Comparing Non-independent Proportions

Jack J. Berry and Gerardo I. Hurtado

March 4, 1994

Abstract

We are all taught in elementary statistics courses how to obtain inferences on a single binomial proportion and on the difference between two binomial proportions using independent samples. This article presents IML programs that implement the methods of Scott and Seber (1983) and Wild and Seber (1993) to handle two situations that are often encountered but not addressed by most statistics books nor a SAS procedure. The first scenario is a single multinomial population where the user wants to compare k proportions. The IML program allows the user to compute individual or simultaneous large-sample confidence intervals of linear combinations of proportions. The second scenario involves several binomial proportions from the same sample where the individual cell sizes are unknown. Because the individual cell sizes are not known, it is impossible to construct exact confidence intervals. Instead, the IML program computes conservative and optimistic large-sample confidence intervals for all the pairwise differences. Based on those results, the program labels the differences as "Significant", "Non-Significant", or "Inconclusive". As in the first scenario, the inferences can be specified as individual or simultaneous.

1 Introduction

Most elementary textbooks in statistics address the problems of estimating a single binomial proportion and the difference between two proportions using independent samples. However, no attention is given to two other often-encountered scenarios. In the first scenario, a number of individuals are classified into more than 2 mutually exclusive categories. The numbers in the categories have a multinomial distribution. We are interested in making inferences concerning the proportions in each category. As an example of this situation, consider the pre-election polls in which we want to compare the proportion of people in favor of particular candidates.

The second scenario considers the case in which several yes/no questions are posed to a group of people and we only have the total number of people that answered "yes" (or "no") to each question. That is, we don't know the frequency of each possible combination of answers, as would appear in a cross-classification table of the questions. Often our interest is in contrasting the proportions of people that answer "yes" (or "no") to the various questions. For example, consider the situation in which a set of non-exclusive yes/no questions is asked to a group of people such as "Do you consider the issues below important in the next election?", with possible issues: "Crime", "The economy", "Foreign Policy", etc.

For each scenario, this article presents some statistical background and a practical example. In the appendix, there is a program written in PROC IML that contains 2 modules, MULTIBIN and MULTINOM, to help you perform the appropriate statistical analysis.

2 Comparing Proportions from a Multinomial Sample

This section describes the first scenario. Suppose that we have a population of N individuals, of which N_i ($i = 1, 2, 3$) prefer candidate i. We take a random sample of size n and find that n_i individuals prefer candidate i. If the sampling fraction n/N is small or we sample with replacement, the joint probability distribution of the n_i's is approximately multinomial and given by

$$f(n_1, n_2, n_3) = \frac{n!}{n_1!n_2!n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}, \quad 0 \leq n_i \leq n, \quad \sum_{i=1}^{3} n_i = n,$$

where $p_i = N_i/N$. An unbiased estimator of p_i is $\hat{p}_i = n_i/n$. Consider a linear combination of the $\hat{p}_i$, say $\mathbf{c}'\hat{\mathbf{p}}$, where $\mathbf{c}$ is a column vector of constants, the quote ($'$) is the transpose operator, and $\hat{\mathbf{p}}$ is a column vector of $\hat{p}_i$ values. An example of the vector $\mathbf{c}$ could be $\mathbf{c}' = \{1 \ -1 \ 0\}$. We know that

$$E[\mathbf{c}'\hat{\mathbf{p}}] = \mathbf{c}'\mathbf{p},$$

and

$$var(\mathbf{c}'\hat{\mathbf{p}}) = \mathbf{c}'\mathbf{V}\mathbf{c},$$

where $\mathbf{V}$ is the covariance matrix of the $\hat{p}_i$,'s with elements $\{v_{ii}\} = \frac{p_i(1-p_i)}{n}$ and $\{v_{ij}\} = -\frac{p_i p_j}{n}$, $i \neq j$. (See Scott and Seber)

If $\mathbf{c}' = \{1 \ -1 \ 0\}$ then $E[\mathbf{c}'\hat{\mathbf{p}}] = E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$, and since $cov(\hat{p}_i, \hat{p}_j) = -\frac{p_i p_j}{n}$,

$$var(\mathbf{c}'\hat{\mathbf{p}}) =$$

$$\{1 \ -1 \ 0\} \ V \ \{1 \ -1 \ 0\}' =$$

$$v_{11} + v_{22} - 2v_{12}$$

$$= var(\hat{p}_1) + var(\hat{p}_2) - 2cov(\hat{p}_1, \hat{p}_2)$$

$$= [p_1(1 - p_1) + p_2(1 - p_2) + 2p_1 p_2]/n$$

$$= [p_1 + p_2 - (p_1 - p_2)^2]/n.$$

For large n, say $\min_i\{n\hat{p}_i(1-\hat{p}_i)\} > 10$, we can use the Central Limit Theorem (see Bishop et al. p. 469) to obtain an approximate 95% confidence interval in the form

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\{[\hat{p}_1 + \hat{p}_2 - (\hat{p}_1 - \hat{p}_2)^2]/n\}^{1/2}.$$

Bonferroni simultaneous $(1-\alpha)100\%$ confidence intervals for k differences can be obtained by using the $100\alpha/(2k)\%$ point of the standard normal distribution, instead of 1.96.

3 A Multinomial Example

Suppose there are 4 candidates running for their party's nomination. The candidates are Anders, Beck, Cash, and Dill. Cash is known to be the front-runner with Anders and Beck battling for second. Finishing in second is crucial if no candidate has a majority (greater than 50%). Cash is despised by the other candidates, who would unite against him in a runoff if that event were to occur. Additionally, the candidate opposing Cash in the runoff would be victorious, since he would have the support of the 2 candidates eliminated in the original election. We define

- p_1 = proportion of party voters favoring Anders
- p_2 = proportion of party voters favoring Beck
- p_3 = proportion of party voters favoring Cash
- p_4 = proportion of party voters favoring Dill

If you are interested only in testing the hypothesis that all the p_i's are equal, Schlotzhauer (1992) shows how you can do this with the CATMOD, or CAPABILITY procedures, or with a DATA step program. However, from the perspective of Anders, there are 2 specific quantities of interest: $p_1 - p_2$, and $p_1 + p_2 + p_4 - p_3$. If both quantities are positive, Anders will receive the nomination. That is, he wants to know that there will be a runoff and that he will be in it.

Both quantities are linear combinations of the p_i's. A linear combination of $p_1 \ldots p_k$ is defined as $c_1 p_1 + c_2 p_2 + \ldots + c_k p_k$, where the c_i's are constants. The quantity p_1/p_2 is not a linear combination but if you were interested in determining if $p_1/p_2 = 2$, you could rearrange the equation to read $p_1 - 2p_2 = 0$ and so your quantity of interest is now the linear combination $p_1 - 2p_2$. One way to perform inferences is to construct 95% confidence intervals for the quantities of interest using the module MULTINOM. MULTINOM is designed specifically for inferences that can be expressed as linear combinations of the p_i's, as in this example.

Suppose we now randomly sample 400 registered voters who have a preference and the results are shown below:

```
Anders  114
Beck     71
Cash    175
Dill     40
```

The MULTINOM module requires 4 parameters to be input:

- NVEC is a row vector containing the cell sizes.
- TYPE is a character string specifying whether you want the confidence limits to control individual or overall error rate. The choices are "i" (for individual) or "s" (for simultaneous).
- ALPHA is the type I error rate (such as 0.05).
- CONT is a matrix with each row containing the appropriate parameter constants defining the linear combinations of interest. If CONT is set to

missing, MULTINOM automatically constructs intervals for all pairwise $(p_i - p_j)$ differences.

To construct individual 95% confidence intervals, invoke IML, define or load the module MULTINOM, and add the following SAS/IML statements:

```
nvec={114 71 175 40};
alpha=0.05;
cont={1 -1 0 0,           /* 1st row for p1-p2        */
      1 1 -1 1};          /* 2nd row for p1+p2+p4-p3  */

run multinom(nvec,'i',alpha,cont);
```

The results are below:

```
               User-Defined Linear Combinations

        95 % Individual Large-Sample Confidence Limits

            Ci            Estimate      LCL      UCL


             1 -1  0  0     0.1075   0.0417   0.1733
             1  1 -1  1     0.1250   0.0278   0.2222
```

The values under the column "Estimate" are point estimates for the linear combination displayed in the first column under "Ci". The values under the columns "LCL" and "UCL" are lower and upper confidence limits respectively and together form the confidence interval. Since we are probably really interested in inferences on the 2 quantities at an overall type I error rate of 0.05, it actually makes more sense to obtain simultaneous confidence intervals. The method MULTINOM uses is a simple Bonferroni adjustment. The resulting intervals will be "conservative" in that they actually will have an overall type I error rate less than or equal to 0.05. Consequently, they will be slightly wider than the theoretical "exact" intervals, which are very difficult or impossible to obtain. The IML code involves the following statement:

```
run multinom(nvec,'s',alpha,cont);
```

```
                User-Defined Linear Combinations

        95 % Simultaneous Large-Sample Confidence Limits

              Ci            Estimate      LCL       UCL


               1 -1  0  0     0.1075    0.0322    0.1828
               1  1 -1  1     0.1250    0.0138    0.2362
```

If we examine the results from the simultaneous confidence limits, we see that the estimate for $p_1 - p_2$ is positive and the confidence interval does not include 0. A confidence interval that does not include 0 is equivalent to a significant result from a two-sided hypothesis test at the same type I error rate. Therefore, there is sufficient evidence to conclude a greater proportion of voters favor Anders over Beck and so Anders will be in the runoff if there is one. Specifically, we "are confident" the difference in proportions is between 3.22% and 18.28%. Similarly, there is sufficient evidence to claim a greater proportion favoring the others than those favoring Cash and so there will be a runoff and the candidate opposing Cash will win. We "are confident" the difference between Cash supporters and his combined opposition is between 1.38% and 23.62%. Overall, we are "95% confident" of the 2 inferences. If you are campaigning for the candidate Anders, you should be confident that there will be a runoff, that he will be in it, and that he will win. Of course, this assumes that the voter's preferences will not change from the survey time to the election times. A good strategy might be to curtail the number of Anders' press conferences.

4 Comparing Proportions from Several Binomial Samples

This section describes the second scenario. Consider the situation in which we pose several yes/no questions to a group of people. Since each question has only 2 possible responses, each individual can respond to the k questions in one of 2^k ways. Let $p_i,\ i = 1, 2, ..., k$ represent the probability an individual responds "yes" to question i, we want to obtain inferences on $p_i - p_j, i \neq j$. Suppose that there are N individuals in the population from which we obtain a sample of size n. Let $n_I,\ I = 1, 2, ..., 2^k$ represent the people in the corresponding categories. If we know all the n_I, we could test for "marginal homogeneity" (see Agresti, p. 390) or analyze this situation as a multinomial problem with 2^k categories and use the methodology from section 2.

Often, we are not fortunate enough to have all the n_I and have only the marginal totals. The method that we are about to describe can be used in this situation to construct confidence intervals for $p_i - p_j, i \neq j$.

For simplicity, suppose we have only k=2 questions and so only $2^2 = 4$ possible ways an individual can respond. Let $p_1 = p_{11} + p_{12}$ be the proportion in the population that answer "yes" to the first question, with p_{11} the proportion answering "yes" to both questions and p_{12} the proportion answering "yes" to the first question and "no" to the second question. Let $p_2 = p_{11} + p_{21}$ be the proportion in the population that answer "yes" to the second question, with p_{21} the proportion answering "no" to the first question and "yes" to the second question. Let p_{22} be the proportion answering "no" to both questions.

The chart below illustrates the idea:

Question 2

		Yes	No	
Question 1	Yes	p_{11}	p_{12}	p_1
	No	p_{21}	p_{22}	
		p		

In this situation, $n_1 = n_{11} + n_{12}$ and $n_2 = n_{11} + n_{21}$ but we do not know the cell sizes n_{ij}, just the marginal totals.

If the sample fraction n/N is small or we sampled with replacement, the distribution of n_{ij} is approximately multinomial. An unbiased estimate of $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 = \frac{n_1}{n} - \frac{n_2}{n}$$
$$= \frac{n_{11} + n_{12}}{n} - \frac{n_{11} + n_{21}}{n}$$
$$= \frac{n_{12}}{n} - \frac{n_{21}}{n}$$
$$= \hat{p}_{12} - \hat{p}_{21},$$

with

$$var(\hat{p}_1 - \hat{p}_2) = var(\hat{p}_{12} - \hat{p}_{21}) =$$
$$var(\hat{p}_{12}) + var(\hat{p}_{21}) - 2cov(\hat{p}_{12}, \hat{p}_{21})$$
$$= \frac{1}{n}[p_{12}(1 - p_{12}) + p_{21}(1 - p_{21}) + 2p_{12}p_{21}]$$
$$= \frac{1}{n}[p_{12} + p_{21} - (p_{12} - p_{21})^2], \tag{1}$$

since $cov(\hat{p}_{12}, \hat{p}_{21}) = -\frac{p_{12}p_{21}}{n}$.

We can't compute an estimate of $var(\hat{p}_1 - \hat{p}_2)$ because we don't have an estimate for $p_{12} + p_{21}$.

Since $p_1 = p_{11} + p_{12}$ and $p_2 = p_{11} + p_{21}$, by adding these two expressions we obtain

$$p_{12} + p_{21} = p_1 + p_2 - 2p_{11}.$$

Also,

$$p_{12} + p_{21} = (2p_{12} - p_{12}) + (2p_{21} - p_{21}) + (2p_{11} - p_{11} - p_{11})$$
$$= 2(p_{12} + p_{21} + p_{11}) - (p_{11} + p_{12}) - (p_{11} + p_{21}).$$

Since $p_{22}+p_{12}+p_{21}+p_{11}=1$, $p_1=p_{11}+p_{12}$, and $p_2=p_{11}+p_{21}$, then

$$p_{12}+p_{21}=2(1-p_{22})-p_1-p_2$$

$$=2-p_1-p_2-2p_{22}=(1-p_1)+(1-p_2)-2p_{22}.$$

Denoting $q_i=1-p_i,\ i=1,2.$, we have

$$p_{12}+p_{21}=q_1+q_2-2p_{22}.$$

Hence, from (1),

$$var(\hat{p}_1-\hat{p}_2)=\frac{1}{n}[q_1+q_2-2p_{22}-(p_1-p_2)^2],$$

and also

$$var(\hat{p}_1-\hat{p}_2)=\frac{1}{n}[p_1+p_2-2p_{11}-(p_1-p_2)^2].$$

Removing the unknown nonnegative quantities $2p_{11}$ and $2p_{22}$ from these last two equations we see that

$$var(\hat{p}_1-\hat{p}_2)\leq\frac{1}{n}[q_1+q_2-(p_1-p_2)^2],$$

and

$$var(\hat{p}_1-\hat{p}_2)\leq\frac{1}{n}[p_1+p_2-(p_1-p_2)^2].$$

and so

$$var(\hat{p}_1-\hat{p}_2)\leq\frac{1}{n}[min(p_1+p_2,q_1+q_2)-(p_1-p_2)^2]. \tag{2}$$

We now have an "upper bound" for the variance, a value we know is greater than or equal to the true variance.

Since, for any nonnegative numbers a and b, $a+b\geq\mid a-b\mid$, then $p_{12}+p_{21}\geq\mid p_{12}-p_{21}\mid$. Substituting the quantity $\mid p_{12}-p_{21}\mid$ for $p_{12}+p_{21}$ in equation (1), yields a "lower bound" for the variance, a value we know is less than or equal to the true variance:

$$var(\hat{p}_1-\hat{p}_2)\geq\frac{1}{n}\mid p_{12}-p_{21}\mid(1-\mid p_{12}-p_{21}\mid), \tag{3}$$

and since $p_{12} - p_{21} = p_1 - p_2$, we have

$$var(\hat{p}_1 - \hat{p}_2) \geq \frac{1}{n} \mid p_1 - p_2 \mid (1 - \mid p_1 - p_2 \mid), \tag{4}$$

Therefore, if we know $\hat{p}_i \;\; i = 1, 2$, we can construct "conservative" and "optimistic" confidence intervals for $p_1 - p_2$ by using (2) and (4) respectively. The conservative one uses the upper bound and is wider than the one that we would use if we knew all n_{ij}, and the optimistic one uses the lower bound and is narrower. Of particular interest is whether or not a given interval contains zero. An interval containing zero is equivalent to a non-significant result from the two-sided hypothesis test that there is a difference in proportions. Similarly, an interval that does not contain zero is equivalent to a significant result. The "exact" confidence interval, (based on (1)), the conservative interval, and the optimistic interval are all centered at the same value $\hat{p}_i - \hat{p}_j$. Consequently, if the confidence interval based on the wider conservative bound doesn't contain zero then the exact confidence interval doesn't contain it either, but if the conservative one contains zero then we don't know if the exact one contains it or not. Similarly, if the narrower optimistic confidence interval contains zero, then the exact interval contains zero also. But if it doesn't contain zero then we don't know if the exact one contains zero or not. If the conservative confidence interval contains zero and the optimistic one doesn't, we do not know whether or not the exact interval contains zero and so our result is inconclusive. This discussion is summarized in the table below:

	Optimistic Interval Contains Zero?	
	Yes	No
Yes	Non-Significant	Inconclusive

```
Conservative
Interval
Contains
Zero?

              No       Cannot Happen            Significant
```

As in the multinomial scenario, we may want our intervals (if there are more than one) to control overall type I error. The Bonferroni adjustment is a conservative one so it makes sense to do that for the intervals that are conservative originally. No such adjustment should be made on the optimistic intervals in order for their meaning to be preserved. The table above is still applicable when the intervals are "simultaneous". When the number of intervals being constructed simultaneously is large, the conservative intervals will be much wider than the optimistic intervals and so there may be many inconclusive results.

5 A Multiple Binomial Example

Suppose there is a new cold medicine "Sinoblast" on the market. The manufacturers are concerned about three things. They want the drug to reduce coughing, reduce congestion, and not cause an upset stomach. We have 3 "yes/no" questions of interest and define parameters p_1, p_2, and p_3 (which need not sum to 1) below:

- p_1=proportion of users with reduced coughing
- p_2=proportion of users with reduced congestion
- p_3=proportion of users with stomach upset

In descending magnitude, we are interested in the quantities $p_1 - p_3$, $p_2 - p_3$, and $p_1 - p_2$. If we obtain one random sample of "Sinoblast" users, we then have 3 dependent binomial random variables:

- N_1=number of users with reduced coughing
- N_2=number of users with reduced congestion
- N_3=number of users with stomach upset

Suppose our sample of 160 users yields $n_1 = 90$, $n_2 = 87$, and $n_3 = 56$. Ideally, we would like to have access to each individual's responses to the three questions. We could then test for marginal homogeneity or treat the problem as a multinomial problem with $2^3 = 8$ categories. However, if we have only the summary statistics mentioned above, we can use the MULTIBIN module to construct 95% large-sample confidence intervals. The module requires 4 parameters to be input:

- N is a positive integer representing the sample size
- NVEC is a row vector containing the number of "successes" for each question
- TYPE is a character string specifying whether you want the "Conservative" confidence limits to control individual or overall error rate. The choices are "i" (for individual) or "s" (for simultaneous). Only the "Conservative" intervals are adjusted to be simultaneous. The "Optimistic" confidence intervals are not affected by the choice of TYPE.
- ALPHA is the type I error rate (such as 0.05).

To construct simultaneous 95% confidence intervals, invoke IML, define the module MULTIBIN, and add the following SAS/IML statements:

```
n=160;
nvec={90 87 56};
alpha=0.05;

run multibin(n,nvec,'s',alpha);
```

The results are below:

```
95 % Simultaneous Large-Sample Confidence Limits For Pi-Pj

 I  J  Estimate    C LCL    C UCL    O LCL    O UCL  Result

 1  2    0.0188   -.1601   0.1976   -.0023   0.0398   N
 1  3    0.2125   0.0362   0.3888   0.1491   0.2759   S
 2  3    0.1938   0.0186   0.3689   0.1325   0.2550   S
```

The chart gives information concerning all pairwise $(p_i - p_j)$ comparisons. The "Estimate" column displays the point estimates corresponding to the given (i, j) pair. The "C LCL" and "C UCL" column show the conservative lower limits and upper limits, which are used to form the conservative confidence intervals. The "O LCL" and "O UCL" columns show the optimistic lower limits and upper limits and are used to form the optimistic confidence intervals. For a given comparison, the conservative interval will always be wider than the optimistic interval and both intervals will be centered at the point estimate. Consequently, if the conservative interval does not include 0, neither will the optimistic interval and the "Result" column is marked "S" for "Significant". If the optimistic interval contains 0, the conservative interval will also contain 0 and the "Result" column displays an "N" for "Non-significant". If the conservative interval contains 0 but the optimistic interval does not, the "Result" column displays a "?" for "Inconclusive".

If we view the output above, we see that the point estimates for $p_1 - p_3$ and $p_2 - p_3$ are positive and the results are both significant. This means that there is sufficient evidence to claim the proportions of "Sinoblast" users with reduced cough sypmtoms and reduced congestion are greater than the proportion suffering from stomach upset. The $p_1 - p_2$ point estimate is positive but the result is non–significant so we cannot conclude a difference in effectiveness of the drug between reducing coughing and reducing congestion.

6 Using The IML Modules

SAS/IML software consists of the IML (Interactive Matrix Language) procedure. PROC IML is a flexible and powerful programming language that you can use to create your own statistical analysis. To perform the analysis discussed in the previous sections, you will need to obtain the program given in the appendix. You can copy the program manually or, if you have computer access to the Internet, you can obtain the program via anonymous FTP. To use this facility, connect to *ftp.sas.com*. Once you are connected, enter the following responses as prompted:

```
Name (FTP.SAS.COM:userid): anonymous
Password: <your E-Mail address>
```

You then type

```
get /observations/v3n4/berry/multinom
quit
```

Once you have obtained the program in the appendix, you do not need to have much IML expertise. All that you need to do is submit the program, assign your parameter values, and run the appropriate module (MULTINOM or MULTIBIN), as demonstrated in sections 3 and 5. If you are running SAS interactively, you will remain in PROC IML until you submit the statement QUIT; or invoke another SAS procedure.

The program in the appendix defines 2 modules, MULTINOM and MULTIBIN. MULTINOM handles the first scenario and MULTIBIN should be used for the second one. Each module requires 4 input parameters described in the module headers. One of the parameters is the vector of cell sizes. If you are working with a raw dataset, you may want to use a SAS procedure (PROC FREQ, for example) to obtain the cell size vector.

If you plan to use a module frequently and do not want to go through the trouble of redefining it each time, you can use the LOAD and STORE commands described in Chapter 12, "Storage Features" in SAS/IML Software: Usage and Reference, Version 6, First Edition.

7 Appendix

```
proc iml;

 /*-------------------------------------------------------------*/
 /*                 The MULTINOM Module                         */
 /*                                                             */
 /* This module constructs large-sample confidence limits for   */
 /* various linear combinations of the K multinomial            */
 /* proportions.                                                */
 /*                                                             */
 /* Input:                                                      */
 /*    1) NVEC, a 1xK vector containing the cell sizes.         */
 /*    2) TYPE, a character string that specifies whether you   */
 /*       want the confidence limits to control individual or   */
```

```
/*      overall error rate.  The choices are "i"             */
/*      (for individual) or "s" (for simultaneous).          */
/*   3) ALPHA, the type I error rate.                        */
/*   4) CONT, a MxK matrix containing the M linear           */
/*      combinations of interest.  If CONT is set to missing, */
/*      MULTINOM automatically constructs confidence         */
/*      intervals for all K(K-1)/2 pairwise differences      */
/*-----------------------------------------------------------*/

start multinom(nvec,type,alpha,cont);
reset noname fw=2;
type=upcase(type);
if min(nvec)<0 | sum(abs(round(nvec,1)-nvec))>1e-4 then do;
   print "ERROR: Cell Sizes Must Be Nonnegative Integers";
   return;
end;
if type^='I' & type^='S' then do;
   print "ERROR: TYPE Parameter Must Be 'I' or 'S'";
   return;
end;
if alpha<0 | alpha>1 then do;
   print "ERROR: ALPHA Must be in [0,1]";
   return;
end;
k=ncol(nvec);
if cont^=. & ncol(cont)^=k then do;
   print "ERROR: Number of Columns in CONT Must Equal K";
   return;
end;
```

```
n=sum(nvec);
pvec=nvec/n;
varcov=(diag(pvec)-pvec`*pvec)/n;
cl=(1-alpha)#100;
c2={"Estimate" "LCL" "UCL"};
if type='I' then label='Individual';
if type='S' then label='Simultaneous';
if cont=. then do;
   c1={"I" "J"};
   print "All Pairwise Differences";
   print cl "%" label "Large-Sample Confidence Limits For Pi-Pj";
   c=round(k#(k-1)/2,1);
   contr=j(c,k,0);
   i=1;  j=2;
   do row = 1 to c;
      index=index//(i||j);
      contr[row,i]=1;
      contr[row,j]=-1;
      j=j+1;
      if j>k then do;
         i=i+1;
         j=i+1;
      end;
   end;
end;
else do;
   c1={"Ci"};
   print "User-Defined Linear Combinations";
   print cl "%" label "Large-Sample Confidence Limits";
```

```
   contr=cont;
   c=nrow(contr);
   index=cont;
end;
if type='I' then z=probit(1-alpha/2);
if type='S' then z=probit(1-alpha/(2#c));
do i = 1 to c;
   ci=contr[i,];
   mean=ci*pvec`;
   var=ci*varcov*ci`;
   inc=z#sqrt(var);
   lcl=mean-inc;
   ucl=mean+inc;
   table=table//(mean||lcl||ucl);
end;
print index[colname=c1 format=2.0] table[colname=c2 format=6.4];
finish multinom;


 /*------------------------------------------------------------*/
 /*                The MULTIBIN Module                          */
 /*                                                             */
 /* This module constructs large-sample conservative &          */
 /* optimistic confidence intervals for the pairwise            */
 /* differences between K binomial proportions when individual  */
 /* cell sizes are unknown.  Based on the two intervals, the    */
 /* the proportion differences are labeled as "significant",    */
 /* "non-significant", or "inconclusive".                       */
 /*                                                             */
```

```
 /* Input:                                                     */
 /*    1) N, the common sample size.                           */
 /*    2) NVEC, a 1xK vector containing the number of          */
 /*       "successes" for each binomial response               */
 /*    3) TYPE, a character string that specifies if you want   */
 /*       individual or simultaneous confidence inferences     */
 /*       Choices are "i" or "s".  Only the "conservative      */
 /*       intervals are adjusted to be simultaneous.           */
 /*    4) ALPHA, the type I error rate.                        */
 /*-----------------------------------------------------------*/

start multibin(n,nvec,type,alpha);
reset noname;
type=upcase(type);
if n<=0 | abs(round(n,1)-n)>1e-4 then do;
   print "ERROR: N Must Be A Positive Noninteger";
   return;
end;
if min(nvec)<0 | max(nvec)>n |
   sum(abs(round(nvec,1)-nvec))>1e-4 then do;
   print "ERROR: NVEC Values Must be Integers & in [0,N]";
   return;
end;
if type^="I" & type^="S" then do;
   print "ERROR: TYPE Parameter Must Be 'I' or 'S'";
   return;
end;
if alpha<0 | alpha>1 then do;
   print "ERROR: ALPHA Must be in [0,1]";
```

```
   return;
end;
k=ncol(nvec);
pvec=nvec/n;
qvec=1-pvec;
c=round((k#(k-1)/2),1);
cl=100#(1-alpha);
zo=probit(1-alpha/2);
if type='I' then do;
   zc=zo;
   label='Individual';
end;
if type='S' then do;
   zc=probit(1-alpha/(2#c));
   label='Simultaneous';
end;
do i = 1 to k-1;
   pi=pvec[1,i];
   qi=qvec[1,i];
   do j = i+1 to k;
      result="  ?";
      pj=pvec[1,j];
      qj=qvec[1,j];
      mean=pi-pj;
      cvar=(1/n)#(min(pi+pj,qi+qj)-(pi-pj)##2);
      cinc=zc#sqrt(cvar);
      clcl=mean-cinc;
      cucl=mean+cinc;
      if clcl>0 | cucl<0 then result="  S";
```

```
      diff=abs(pi-pj);
      ovar=(1/n)#diff#(1-diff);
      oinc=zo#sqrt(ovar);
      olcl=mean-oinc;
      oucl=mean+oinc;
      if olcl<=0 & oucl>=0 then result="  N";
      index=index//(i||j);
      table=table//(mean||clcl||cucl||olcl||oucl);
      rvec=rvec//result;
   end;
end;
ch1={"I" "J"};
ch2={"Estimate" "C LCL" "C UCL" "O LCL" "O UCL"};
ch3={"Result"};
print cl "%" label "Large-Sample Confidence Limits For Pi-Pj";
print index[colname=ch1 format=2.0]
      table[colname=ch2 format=6.4] rvec[colname=ch3];
finish multibin;
```

8 References

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley and Sons, Inc.

Bishop, Y. M. M., Fienberg, S. E., Holland, P W. (1975). *Discrete Multivariate Analysis.* Cambridge, Massachusetts: The MIT Press.

Goodman, L. A. (1965). "On Simultaneous Confidence Intervals for Multinomial Proportions" *Technometrics*, Vol. 7, No. 2, 247–254.

SAS Institute Inc. (1990), *SAS/IML Software, Version 6, First Edition*, Cary, NC: SAS Institute,Inc.

Schlotzhauer, D. (1992). "Computing Goodness-of-fit Chi-square Statistics with SAS Software" *Observations: The Technical Journal for SAS Software Users*, Vol. 1, No. 1, 17–24.

Scott, A. J. and Seber, G. A. F. (1983). "Difference of Proportions From the Same Survey" *The American Statistician*, Vol. 37, No. 4, 319–320.

Wild, C. J. and Seber, G. A. F. (1993). "Comparing Two Proportions From the Same Survey", *The American Statistician*, Vol. 47, No. 3, 178–181.