Interval estimates and their precision

Luboš Marek, Michal Vrabec
University of Economics, Prague, Faculty of Informatics and Statistics, W. Churchill Sq. 4, CZ – 130 67 Prague 3
marek@vse.cz, vrabec@vse.cz

Abstract. A task very often met in practice is computation of confidence interval bounds for the relative frequency within sampling without replacement. A typical situation includes pre-election estimates and similar tasks. In other words, we build the confidence interval for the parameter value \( M \) in the parent population of size \( N \) on the basis of a random sample of size \( n \). There are many ways to build this interval. We can use a normal or binomial approximation. More accurate values can be looked up in tables. We consider one more method, based on MS Excel calculations. In our paper we compare these different methods for specific values of \( M \) and we discuss when the considered methods are suitable. The aim of the article is not a publication of new theoretical methods. This article aims to show that there is a very simple way how to compute the confidence interval bounds without approximations, without tables and without other software costs.

1. Introduction
A frequent task undertaken in public opinion polls is that of estimating election preferences of political parties, that is, estimating a proportion of voters who are going to vote for party "xy" (all of our text will strictly avoid references to any particular parties). The tasks of this type, we meet very often and need not be only about elections. From the viewpoint of statistics it means finding a point or interval estimate of an unknown parameter value for hypergeometric distribution – in principle estimating the relative frequency within sampling without replacement. In such types of problems we usually work with approximations, most often those using the normal distribution. In this paper we are, however, going to show that this approach may often lead to inaccurate results. Moreover, agencies always publish only point estimates, completely ignoring interval estimates, which are more informative, especially in this case, even though the interval estimates are easy to find.

Our task therefore is to estimate an unknown parameter value \( M \) or \( P \) in a parent population of size \( N \) on the basis of a random sample of size \( n \). There are several approaches to such estimation. In this paper we are going to mutually compare several methods and suggest our own very simple, accurate and efficient method for finding the confidence interval bounds of the unknown parameter value \( M \), even without using any specialized statistical software. This means that regarding potential errors we only focus on statistical properties of the estimates (if a correct theory is not used, it would be useless to consider other kinds of errors). In other words, it is the only component of the total error – a discrepancy between the normal distribution (routinely used for the construction of interval estimates) and the exact construction based on the actual distribution of the sample relative frequency (binomial and hypergeometric).
In the following text, we describe the commonly used methods and we compare the results achieved by simulations. We discuss when the considered methods are suitable and we suggest very easy method how to obtain the best results with using MS Excel software.

2. Methodology

Let us first focus on the construction of confidence intervals for relative (and possibly absolute) frequencies within the sampling without replacement. We derive a $100(1 - \alpha)\%$ confidence interval for parameter $P$ (relative frequency) as

$$P' < P < P''$$

where $P$ is the relative frequency, $P'$ is the lower bound of the confidence interval, and $P''$ is its upper bound. The confidence interval for absolute frequency $M$ is given by the formula

$$M' < M < M''$$

where $M$ is the absolute frequency, $M'$ is the lower bound of the confidence interval, and $M''$ is its upper bound. Values $P', P, P''$ and $M', M, M''$ are governed by relations

$$P = \frac{M}{N}, \quad P' = \frac{M'}{N} \quad \text{and} \quad P'' = \frac{M''}{N}.$$  

Value $1 - \alpha$ stands here for the confidence level (most often it is $1 - \alpha = 0.95$, that is, $\alpha = 0.05$, hence we speak about the 95% confidence interval). Let us denote by $\alpha_1$ probability $P(P < P') = \alpha_1$ and by $\alpha_2$ probability $P(P > P'') = \alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$. Usually it is $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$.

The lower bound of the confidence interval $M'$ (for the absolute frequency), or the lower bound $P'$ (for relative frequency) is given by solving the following equation (for given $m$ and $\alpha_1$; here $m$ is the "success" count in a sample of size $n$)

$$\sum_{x=m}^{\infty} \binom{M'}{x} \binom{N-M'}{n-x} \frac{\binom{N}{n}}{\binom{n}{x}} \leq \alpha_1 \quad \text{or} \quad \sum_{x=m}^{\infty} \binom{M' P'}{x} \binom{N - (1 - P')}{n-x} \frac{\binom{N P'}{n}}{\binom{n}{x}} \leq \alpha_1$$

This expression equals $1 - F(m-1)$, where $F$ is the distribution function of the hypergeometric distribution with parameters $M', N$, and $n$; or parameters $NP', N$, and $n$. 


The upper bound $M''$ or $P''$ (for the relative frequency) of the confidence is obtained by solving the following equation (for given $m$ and $\alpha$; here $m$ is the "success" count in a sample of size $n$)

$$\sum_{x=0}^{m} \binom{M''}{x} \binom{N-M''}{n-x} \frac{N}{n} \pm \alpha_{1}, \quad \text{or} \quad \sum_{x=0}^{m} \binom{N-P''}{x} \binom{N(1-P'')}{n-x} \frac{N}{n} \pm \alpha_{1}. \quad (5)$$

This expression equals $F(m)$, where $F$ is the distribution function of the hypergeometric distribution with parameters $M''$, $N$, and $n$; or parameters $NP''$, $N$, and $n$.

Setting $\alpha_{1} = 0$ (i.e., $\alpha_{1} = \alpha$), we obtain the right-hand confidence interval $0 \leq P \leq P''$. On the other hand, setting $\alpha_{1} = 0$ (i.e., $\alpha = \alpha$), we obtain the left-hand confidence interval $P' \leq P \leq 1$. If $\alpha_{1} = \alpha_{2} = \frac{\alpha}{2}$ we get the bilateral confidence interval $P' < P < P''$. Solving the presented equations is rather complicated, which explains why no "fully adequate" tables of values $P'$ and $P''$ have been published in the literature yet. Moreover, sufficiently detailed tables (for values of $N$, $n$, and $m$) would be too large. Hence we use in our examples monographs published in [5].

Methods we briefly describe below can be used for determining the lower and upper bounds of the confidence interval:

- tables,
- normal approximation,
- normal approximation with correction to discontinuity,
- binomial approximation,
- Goal Seek procedure in MS Excel spreadsheet.

2.1. Tables
The tabled values were calculated numerically (without any approximations) and these values therefore appear to be most accurate among all methods considered in this paper. The values can be found in [4], [5] and [6].
2.2. Approximation 1 – normal distribution
If the sample size is sufficiently large and the relative frequency \( M/N \) to be estimated is sufficiently close to \( 1/2 \), say, \( n \cdot M/N > 30 \), cf. [3], the hypergeometric distribution converges, for fixed \( M/N \) and \( n \to \infty \), to the normal distribution. If this is the case, the bounds of the confidence interval can be calculated with the aid of this approximation. In practice we of course work with a finite but sufficiently large sample of size \( n \). The limit distribution of the statistic \( p = m/n \) is then normal with parameters:

\[
\mu = P = \frac{M}{N} \quad \text{and} \quad \sigma^2 = \frac{1}{n} \frac{P(1-P)}{N} \frac{N-n}{N-1}.
\]

The above-cited equations then enable us to derive that the \( 100(1-\alpha)\% \) bilateral confidence interval for the unknown parameter \( P \) is given as

\[
p - u_{\alpha/2} \sqrt{\frac{p(1-p)}{n-1} \frac{N-n}{N}} < p < p + u_{\alpha/2} \sqrt{\frac{p(1-p)}{n-1} \frac{N-n}{N}}.
\]

(6)

2.3. Approximation 2 – normal distribution with correction to discontinuity
The approximation of the hypergeometric distribution by normal may be made more accurate by taking into account the discontinuity (cause by our approximating a discrete distribution by a continuous one), that is, adding a corrective term \( 1/(2n) \). More detailed (statistical) reasoning about this term can be found, e.g., in [4]. This implies

\[
p = \frac{1}{2n} - u_{\alpha/2} \sqrt{\frac{p(1-p)}{n-1} \frac{N-n}{N}} < p < p + \frac{1}{2n} + u_{\alpha/2} \sqrt{\frac{p(1-p)}{n-1} \frac{N-n}{N}}.
\]

(7)

2.4. Approximation 3 – binomial distribution
Here an approximation will again be used; this time based on a relationship between the probability functions of the binomial and hypergeometric distributions. It can be proven that, for \( N \to \infty \) and fixed \( n \), the hypergeometric distribution converges to the binomial distribution "faster" than to normal. The desired confidence interval is then given as

\[
p - \frac{1}{2n} (p - \frac{1}{2n} - p') \sqrt{\frac{N-n}{N}} < p < p + \frac{1}{2n} (p' - p - \frac{1}{2n}) \sqrt{\frac{N-n}{N}},
\]

(8)

where \( p' \) is the lower bound and \( p'' \) is the upper bound of the confidence interval for the binomial distribution's parameter. The values can be calculated with the aid of known formulas or looked up in statistical tables. Details can be found in [1] and [2].

2.5. Goal Seek procedure in MS Excel
The following formulas are easily derived from (4) and (5). In fact we obtain

\[
\alpha_i = 1 - F(m-1),
\]

equivalently

\[
1 - \alpha_i = F(m-1),
\]

implying
$F_m = -F(m-1), \quad (11)$

where $F$ is the distribution function of hypergeometric distribution. We utilize these relationships in the Goal Seek procedure. We look for the 95% confidence interval for parameter $M$. The entire situation is illustrated in Figure 2.

![Figure 2. Goal Seek procedure in MS Excel.](image)

The theory described above can be studied, e.g., in [1], [2] and [3].

3. Simulations

The above theory will be applied to simulated data to mutually compare the results of the considered methods. Such comparison will indicate situations suitable for the use of this or that method.

3.1. Simulation 1

Let us assume a random sample of size 100 (sampling without replacement from the parent population of size $N = 500$) in which 50 units pertain to the observed attribute, that is, $n = 100$, $m = 50$, and consequently $p = 0.5$. The unknown value $P$ of the underlying relative frequency lies between the bounds $(P', P'')$ of the confidence interval with the confidence value of 0.95 (alternatively we could consider the absolute frequency value $M$). Most exact values for the bounds of the confidence interval can be looked up in tables. All other methods (based on approximations) provide less accurate results. The obtained results are shown in the Table 1.

**Table 1. Absolute frequency – Simulation 1.**

| Absolute frequency method          | lower $M'$ | point $M$ | upper $M''$ | difference $M''-M'$ |
|-----------------------------------|------------|-----------|-------------|---------------------|
| Tables                            | 205        | 250       | 296         | 91                  |
| Approximation 1 – normal          | 206        | 250       | 294         | 88                  |
| Approximation 2 – normal with correction | 203     | 250       | 297         | 94                  |
| Approximation 3 – binomial        | 209        | 250       | 296         | 87                  |
| MS Excel                          | 204        | 250       | 296         | 92                  |

**Table 2. Relative frequency – Simulation 1.**

| Absolute frequency method | lower $P'$ | point $P$ | upper $P''$ | difference $P''-P'$ |
|---------------------------|------------|-----------|-------------|---------------------|
| Tables                    | 0.409      | 0.5       | 0.591       | 0.182               |
The 95% both sided confidence interval solutions for the relative frequency $P$ can, for all the considered methods, be graphically displayed as line segments representing the interval lengths:

Figure 3. Method comparison in Simulation 1.

Intervals are ordered from the bottom upwards. That is, values from the tables are at the bottom, Excel results are at the top. At first sight we can see, from both the Table and the Figure, that at a relative frequency value of 0.5 in the parent population, all intervals are about the same and the results are comparable with minor variations; and the tables and MS Excel provide identical results.

### 3.2. Simulation 2

Let us again assume a random sample of size 100 (sampling without replacement from the parent population of size $N = 10000$), in which 50 units pertain to the observed attribute, that is, $n = 100$, $m = 5$, and consequently $p = 0.05$. The unknown value $P$ of the underlying relative frequency lies between the bounds $(P', P'')$ of the confidence interval with the confidence value of 0.95 (alternatively we could consider the absolute frequency value $M$). Most exact values for the bounds of the confidence interval can again be looked up in tables. All other methods (based on approximations) provide less accurate results. The obtained results are shown in the Table.

#### Table 3. Absolute frequency – Simulation 2.

| Absolute frequency method | lower point $M'$ | point $M$ | upper point $M''$ | difference $M''-M'$ |
|---------------------------|------------------|-----------|-------------------|---------------------|
| Tables                    | 170              | 500       | 1130              | 960                 |
| Approximation 1 – normal  | 100              | 500       | 927               | 827                 |
| Approximation 2 – normal with correction | 23               | 500       | 977               | 954                 |
| Approximation 3 – binomial| 248              | 500       | 1124              | 876                 |
We can again see the 95% both-sided confidence intervals for the relative frequency as line segments representing the interval lengths:

![Figure 4. Method comparison in Simulation 2.](image)

At first sight we can now see, both from the Table and the Figure, that the achieved results are sharply different and no general comparability takes place. Only two methods, namely, tables and MS Excel, provide identical results.

### 4. Conclusions based on simulations

- If the $P$ parameter value is close to 0.5 (it is even sufficient for it to be between 0.1 and 0.9 (with respect to the sizes of the parent population and the sample), the achieved results are very similar to each other and the methods may be deemed comparable. It is nearly irrelevant which of them is used. No substantial difference exists between the values from tables and those from MS Excel. If we apply this principle to the data from opinion polls, this case covers the instances
of parties with large preference proportions, but not those whose preferences are smaller than 10%.
• If parameter $P$ fulfills the condition $P < 0.1$ or $P > 0.9$ (both meaning essentially the same situation), the tabled values should be used, or alternatively the Goal Seek procedure in MS Excel.
• The Goal Seek procedure in MS Excel provides good results in all cases, regardless of the value of $P$. This procedure has another advantage – cost savings on software, because no specialized statistical software is necessary for it. MS Excel is more or less standard installation on any PC, so costs due to its use are practically none. Another argument in favor of the MS Excel procedure is its simplicity and availability.
• the correct procedure (estimation of parameter $P$) has political and economic impacts - estimating election preferences of political parties, market research, public opinion polls and others.

References
[1] Buonacorsi J P 1987 A note on confidence intervals for proportions in finite populations (American Statistician, 41, p. 215 – 218)
This reference has two entries but the second one is not numbered (it uses the ‘Reference (no number)’ style.
[2] Burstein H 1975 Finite population correction for binomial confidence limits (JASA, 70, p. 67 – 69)
[3] Cochran W G 1977 Sampling Techniques, 3rd edition (New York, Wiley)
[4] Čermák V and Vrabec M, 1998 Sampling Survey Theory (in Czech: Teorie výběrových šetření). University of Economics, Prague. Faculty of Informatics and Statistics
[5] Chung J H and Delury D B, 1950 Confidence Limits for the Hypergeometric Distribution (Toronto, University of Toronto Press)
[6] Quesenberry C P and Hurst D C 1964 Large sample simultaneous confidence intervals for multinomial proportions (Technometrics, 6, p. 191 – 195)

We would like to acknowledge financial support from P402/12/G097.