables from truncated Poisson distributions on the interval  $[b_i, a_i]$  also with mean  $\lambda_i = n\pi_i$ .

Because  $\lambda_i = n\pi_i$  is usually unknown, we use  $\hat{\lambda}_i = n\hat{\pi}_i$  to estimate the moments and rely on an Edgeworth expansion to estimate the probability mass function for the approximate truncated Poisson distribution as outlined in Appendix A. We search for a positive integer  $c$  by solving

$$\rho(c) = P(p_i - c/n \leq \pi_i \leq p_i + c/n ; i = 1, \dots, k) \approx 1 - \alpha.$$

That is, we search for a single number  $c$  that will give us approximately the correct coverage probability for the entire set of events  $A_1, \dots, A_k$ . We use the same value of  $c$  for all proportions so that the method works best when we have nearly equal proportions across all  $k$  cells. We give equal weight to each proportion and, therefore, the interval lengths are the equal. Because the interval lengths are equal, the Sison and Glaz [3] method relies on fairly even dispersion of the cell counts across all categories. The Sison and Glaz [3] method is quite different from the Wilson or Quesenberry and Hurst methods that provide potentially different interval lengths based on the variance associated with a particular cell count.

Sison and Glaz [3] suggested finding an integer such that  $\rho(c) \leq 1 - \alpha \leq \rho(c + 1)$  and using an interpolation adjustment to form the simultaneous confidence intervals

$$(p_i - c/n \leq \pi_i \leq p_i + c/n + 2\delta/n)$$

where  $\delta = [(1 - \alpha) - \rho(c)] / [\rho(c + 1) - \rho(c)]$ . They note that the adjustment is necessary to correct for skewness. An alternative approach that is easy to implement with the SAS macro we provided is to compute slightly wider intervals as

$$(p_i - c/n - 1/n \leq \pi_i \leq p_i + c/n + 1/n)$$

that is equivalent to using  $c + 1$  instead of  $c$  where  $\rho(c) \leq 1 - \alpha \leq \rho(c + 1)$ . If  $n$  is relatively large, these intervals will be only slightly more conservative than those proposed by Sison and Glaz [3] but will ensure that the coverage probability is at least as large as the specified level, assuming  $np$  is a good approximation of  $n\pi$ . The intervals we propose, however, do not account for skewness as do those of Sison and Glaz [3].

### 3. A SAS macro for two-sided multinomial confidence intervals

We have written a SAS macro using PROC IML that takes multinomial cell counts as input and constructs simultaneous confidence intervals for multinomial parameters as output. The algorithm, based on the methods of Sison and Glaz [3], estimates the moments of truncated Poisson distributions to employ an Edgeworth expansion for estimating the coverage probabilities for the observed cell counts. A simple search yields a set of intervals with user-specified coverage probability for the joint confidence region. The macro is given in Appendix B and the calling routine is given in Appendix C. The macro definition provides the user options for specifying the *dataset* and the desired coverage probability  $1 - \alpha$ . For small proportions, it may be necessary to change the output to include more decimal places using the *decimal* option.

The data are entered as a  $k \times 1$  vector using a SAS DATA step. The calling routine follows the data step and is called with the statement `%sison(data = dataset, alpha = alpha)`. The program computes the confidence intervals for the proportions based on the truncated Poisson method and outputs the upper and lower bounds. As an example of the calling routine, the code in Appendix C computes one of the examples given by Sison and Glaz [3]. Sample output from the macro is given in Appendix D.

The "moments" function takes input values of  $\hat{\lambda}_i = np_i$  for each of the  $k$  observed cell counts and uses a common integer  $c$  in constructing the intervals. If  $b_i = \lambda_i - c \leq 0$ , we assume  $b = 0$  and use  $P(Z \leq a)$  for the denominator; otherwise, we use  $P(Z \leq a) - P(Z$

$\leq b - 1$ ). The Poisson probabilities are computed using the available SAS function  $P(Z \leq z) = \text{poisson}(\text{lambda}, z)$ . The "moments" function then computes the factorial moments storing them in the vector, mu, setting to 0 those that are undefined. The central moments are computed from the factorial moments and stored in the "mom" vector. The "truncpoi" function first calls the "moments" function for each of the k observations and stores the central moments. The various components of the Edgeworth expansion are computed and the coverage probability for the particular choice of c is computed in the "truncpoi" function.

The main routine compares the coverage probability for c with the previously computed coverage probability for c - 1. The algorithm allows c to range from 1 to n. When  $\rho(c) \leq 1 - \alpha \leq \rho(c + 1)$ , the algorithm stops. The correction factor  $\delta$  is calculated and the endpoints are presented as part of the output.

#### **4. Example**

The example used to illustrate the calling routine was chosen to complement data from Sison and Glaz [3] where cell counts represent the number of "personal crimes committed in the city of New Orleans on each of the seven days in a randomly selected week in 1984". The cell counts for each of the 7 days are given in the calling routine in Appendix C along with the output for this example in Appendix D. Using a Pentium 200 MHz processor running under Windows 95, the macro used approximately 0.4 seconds for this example. Sison and Glaz [3] through simulation estimated the average coverage probability to be 0.939 for their method and approximately 0.935 for Goodman's [10] approach and there is no clear choice based on their simulations. In this case, the

proportions are nearly equal and the sample size is adequate for the normal theory methods.

The output gives the estimated coverage probabilities using  $c$  and  $c + 1$  as 0.953 and 0.932 for this particular data set. These are not, however, the same as the expected coverage probabilities that Sison and Glaz [3] simulated and should not be interpreted as such. The coverage probability for this particular sample using the Sison and Glaz [3] method is not calculated since the Poisson function relies on discrete counts that are between  $c$  and  $c + 1$ , but the estimated coverage probability for this particular observation is somewhere between  $P(c)$  and  $P(c + 1)$ , provided the observed proportions are near the true population proportions. The volume is also based on this particular sample and is included as a comparison between the two methods. Thus, the reported volume is not to be interpreted as an expected volume.

The observed proportions and confidence intervals using both the recommended correction  $\delta$  and those based on  $c + 1$  are also listed. Necessarily, the intervals using  $\delta$  are slightly skewed with center at  $p + \delta/n$ . The intervals based on  $c + 1$  use endpoints  $\pm (c/n + 1/n)$ . The macro also outputs the mid-points of the Sison and Glaz [3] intervals that, similar to the weighted estimators discussed by Agresti and Coull [4], may also be used to estimate the proportions.

For comparison, Table 1 gives other intervals discussed by May and Johnson [1].

-----Table 1 goes about here-----

These include intervals proposed by Quesenberry and Hurst, Goodman, and Fitzpatrick and Scott. We point out that, for this example, the volume for the Sison and Glaz intervals is smallest.

## 5. Conclusion

As discussed in previous articles [1, 11], the Sison and Glaz [3] intervals are not always appropriate. If the number of cells is small, say no more than 10, the Goodman [10] intervals perform well. The Wald intervals should generally be avoided unless the number of observations per cell is very large. Note that a large overall sample size does not guarantee coverage probability for any of the methods.

The Sison and Glaz [1, 3] intervals perform well when the number of categories is large and the observations are distributed evenly across the  $k$  cells. In these situations, the overall coverage volume is consistently smaller than other methods and the coverage probability is near the desired level. In the development of the Sison and Glaz [3] methods, however, recall that we added the same constant to all cells. Thus, we estimate each of the proportions with the same precision analogous to assuming equal variance for a normal-theory model. If some cell counts dominate, the Sison and Glaz [3] intervals do not perform well. On the other hand, if the cell counts are roughly equal, the method outperforms normal-theory methods, especially when the number of cells is large and the number in each cell is small.

In addition to the Sison and Glaz method, we proposed a slight modification. While the Quesenberry and Hurst, Goodman, and Sison and Glaz intervals are asymmetric about the observed proportions, the proposed intervals are symmetric. The volume of the Sison and Glaz intervals is  $[2(c+\delta)/n]^k$  where  $\delta = [(1 - \alpha) - \rho(c)] / [\rho(c + 1) - \rho(c)]$  and for the proposed intervals it is  $[2(c+1)/n]^k$ . Because  $c$  is chosen so that

$\rho(c) \leq 1 - \alpha \leq \rho(c + 1)$ , we have  $0 \leq \delta \leq 1$  and the volume of the Sison and Glaz intervals is smaller than the proposed intervals.

## Appendix A

It is convenient for programming purposes to view the term  $\frac{n^n e^{-n}}{n!}$  as a Poisson probability with location parameter  $\lambda = n$  and observed count  $n$ . Because of the assumed independence of the  $A_i$  events, the joint probability  $\prod_{i=1}^k P(b_i \leq Z_i \leq a_i)$  is the product of Poisson probabilities with parameters  $\lambda_i = n\pi_i$  summed over the intervals  $[b_i, a_i]$  for all  $i = 1, \dots, k$ .

Computing  $P(W = n)$  is somewhat more difficult. Sison and Glaz [3] used Levin's [12] suggested Edgeworth expansion

$$f\left(\frac{n - \sum_{i=1}^k \mu_i}{\sqrt{\sum_{i=1}^k \sigma_i^2}} \left( \frac{1}{\sqrt{\sum_{i=1}^k \sigma_i^2}} \right)\right)$$

where  $\mu_i$  and  $\sigma_i^2$  are the 1<sup>st</sup> and 2<sup>nd</sup> central moments from the  $i^{\text{th}}$  truncated Poisson distribution with location parameters  $\lambda_i = n\pi_i$ . Note that  $W = n$  is the observed sum from  $k$  truncated Poisson distributions with  $E(W) = \sum_{i=1}^k \mu_i$  and variance  $\sum_{i=1}^k \sigma_i^2$ , assuming independence. Thus,

$$X = \frac{W - \sum_{i=1}^k \mu_i}{\sqrt{\sum_{i=1}^k \sigma_i^2}}$$

is a normalized sum of truncated Poisson random variables whose derivative with respect to  $W$  is  $1/\sqrt{\sum_{i=1}^k \sigma_i^2}$ . We can approximate the probability mass function of  $X$  by "scaling" the normal distribution through an Edgeworth expansion.

Patel and Read [13] discussed the Edgeworth expansion in general and also gave references and details using

$$f(x) = \phi(x) \left[ 1 + \gamma_1 \left( \frac{H_3(x)}{3!} \right) + \gamma_2 \left( \frac{H_4(x)}{4!} \right) + 10\gamma_1^2 \left( \frac{H_6(x)}{6!} \right) + \dots \right]$$

where

$$\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

is the probability density function of the standard normal and

$$H_0(x) = 1$$

$$H_3(x) = x^3 - 3x$$

$$H_4(x) = x^4 - 6x^2 + 3$$

$$H_6(x) = x^6 - 15x^4 + 45x^2 - 15$$

are Tchebyshev-Hermite polynomials. To compute  $\gamma_1$  and  $\gamma_2$ , let  $\mu_i$ ,  $\sigma_i^2$ ,  $\mu_{3,i}$  and  $\mu_{4,i}$  represent the first 4 central moments of the truncated Poisson distribution. Let

$$\gamma_1 = \frac{\sum_{i=1}^k \mu_{3,i}}{\left[ \sum_{i=1}^k \sigma_i^2 \right]^{3/2}}$$

and

$$\gamma_2 = \frac{\sum_{i=1}^k \mu_{4,i} - 3\sigma_i^4}{\left[ \sum_{i=1}^k \sigma_i^2 \right]^2}.$$

Levin [12] and Patel and Read [13] labeled  $\gamma_1$  the "coefficient of skewness" and  $\gamma_2$  the "coefficient of excess". We need to compute the first 4 central moments of the truncated Poisson distribution. Hjorth [14] discussed the relationship between the Edgeworth expansion and the moment generating function that the reader may find useful.

Mood, Graybill and Boes [15] noted that for discrete distributions the factorial moments may be easier to calculate than the central moments. The central moments can be calculated from the factorial moments using standard formulae. For  $r > 0$ , the  $r^{\text{th}}$  factorial moment of the truncated Poisson distribution on  $[b, a]$  with location parameter  $\lambda$  is

$$\mu_{(r)} = \lambda^r \left[ 1 - \frac{\sum_{v=a-r+1}^a \frac{\lambda^v e^{-\lambda}}{v!} - \sum_{v=b-r}^{b-1} \frac{\lambda^v e^{-\lambda}}{v!}}{\sum_{v=b}^a \frac{\lambda^v e^{-\lambda}}{v!}} \right].$$

For ease of notation, we give a general formula but each of the  $i = 1, \dots, k$  observations will have a separate set of moments. Note that the expression  $\frac{\lambda^v e^{-\lambda}}{v!}$  is a Poisson probability and is relatively easy to compute using standard Poisson probability functions.

The first factorial moment of the truncated Poisson is

$$\begin{aligned} \mu_{(1)} &= \lambda \left[ 1 - \frac{\frac{\lambda^a e^{-\lambda}}{a!} - \frac{\lambda^{b-1} e^{-\lambda}}{(b-1)!}}{\sum_{v=b}^a \frac{\lambda^v e^{-\lambda}}{v!}} \right] \\ &= \lambda \left[ 1 - \frac{P(Z = a) - P(Z = b-1)}{P(b \leq Z \leq a)} \right] \\ &= \lambda \left[ 1 - \frac{[P(Z \leq a) - P(Z \leq a-1)] - [P(Z \leq b-1) - P(Z \leq b-2)]}{P(b \leq Z \leq a)} \right] \end{aligned}$$

where  $P(Z \leq z)$  is a cumulative Poisson probability. It is convenient to formulate the factorial moments as

$$\mu_{(r)} = \lambda^r \left[ 1 - \frac{[P(Z \leq a) - P(Z \leq a-r)] - [P(Z \leq b-1) - P(Z \leq b-1-r)]}{P(b \leq Z \leq a)} \right].$$

Note that the  $\mu_{(r)}$  are functions of cumulative Poisson probabilities. The moments are undefined for  $P(Z = z)$  when  $z < 0$  and it is customary to assume  $\mu_{(r)} = 0$  in such cases.

The central moments are obtained from the factorial moments as follows:

$$\mu = \mu_{(1)}$$

$$\mu_2 = \mu_{(2)} + (\mu - \mu^2)$$

$$\mu_3 = \mu_{(3)} + \mu_{(2)}(3 - 3\mu) + (\mu - 3\mu^2 + 2\mu^3)$$

$$\mu_4 = \mu_{(4)} + \mu_{(3)}(6 - 4\mu) + \mu_{(2)}(7 - 12\mu + 6\mu^2) + (\mu - 4\mu^2 + 6\mu^3 - 3\mu^4).$$

Given the means of the Poisson distribution, we can find the first 4 factorial moments from which we compute the first 4 central moments of the truncated Poisson distribution. The central moments are used in the Edgeworth expansion where we must find  $P(W = n)$ .

## Appendix B

```

%macro sison(data=      _last_,
                   al pha=0.05,
                   deci mal=12.4);

proc iml;
  reset nocenter;
  reset noname;
  use &data;
  setin &data;
  read all into x;
  data={&data};
  al pha={&al pha};
  deci mal={&deci mal};
  n=sum(x);
  k=nrow(x);
  p=x/n;
  probn=1/(poi sson(n, n)-poi sson(n, n-1));
  sqrt2pi=sqrt(2)*gamma(0.5);
  print 'Data set used:      ' data;
  print 'Al pha =          ' al pha;
  print 'Number of cells=' k;
  print 'N =              ' n;
  print 'Observed cell counts ', x;

  start moments;
  a=l ambda+c;
  b=l ambda-c;
  if b<0 then b=0;
  poi sl ama=poi sson(l ambda, a);
  poi sl amb=poi sson(l ambda, b-1);
  if b>0 then den=poi sl ama-poi sl amb;
  if b=0 then den=poi sl ama;
  mu=j(4, 1, 0); mom=j(5, 1, 0);
  do r = 1 to 4;
    poi sA=0;
    poi sB=0;
    if a-r>=0 then poi sA=poi sl ama-poi sson(l ambda, a-r);
    if a-r< 0 then poi sA=poi sl ama;
    if b-r-1>=0 then poi sB=poi sl amb-poi sson(l ambda, b-r-1);
    if b-r-1<0 && b-1>=0 then poi sB=poi sl amb;
    if b-r-1<0 && b-1<0 then poi sB=0;
    mu[r]=(l ambda**r)*(1-(poi sA-poi sB)/den);
  end;
  mu1=mu[1]; mu1_2=mu1**2; mu1_3=mu1**3; mu1_4=mu1**4;
  mu2=mu[2]; mu3=mu[3]; mu4=mu[4];

  mom[1]=mu1;
  mom[2]=mu2+mu1-mu1_2;
  mom[3]=mu3+mu2*(3-3*mu1)+(mu1-3*mu1_2+2*mu1_3);
  mom[4]=mu4+mu3*(6-4*mu1)+mu2*(7-12*mu1+6*mu1_2)
        +mu1-4*mu1_2+6*mu1_3-3*mu1_4;
  mom[5]=den;
  fini sh;

```

```

start truncpoi;
m=j(k, 5, 0);
do i=1 to k;
  lambda=x[i];
  run moments;
  do j=1 to 5;
    m[i, j]=mom[j];
  end;
end;
s1=m[+, 1];
s2=m[+, 2];
s3=m[+, 3];
do i=1 to k;
  m[i, 4]=m[i, 4]-3*m[i, 2]**2;
end;
s4=m[+, 4];
z=(n-s1)/sqrt(s2);
g1=s3/(s2**(3/2));
g2=s4/(s2**2);
z_2=z**2;
z_3=z**3;
z_4=z**4;
z_6=z**6;
pol y=1+g1*(z_3-3*z)/6+g2*(z_4-6*z_2+3)/24
      +g1**2*(z_6-15*z_4+45*z_2-15)/72;
f=pol y*exp(-z_2/2)/sqrt(2*pi);
probx=1;
do i=1 to k;
  probx=probx*m[i, 5];
end;
p=probn*probx*f/sqrt(s2);
fini sh;

start;
pold=0;
do c=1 to n;
  run truncpoi;
  if p > 1-alpha && pold < 1-alpha then goto done;
  pold=p;
end;

done;
del ta=(1-alpha-pold)/(p-pold);
out=j(k, 5, 0);
num=j(k, 1, 0);
c=c-1;
vol 1=1;
vol 2=1;
do i=1 to k;
  num[i, 1]=i;
  obsp=x[i]/n;
  cn=c/n;
  onen=1/n;
  out[i, 1]=obsp;
  out[i, 2]=obsp-cn;
  out[i, 3]=obsp+cn+2*del ta/n;
  if out[i, 2]<0 then out[i, 2]=0;
  if out[i, 3]>1 then out[i, 3]=1;
  out[i, 4]=obsp-cn-onen;
  out[i, 5]=obsp+cn+onen;
  if out[i, 2]<0 then out[i, 2]=0;
  if out[i, 3]>1 then out[i, 3]=1;
  vol 1=vol 1*(out[i, 3]-out[i, 2]);
  vol 2=vol 2*(out[i, 5]-out[i, 4]);
end;
c1={' PROPORTION', ' LOWER(SG)', ' UPPER(SG)', ' LOWER(C+1)', ' UPPER(C+1) '};
cov=100*(1-alpha);
sg=(x+del ta)/n;
c2={' SG-midpoint' };
print '-----';
print cov% SIMULTANEOUS CONFIDENCE INTERVALS';
print '      BASED ON THE METHODS OF SISON AND GLAZ';
print '-----';
print ' C = ' c;
print ' P(c+1) = ' p(|format=5.4|);
print ' P(c) = ' pold(|format=5.4|);
print ' del ta = ' del ta(|format=5.4|);
print ' Volume(SG) = ' vol 1;
print ' Volume(C+1) = ' vol 2;
print num(|format=3.0|) out(|col name=c1 format=&decimal |);

```

```
    print num(|format=3.0|) sg(|col name=c2 format=&decimal |);  
finish;  
run;  
quit;  
  
%mend;  
  
quit;  
run;
```

## Appendix C

```
data one;  
  input x @@;  
cards;  
56 72 73 59 62 87 58  
  
run;  
%si son(data=one, al pha=0.05);  
run;  
qui t;
```

Appendix D

Data set used: ONE  
 Alpha = 0.05  
 Number of cells= 7  
 N = 467

Observed cell counts  
 56  
 72  
 73  
 59  
 62  
 87  
 58

-----  
 95 % SIMULTANEOUS CONFIDENCE INTERVALS  
 BASED ON THE METHODS OF SISON AND GLAZ  
 -----

C = 19  
 P(c+1) = .9525  
 P(c) = .9320  
 del ta = .8771  
 Volume(SG) = 3.2393E-8  
 Volume(C+1) = 3.3822E-8

	PROPORTION	LOWER(SG)	UPPER(SG)	LOWER(C+1)	UPPER(C+1)
1	0.1199	0.0792	0.1644	0.0771	0.1627
2	0.1542	0.1135	0.1986	0.1113	0.1970
3	0.1563	0.1156	0.2008	0.1135	0.1991
4	0.1263	0.0857	0.1708	0.0835	0.1692
5	0.1328	0.0921	0.1772	0.0899	0.1756
6	0.1863	0.1456	0.2307	0.1435	0.2291
7	0.1242	0.0835	0.1686	0.0814	0.1670

SG-midpoint

1	0.1218
2	0.1561
3	0.1582
4	0.1282
5	0.1346
6	0.1882
7	0.1261

## References

- [1] W.L. May and W.D. Johnson, A SAS<sup>®</sup> macro for constructing simultaneous confidence intervals for multinomial proportions, *Computer Methods and Programs in Biomedicine*, 53 (1997) 153-162.
- [2] SAS Institute, Inc., SAS/IML<sup>®</sup> User's Guide for Personal Computers, Version 6 edn. (SAS Institute, Inc., Cary, NC, 1985).
- [3] C.P. Sison and J. Glaz, Simultaneous confidence intervals and sample size determination for multinomial proportions, *J. Am. Stat. Assoc.* 90 (1995) 366-369.
- [4] A. Agresti and B.A. Coull, Approximate is better than "exact" for interval estimation of binomial proportions, *The American Statistician* 52 (1998) 119-126.
- [5] E.B. Wilson, Probable inference, the law of succession and statistical inference, *J. Am. Stat. Assoc.* 22 (1927) 209-212.
- [6] A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Trans. Am. Math. Soc.* 54 (1943) 426-482.
- [7] K. Pearson, On the criteria that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philos. Mag.* 50 (1900) 157-175.
- [8] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, 2<sup>nd</sup> edn. (Prentice Hall, New Jersey, 1988).
- [9] C.P. Quesenberry and D.C. Hurst, Large sample simultaneous confidence intervals for multinomial proportions, *Technometrics* 6 (1964) 191-195.
- [10] L.A. Goodman, On simultaneous confidence intervals for multinomial proportions, *Technometrics* 7 (1965) 247-254.
- [11] W.L. May and W.D. Johnson, Properties of simultaneous confidence intervals for multinomial proportions, *Commun. Statist.-Simula.* 26 (1997) 495-518.
- [12] B. Levin, A representation for multinomial cumulative distribution functions, *The Annals of Statistics* 9 (1981) 1123-1126.
- [13] J.K. Patel and C.B. Read, *Handbook of the Normal Distribution*, 2<sup>nd</sup> edn. (Marcel Dekker, Inc., New York, 1996).
- [14] A.M. Mood, F.A. Graybill and D.C. Boes, *Introduction to Theory of Statistics*, 3<sup>rd</sup> edn. (McGraw-Hill, Inc., New York, 1974).

[15] J.S.U. Hjorth, *Computer Intensive Statistical Methods; Validation, Model Selection and Bootstrap* (Chapman & Hall, London, 1994).

**Table 1:** 95% confidence interval limits (lower and upper) for the proportions of personal crimes committed in the city of New Orleans on each of the seven days of a randomly selected week in 1984.

Week	Quesenberry and Hurst		Goodman		Fitzpatrick and Scott		Sison and Glaz	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
1	0.0763	0.1835	0.0852	0.1663	0.0677	0.1721	0.0792	0.1644
2	0.1040	0.2225	0.1145	0.2044	0.1020	0.2064	0.1135	0.1986
3	0.1058	0.2249	0.1164	0.2067	0.1041	0.2085	0.1156	0.2008
4	0.0814	0.1909	0.0906	0.1735	0.0741	0.1786	0.0857	0.1708
5	0.0865	0.1982	0.0961	0.1807	0.0805	0.1850	0.0921	0.1772
6	0.1309	0.2582	0.1428	0.2394	0.1341	0.2385	0.1456	0.2307
7	0.0797	0.1884	0.0888	0.1711	0.0720	0.1764	0.0835	0.1686
Volume	$2.6 \times 10^{-7}$		$3.7 \times 10^{-8}$		$1.4 \times 10^{-7}$		$3.2 \times 10^{-8}$	