Large sample simultaneous confidence intervals for any combination of cell probabilities*†‡

Tommy Norberg§¶
CTH & GU

January 25, 1996

Abstract

The note provides a method for obtaining simultaneous confidence intervals for any combination of cell probabilities of a multinomial distribution. Large sample size is assumed.

1 Introduction

Consider an experiment in which k disjoint events, $A_1, \ldots, A_k$ say, may occur. Suppose it is independently repeated n times. Let $n_1, \ldots, n_k$ denote the observed frequencies. Then $n_1, \ldots, n_k$ is distributed according to a multinomial distribution with parameters $n = n_1 + \ldots + n_k$ and $p_1, \ldots, p_k$, where $p_i = \mathbb{P}(A_i) > 0$, $i = 1, \ldots, k$. Thus the probability mass function of $n_1, \ldots, n_k$ is

$$f(n_1, \ldots, n_k) = \binom{n}{n_1, \ldots, n_k} p_1^{n_1} \cdots p_k^{n_k}$$

It is well known that, for n sufficiently large, the χ^2-distance $\sum_{i=1}^k (n_i - np_i)^2/(np_i)$ is distributed approximately as a χ^2-variable with $k-1$ degrees of freedom. Hence the set

$$C = \left\{ p_1, \ldots, p_k : \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \le c_{k-1,\alpha} \right\}$$

*Supported by the Swedish Natural Science Research Council

†*AMS 1991 Subject Classification:* Primary 62D05; secondary 62G15, 62G20.

‡*Key words and phrases:* Multinomial probabilities, confidence region, asymptotic confidence level

§Department of Mathematics, Chalmers University of Technology and Göteborg University, S-412 96 Göteborg, Sweden

¶*Email:* tommy@math.chalmers.se; *www:* http://www.math.chalmers.se/~tommy/

is an approximate $100(1-\alpha)\%$ confidence region for the cell probabilities $p_1, \dots, p_k$, provided the critical value $c_{k-1,\alpha}$ is choosen to satisfy

$$\mathbb{P}\{\chi^2_{k-1} > c_{k-1,\alpha}\} = \alpha$$

where χ^2_{k-1} denotes a random variable having a χ^2-distribution with $k-1$ degrees of freedom. Below we will assume that n is large enough for this approximation to be valid.

Quesenberry and Hurst [Quesenberry & Hurst 1964] have shown that the k confidence intervals

$$p_{i1} \le p_i \le p_{i2}, \quad i = 1, \dots, k \tag{1}$$

where

$$p_{i1} = \frac{2n_i + c_{k-1,\alpha} - \sqrt{c_{k-1,\alpha}(c_{k-1,\alpha} + 4n_i(n-n_i)/n)}}{2(n + c_{k-1,\alpha})}$$

$$p_{i2} = \frac{2n_i + c_{k-1,\alpha} + \sqrt{c_{k-1,\alpha}(c_{k-1,\alpha} + 4n_i(n-n_i)/n)}}{2(n + c_{k-1,\alpha})}$$

have a simultaneous confidence coefficient approximately $1-\alpha$. Note that the bounds p_{i1}, p_{i2} may be obtained as the two solutions of the quadratic equation

$$(n_i - np_i)^2 = c_{k-1,\alpha} np_i(1-p_i)$$

Hence

$$\mathbb{P}\left(\bigcap_{i=1}^{k} \{p_i \in C_i\}\right) \gtrapprox 1 - \alpha \tag{2}$$

is an alternative formulation of Quesenberry and Hurst's result. Here

$$C_i = \left\{ p_i : \frac{(n_i - np_i)^2}{np_i(1-p_i)} \le c_{k-1,\alpha} \right\}$$

for $i = 1, \dots, k$.

Suppose now that an experimenter is interested, not only in the marginal probabilities $p_1, \dots, p_k$, but also in some probabilities of the kind

$$p_B = \sum_{i \in B} p_i = \mathbb{P}\left(\bigcup_{i \in B} A_i\right)$$

where $\emptyset \ne B \subsetneqq \{1, \dots, k\}$. Define

$$C_B = \left\{ p_B : \frac{(n_B - np_B)^2}{np_B(1-p_B)} \le c_{k-1,\alpha} \right\}$$

where $n_B = \sum_{i \in B} n_i$. Our aim is to prove the following extension of (2):

$$\mathbb{P}\left(\bigcap_{B} \{p_B \in C_B\}\right) \gtrapprox 1 - \alpha \tag{3}$$

Thus the whole collection of sets C_B, where $\emptyset \neq B \subsetneq \{1, \ldots, k\}$, forms a family of confidence intervals having simultaneous confidence coefficient approximately $1-\alpha$. For a proof, refer to Section 2.

Goodman [Goodman 1965] argued that Quesenberry and Hurst's intervals (1) can be made shorter in general, by replacing $c_{k-1,\alpha}$ with $c_{1,\alpha/k}$, so that, for any i,

$$\mathbb{P}\{p_{i1} \leq p_i \leq p_{i2}\} \gtrapprox 1 - \frac{\alpha}{k}$$

and then, by the Bonferroni inequality,

$$\mathbb{P}\{p_{i1} \leq p_i \leq p_{i2},\ 1 \leq i \leq k\} \gtrapprox 1 - \alpha$$

This approach is not possible for us, since the number of intervals in (3) is large even for moderate k.

We have already remarked that

$$\frac{(n_B - np_B)^2}{np_B(1-p_B)} = c_{k-1,\alpha}$$

if, and only if,

$$p_B = \frac{2n_B + c_{k-1,\alpha} \pm \sqrt{c_{k-1,\alpha}(c_{k-1,\alpha} + 4n_B(n-n_B)/n)}}{2(n + c_{k-1,\alpha})}$$

Thus the center of the interval C_B is

$$\frac{2n_B + c_{k-1,\alpha}}{2(n + c_{k-1,\alpha})} = \hat{p}_B + (\frac{1}{2} - \hat{p}_B)\frac{c_{k-1,\alpha}}{n + c_{k-1,\alpha}}$$

where $\hat{p}_B = n_B/n$. Moreover, the width of the interval C_B is (after algebraic manipulations) seen to equal

$$\sqrt{c_{k-1,\alpha}}\frac{\sqrt{c_{k-1,\alpha} + 4n\hat{p}_B(1-\hat{p}_B)}}{n + c_{k-1,\alpha}}$$

For large enough n this is approximately

$$2\sqrt{c_{k-1,\alpha}}\sqrt{\frac{\hat{p}_B(1-\hat{p}_B)}{n}} \tag{4}$$

Divide by the width of the corresponding per comparison interval (which you get by replacing $c_{k-1,\alpha}$ with $c_{1,\alpha}$) and let the sample size n tend to infinity, to get

$$\sqrt{\frac{c_{k-1,\alpha}}{c_{1,\alpha}}} \tag{5}$$

Thus the simultaneous intervals are asymptotically $\sqrt{c_{k-1,\alpha}/c_{1,\alpha}}$ times as long as the per comparison ones.

A plausible conclusion from (4) is that the intervals

$$p_B = \hat{p}_B \pm \sqrt{c_{k-1,\alpha}}\sqrt{\frac{\hat{p}_B(1-\hat{p}_B)}{n}}, \quad \emptyset \neq B \subsetneqq \{1,\ldots,k\} \tag{6}$$

have simultaneous confidence coefficient approximately $1-\alpha$. That this is so follows immediately by Theorem 3 of Gold [Gold 1963].

Note that (3) does not follow from Theorem 3 of [Gold 1963] (at least not immediately). Instead it is a link between Quesenberry and Hurst's result (1) [Quesenberry & Hurst 196 and Gold's Theorem 3 [Gold 1963].

It is interesting to compare (6) with the usual per comparison confidence interval

$$p_B = \hat{p}_B \pm \sqrt{c_{1,\alpha}}\sqrt{\frac{\hat{p}_B(1-\hat{p}_B)}{n}}$$

for p_B based on the fact that $(\hat{p}_B - p_B)/\sqrt{\hat{p}_B(1-\hat{p}_B)/n}$ is asymptotically normal with mean 0 and variance 1.

2 Proof

First note that

$$\frac{(n_B - np_B)^2}{np_B(1-p_B)} = \frac{(n_B - np_B)^2}{np_B} + \frac{(n_{B^c} - np_{B^c})^2}{np_{B^c}}$$

where $n_{B^c} = n - n_B = \sum_{i\notin B} n_i$ and $p_{B^c} = 1 - p_B = \sum_{i\notin B} p_i$. Next note that

$$\frac{(n_B - np_B)^2}{np_B} = \frac{\left(\sum_{i\in B}(n_i - np_i)\right)^2}{\sum_{i\in B} np_i} \leq \sum_{i\in B}\frac{(n_i - np_i)^2}{np_i}$$

The inequality follows by the following chain of equivalences

$$\frac{(x_1+x_2)^2}{a_1+a_2} \leq \frac{x_1^2}{a_1} + \frac{x_2^2}{a_2}$$

$$\iff \quad (x_1+x_2)^2 \leq x_1^2\left(1+\frac{a_2}{a_1}\right) + x_2^2\left(1+\frac{a_1}{a_2}\right)$$

$$\iff \quad 2x_1x_2 \leq x_1^2\frac{a_2}{a_1} + x_2^2\frac{a_1}{a_2} = (ax_1)^2 + (x_2/a)^2$$

$$\iff \quad 0 \leq (ax_1)^2 - 2(ax_1)(x_2/a) + (x_2/a)^2 = (ax_1 - x_2/a)^2$$

(where $a = \sqrt{a_2/a_1}$) and induction. The fact that

$$\frac{(n_{B^c} - np_{B^c})^2}{np_{B^c}} \leq \sum_{i \notin B} \frac{(n_i - np_i)^2}{np_i}$$

follows also. Hence

$$\frac{(n_B - np_B)^2}{np_B(1 - p_B)} \leq \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

showing that if $p_1, \ldots, p_k \in C$, then $p_B \in C_B$. This proves our claim that (3) holds true for sufficiently large n.

3 Example

Consider a typical political opinion survey in a country having seven political parties in its parliament. Table 1 shows a thought example in which $n = 3366$ people were asked to say which political party they prefer for the moment. The error bounds in its right most column are per comparison 95% confidence intervals. Typically this would be the error claimed by the institute making the survey. (Though the improvement is not very big, most survey institutes nowadays use more subtle stratified estimators having less variance than the naive ones we consider.)

party i	n_i	$\hat{p}_i$	error bounds
1	1227	0.365	$[0.348, 0.381]$
2	846	0.251	$[0.237, 0.267]$
3	375	0.111	$[0.101, 0.123]$
4	348	0.103	$[0.093, 0.115]$
5	249	0.074	$[0.065, 0.084]$
6	201	0.060	$[0.052, 0.069]$
7	120	0.036	$[0.029, 0.043]$
	3366	1.000	

Table 1: Result of thought political opinion survey. The error bounds are 95% per comparison confidence intervals.

Compare with Table 2, which shows the estimated support of various political configurations together with their 95% simultaneous (or experimentwise) confidence intervals.

It is interesting to note that in order to get the error bounds of Table 1 to coincide with the simultaneous confidence intervals of Table 2, either the latter would have to have a confidence coefficient as low as $\mathbb{P}\{\chi_6^2 \leq c_{1,.05}\} = .30$ or the

political constellation	estimated support	confidence interval
1	0.365	$[0.335, 0.395]$
2	0.251	$[0.225, 0.279]$
3	0.111	$[0.093, 0.133]$
4	0.103	$[0.086, 0.124]$
5	0.074	$[0.059, 0.092]$
6	0.060	$[0.046, 0.076]$
7	0.036	$[0.025, 0.049]$
$1, 3$	0.476	$[0.445, 0.507]$
$1, 5$	0.439	$[0.408, 0.469]$
$3, 4$	0.215	$[0.190, 0.241]$
$2, 5, 6, 7$	0.421	$[0.390, 0.451]$
$2, 6, 7$	0.347	$[0.318, 0.377]$

Table 2: 95% simultaneous confidence intervals for various political constellations.

former one as high as $\mathbb{P}\{\chi_1^2 \leq c_{6,.05}\} = .9996$. This large discrepancy is of course due to the fact that there are as many as seven political parties in the survey.

Another way of measuring the loss is to calculate the asymptotic ratio between the lengths of the simultaneous and the per comparison 95% confidence intervals

$$\sqrt{\frac{c_{6,.05}}{c_{1,.05}}} \approx 1.81$$

(cf (5)). Thus, the simultaneous confidence intervals are 1.81 times as long as the per comparison ones (in the limit as $n \to \infty$).

References

[Gold 1963] Ruth Z. Gold: Test auxiliary to χ^2 tests in a Markov chain. *Ann. Math. Statist.*, **34**, 56-74 (1963).

[Goodman 1965] Leo A. Goodman: On simultaneous confidence intervals for multinomial proportions. *Technometrics* **7**, 247-254 (1965).

[Quesenberry & Hurst 1964] C. P. Quesenberry & D. C. Hurst: Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics* **6**, 191-195 (1964).

Extended abstract

Assume the frequencies $n_1, \ldots, n_k$ to be distributed according to a multinomial distribution with parameters $n = n_1 + \ldots + n_k$ and $p_1, \ldots, p_k > 0$, with $\sum_i p_i = 1$. For large n, the χ^2-distance $\sum_{i=1}^k (n_i - np_i)^2/(np_i)$ is approximately χ^2-distributed with $k-1$ degrees of freedom. Hence the set

$$C = \left\{ p_1, \ldots, p_k : \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \leq c \right\}$$

is an approximate $100(1-\alpha)\%$ confidence region for the probabilities $p_1, \ldots, p_k$, provided the critical value c is chosen to satisfy $\mathbb{P}\{\chi^2_{k-1} > c\} = \alpha$, where χ^2_{k-1} denotes a random variable having a χ^2-distribution with $k-1$ degrees of freedom.

For $\emptyset \neq B \subsetneqq \{1, \ldots, k\}$ define $p_B = \sum_{i \in B} p_i$ and let $n_B = \sum_{i \in B} n_i$. The aim of this note is to prove that if $p_1, \ldots, p_k \in C$, then

$$p_B \in C_B = \left\{ p_B : \frac{(n_B - np_B)^2}{np_B(1-p_B)} \leq c \right\}$$

Hence the sets C_B, $\emptyset \neq B \subsetneqq \{1, \ldots, k\}$, forms a family of confidence intervals having simultaneous confidence coefficient approximately $100(1-\alpha)\%$.