CONFIDENCE INTERVAL FOR THE WEIGHTED SUM OF TWO BINOMIAL PROPORTIONS

Abstract. Interval estimation of the probability of success in a binomial model is considered. The classical confidence interval is compared with a confidence interval which uses information about non-homogeneity of the sample.

Introduction. Suppose that there are two suppliers of an item with unknown defectiveness $\theta_1$ and $\theta_2$, respectively. It is known that the share of supply provided by the first supplier is $w_1$, while the share of the second one equals $w_2 = 1 - w_1$. It is of interest to estimate the overall defectiveness $\theta = w_1 \theta_1 + w_2 \theta_2$ on the basis of a sample of size $n$.

Similar problems have already been considered. Especially, the problem of constructing confidence intervals for the difference of probabilities of success has been widely studied, due to its numerous applications in biostatistics and elsewhere—see e.g. Anbar (1983), Newcombe (1998), Zhou, Tsao & Qin (2004). Decrouez & Robinson (2012) considered the more general problem of estimating a linear combination $w_1 \theta_1 + w_2 \theta_2$ ($w_1, w_2 \neq 0$). They constructed and compared the Wald interval, the Haldane and Jeffreys–Perks interval, the modified Wald interval, the score interval and the likelihood-ratio interval. Unfortunately the confidence levels of those confidence intervals are less than the nominal one, which is in contradiction to Neyman’s (1934) definition of confidence intervals.

The approach of Decrouez & Robinson (2012) is as follows: first we estimate $\theta_1$ and $\theta_2$ and then we construct a confidence interval for $\theta =$

---

2010 Mathematics Subject Classification: Primary 62F25.

Key words and phrases: confidence interval, binomial distribution, binomial proportions.

Received 8 September 2017.

Published online *.

DOI: 10.4064/am2349-12-2017 [1] © Instytut Matematyczny PAN, 2018
In what follows, another approach is applied. We want to construct a confidence interval for \( \theta \), and we are not interested in estimating \( \theta_1 \) and \( \theta_2 \). Note that for given \( \theta \) there are infinitely many \( \theta_1 \) and \( \theta_2 \) giving \( \theta \). Hence averaging with respect to \( \theta_1 \) and \( \theta_2 \) is applied.

We confine ourselves to the case \( w_1, w_2 > 0 \) and \( w_1 + w_2 = 1 \), because of its nice interpretation mentioned above.

**Confidence interval.** Let \( \xi \) be the number of successes in \( n \) trials. This is a random variable binomially distributed. The statistical model for \( \xi \) is

\[
\{\{0, 1, \ldots, n\}, \{\text{Bin}(n, \theta), \theta \in (0, 1)\}\}
\]

and \( \hat{\theta}_c = \xi/n \) is the unbiased estimator with minimal variance of the parameter \( \theta \). Its variance equals \( \theta(1 - \theta)/n \). Let \( \hat{\theta}_c = u \) be observed. The Clopper–Pearson symmetric confidence interval at confidence level \( \gamma \) for \( \theta \) has the form \((\hat{\theta}_c^L(u), \hat{\theta}_c^U(u))\), where

\[
\hat{\theta}_c^L(u) = \begin{cases} 
0 & \text{for } u = 0, \\
B^{-1}(n(1 - u) + 1, nu; (1 + \gamma)/2) & \text{for } u > 0, 
\end{cases}
\]

\[
\hat{\theta}_c^U(u) = \begin{cases} 
1 & \text{for } u = 1, \\
B^{-1}(n(1 - u), nu + 1; (1 - \gamma)/2) & \text{for } u < 1. 
\end{cases}
\]

Here \( B^{-1}(\cdot, \cdot; \cdot) \) is the quantile of the beta distribution.

Suppose now that \( n_1 \) trials have been conducted with probability of success \( \theta_1 \), and \( n_2 \) trials with probability of success \( \theta_2 \). We are interested in estimating of \( \theta = w_1 \theta_1 + w_2 \theta_2 \) for known \( 0 < w_1 < 1 \) and \( w_2 = 1 - w_1 \). Let \( \xi_1 \sim \text{Bin}(n_1, \theta_1) \), \( \xi_2 \sim \text{Bin}(n_2, \theta_2) \) and consider the random variable

\[ \hat{\theta}_w = w_1 \frac{\xi_1}{n_1} + w_2 \frac{\xi_2}{n_2}. \]

**Theorem.** The estimator \( \hat{\theta}_w \) is an unbiased estimator of \( \theta \).

**Proof.** Note that for a given \( \theta \) there are infinitely many \( \theta_1 \) and \( \theta_2 \) giving \( \theta \). Hence we average with respect to \( \theta_1 \) assuming the uniform distribution of \( \theta_1 \) on the interval \((0 \vee \frac{\theta - w_2}{w_1}, 1 \wedge \frac{\theta}{w_1})\). Let

\[
a_\theta = 0 \vee \frac{\theta - w_2}{w_1}, \quad b_\theta = 1 \wedge \frac{\theta}{w_1} \quad \text{and} \quad L_\theta = b_\theta - a_\theta.
\]

We have

\[
E_\theta \hat{\theta}_w = E_\theta \left( w_1 \frac{\xi_1}{n_1} + w_2 \frac{\xi_2}{n_2} \right) = \frac{1}{L_\theta} \int_{a_\theta}^{b_\theta} \left( \frac{w_1}{n_1} E_{\theta_1} \xi_1 + \frac{w_2}{n_2} E_{\theta - w_1 \theta_1} \xi_2 \right) d\theta_1 = \theta
\]

for all \( \theta \).
The variance of $\hat{\theta}_w$ equals
\[
D^2_\hat{\theta} \hat{\theta}_w = D^2_\theta \left( \frac{w_1 \xi_1}{n_1} + \frac{w_2 \xi_2}{n_2} \right)
\]
\[
= \frac{1}{L(\theta)} \int_{a(\theta)}^{b(\theta)} \left( \frac{w^2_1}{n_1^2} D^2_\theta \xi_1 + \frac{w^2_2}{n_2^2} D^2_{\theta - w_1 \xi_2} \xi_2 \right) d\theta
\]
\[
= \frac{2n_1 w_1 (b^3_\theta - a^3_\theta) + 3w_2 (n_1 (1 - 2\theta) - nw_1) (b^2_\theta - a^2_\theta)}{6n_1 (n - n_1)(b_\theta - a_\theta)}
\]
\[
+ \frac{6n_1 \theta (\theta + w - 1) (b_\theta - a_\theta)}{6n_1 (n - n_1)(b_\theta - a_\theta)}.
\]

Let $\hat{\theta}_w = u$ be observed. The (symmetric) confidence interval for $\theta$ at confidence level $\gamma$ is $(\theta^{wL}_w(u), \theta^{wU}_w(u))$, where
\[
\theta^{wL}_w(u) = \begin{cases} 0 & \text{for } u = 0, \\ \max \{ \theta : P_\theta \{ \hat{\theta}_w < u \} = (1 + \gamma)/2 \} & \text{for } u > 0, \end{cases}
\]
\[
\theta^{wU}_w(u) = \begin{cases} 1 & \text{for } u = 1, \\ \min \{ \theta : P_\theta \{ \hat{\theta}_w \leq u \} = (1 - \gamma)/2 \} & \text{for } u < 1. \end{cases}
\]

We have
\[
P_\theta \{ \hat{\theta}_w \leq u \} = P_\theta \left\{ w_1 \frac{\xi_1}{n_1} + w_2 \frac{\xi_2}{n_2} \leq u \right\}
\]
\[
= \frac{1}{L(\theta)} \int_{a(\theta)}^{b(\theta)} \sum_{i_2=0}^{n_2} P_{\theta_1} \left\{ \xi_1 \leq \frac{n_1}{w_1} \left( u - w_2 \frac{i_2}{n_2} \right) \right\} P_{\theta_2} \{ \xi_2 = i_2 \} d\theta_1
\]
and
\[
P_\theta \{ \hat{\theta}_w < u \} = P_\theta \left\{ w_1 \frac{\xi_1}{n_1} + w_2 \frac{\xi_2}{n_2} < u \right\}
\]
\[
= \frac{1}{L(\theta)} \int_{a(\theta)}^{b(\theta)} \sum_{i_2=0}^{n_2} P_{\theta_1} \left\{ \xi_1 \leq \frac{n_1}{w_1} \left( u - w_2 \frac{i_2}{n_2} \right) - 1 \right\} P_{\theta_2} \{ \xi_2 = i_2 \} d\theta_1.
\]

In the integrals above, $\theta_2 = (\theta - w_1 \xi_1)/w_2$.

For given $\theta \in (0, 1)$, the expected length of the confidence interval equals
\[
l^w(\theta) = \sum_{u \in U} (\theta^w_u(u) - \theta^{wU}_w(u)) g(u; n_1, n_2, \theta) 1_{(\theta^w_L(u), \theta^w_U(u))}(\theta),
\]
where
\[
U = \left\{ u = w_1 \frac{k_1}{n_1} + w_2 \frac{k_2}{n_2} : k_1 = 0, 1, \ldots, n_1, k_2 = 0, 1, \ldots, n_2 \right\}
\]
and
\[ g(u; n_1, n_2, \theta) = P_\theta \{ \hat{\theta}_w = u \} = P_\theta \left\{ \frac{w_1 \xi_1}{n_1} + \frac{w_2 \xi_2}{n_2} = u \right\} \]
\[ = \frac{1}{L(\theta)} \int \sum_{i_2=0}^{n_2} P_{\theta_1} \left\{ \xi_1 = \frac{n_1}{w_1} \left( u - \frac{i_2}{n_2} \right) \right\} \frac{P_{\theta-w_1\theta_1} \{ \xi_2 = i_2 \} \ d\theta_1. \]

It is interesting to compare the expected length \( l^w(\theta) \) with the expected length \( l^c(\theta) \) of the standard confidence interval:
\[ l^c(\theta) = \sum_{u=1}^{n} (\theta_{c_L}(u) - \theta_{c_U}(u)) f(u; n, \theta) 1_{(\theta_{c_L}(u), \theta_{c_U}(u))}(\theta), \]
where
\[ f(u; n, \theta) = \binom{n}{u} \theta^u (1 - \theta)^{n-u}. \]

For such comparison it is enough to compare \( F_{\hat{\theta}_c}^{-1}(1-\gamma) - F_{\hat{\theta}_w}^{-1}(1-\gamma) \) and \( F_{\hat{\theta}_c}^{-1}(\gamma) - F_{\hat{\theta}_w}^{-1}(\gamma) \), where \( F_{\hat{\theta}_c} \) and \( F_{\hat{\theta}_w} \) are the cdf’s of \( \hat{\theta}_c \) and \( \hat{\theta}_w \), respectively. Note that the distributions of \( \hat{\theta}_c \) and \( \hat{\theta}_w \) are unimodal with the same expected value. Hence, for the length comparison it is sufficient to compare the variances of \( \hat{\theta}_c \) and \( \hat{\theta}_w \).

The variances of estimators are symmetric with respect to \( \theta = 0.5 \) and take on the maximal value at that point. Hence it is enough to compare \( D_{0.5}^2 \hat{\theta}_c \) and \( D_{0.5}^2 \hat{\theta}_w \). The detailed analysis of the variance \( D_{0.5}^2 \hat{\theta}_w \) may be found in Zieliński (2016).

For \( w_1 \leq 0.5 \) we have
\[ D_{0.5}^2 \hat{\theta}_c = \frac{1}{4n} \quad \text{and} \quad D_{0.5}^2 \hat{\theta}_w = \frac{1}{4n} f(1-2w_1) + \frac{2w_1^2/3}{f(1-f)}, \]
where \( f = n_1/n \). It is easy to check that the optimal \( f \), i.e. one minimizing \( D_{0.5}^2 \hat{\theta}_w \), equals
\[ f^* = \left( 1 + \sqrt{1.5 - 3w_1 + w_1^2} \right)^{-1} \]
Values of \( f^* \) for different \( w_1 \) are given in Table 1.

<table>
<thead>
<tr>
<th>( w_1 )</th>
<th>0.10</th>
<th>0.20</th>
<th>0.30</th>
<th>0.40</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f^* )</td>
<td>0.0833</td>
<td>0.1710</td>
<td>0.2653</td>
<td>0.3710</td>
<td>0.5000</td>
</tr>
</tbody>
</table>

Some numerical results for \( n = 100, w_1 = 0.3, n_1 = 27 \) and \( \gamma = 0.95 \) are shown in Figures 1 and 2. In Figure 1 the coverage probability of the confidence interval \( (\theta_{c_L}^w, \theta_{c_U}^w) \) (solid line) as well as the coverage probability of \( (\theta_{c_L}^c(u), \theta_{c_U}^c(u)) \) (dotted line) are presented. In Figure 2 their lengths are compared.
Example. Consider the problem of interval estimation of the support for a political party. The easiest and standard way is to take a sample of size $n$, count the “yes, I will vote for that party” answers and divide that number by the sample size. But it appears that the support for the party may depend on a geographical region, sex of the voter, economic status etc. The question is whether such information may improve estimation.

Suppose we want to estimate the support for a political party (it will be referred to as party “A”) in Poland. In Poland there are more than 30000000 people who can vote (according to official statistics, in 2011 there were $N = 30762931$ voters \(^{(1)}\)). The standard way of estimation is to take a sample of size $n = 1000$ using the scheme of simple sampling without replacement. Let $\xi$ denote the number of “yes” answers. The confidence

interval for the support for the party may be calculated as

\[
B^{-1}\left(n - \xi + 1, \xi; \frac{1+\gamma}{2}\right), B^{-1}\left(n - \xi, \xi + 1; \frac{1-\gamma}{2}\right).
\]

Theoretically \(\xi\) has a hypergeometric distribution, but here the binomial approximation is applied. The consequences of such an approximation (as well as of a normal approximation) are discussed in Zieliński (2011). In what follows, for illustrative purposes the binomial approximation will be used.

Suppose in 2011 party “A” won in 27 out of 41 regions. In those regions there were 20224144 people who can vote, while in the remaining regions there were 10540517 voters. To improve estimation of the support of party “A” we divide Poland into two strata: the first one of weight \(w_1 = \frac{10540517}{30762931} = 0.342636955\) and the second one of weight \(w_2 = \frac{20222414}{30762931} = 0.657363045\). Due to the formula (*) the sample of size \(n = 1000\) is divided into two subsamples of sizes \(n_1 = 309\) and \(n_2 = 691\) from the first and the second stratum respectively.

Suppose that in the whole sample, 200 “yes” answers were obtained. The point estimate of the support equals 0.2 and the Clopper–Pearson confidence interval (at confidence level 0.95) is (0.17562, 0.22616); its length equals 0.05054. If in the sample of size 309 from the first stratum there were 25 “yes” answers and in the sample of size 691 from the second stratum there were 175 “yes” answers, then the confidence interval for the support is (0.17046, 0.21813) and its length is 0.04767. Note that the stratified confidence interval is shorter than the confidence interval based on the non-stratified sample. Table 2 gives other stratified confidence intervals for other possible results of the pool, assuming that the overall number of positive answers is 200.

<table>
<thead>
<tr>
<th>(\xi_1)</th>
<th>(\xi_2)</th>
<th>(left, right)</th>
<th>length</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>175</td>
<td>(0.17046, 0.21813)</td>
<td>0.04767</td>
</tr>
<tr>
<td>50</td>
<td>150</td>
<td>(0.17424, 0.22219)</td>
<td>0.04795</td>
</tr>
<tr>
<td>75</td>
<td>125</td>
<td>(0.17803, 0.22625)</td>
<td>0.04822</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>(0.18182, 0.23030)</td>
<td>0.04848</td>
</tr>
<tr>
<td>125</td>
<td>75</td>
<td>(0.18562, 0.23435)</td>
<td>0.04873</td>
</tr>
<tr>
<td>150</td>
<td>50</td>
<td>(0.18942, 0.23840)</td>
<td>0.04898</td>
</tr>
<tr>
<td>175</td>
<td>25</td>
<td>(0.19321, 0.24243)</td>
<td>0.04921</td>
</tr>
</tbody>
</table>

In Tables 3 and 4 similar results are shown for the overall number of “yes” answers equal to 300 and 400 respectively.

The confidence intervals based on the stratified sample are significantly shorter than the Clopper–Pearson confidence intervals. The last columns of
Tables 2, 3 and 4 give the ratio of the length of the stratified confidence interval to the length of the non-stratified one.

**Concluding remarks.** The confidence interval for the probability of success using information on the non-homogeneity of the sample is better, i.e. shorter, than the standard Clopper–Pearson confidence interval. Unfortunately, closed formulae for such confidence intervals are not available. Nevertheless, for given $w_1$, $n$, $n_1$ and observed $\xi_1$ and $\xi_2$ the confidence interval may be easily obtained with standard mathematical software (for example Mathematica, MathLab etc.).

Comparison with other confidence intervals (see Decrouez & Robinson, 2012) does not make sense, because those confidence intervals do not keep the prescribed confidence level.
The results of this paper may be generalized to the case of arbitrary \( w_1 \) and \( w_2 \), for example \( w_1 = 1 = -w_2 \), i.e. the difference of the probabilities of success. Work on this subject is in progress.

References

- C. J. Clopper and E. S. Pearson (1934), *The use of confidence or fiducial limits illustrated in the case of the binomial*, Biometrika 26, 404–413.

Wojciech Zieliński
Department of Econometrics and Statistics
Warsaw University of Life Sciences
Nowoursynowska 159
02-787 Warszawa, Poland
E-mail: wojciech.zielinski@sggw.pl
http://wojtek.zielinski.statystyka.info