

Confidence Intervals
for a Binomial Proportion

A. Boomsma

Department of Statistics & Measurement Theory
University of Groningen

December 16, 2005

`confbin.tex`

Confidence intervals

The concept and computation of confidence intervals

A *confidence interval* for a population parameter consists of a range of values, restricted by a lower and an upper limit. For all options the size of the interval depends, among other things, on the *sample size*(s) and on the so-called *confidence coefficient* $1 - \alpha$. It was Jerzy Neyman (1934, p. 562) who proposed the terms confidence interval and confidence coefficient, but Pierre Simon Laplace (1749–1827) had introduced confidence interval procedures already in 1814.

By definition, a confidence interval $\theta_l \leq \theta \leq \theta_u$ for an unknown parameter θ , with *unreliability* α , entails all values θ_0 for which the null hypothesis $H_0 : \theta = \theta_0$ would not have been rejected in the observed sample when a two-sided test with unreliability α (i.e., the Type I error) would have been applied. Any value θ_0 smaller than the lower bound θ_l in the sample at hand is ‘improbably small’, and any $\theta_0 > \theta_u$ is ‘improbably large’ (cf. Van den Ende & Verhoef, 1973, p. 239f.). Given some best estimate of θ in a given sample, two numbers θ_l and θ_u have to be calculated that meet the required property.

In addition, the interpretation of a confidence interval has to be understood in a frequentist sense, i.e., in a framework of repeatedly taking samples of the same size n from the same population distribution, and calculating an interval with confidence coefficient $1 - \alpha$ for some unknown but fixed parameter of interest in each of these samples. Confidence intervals are then constructed in such a way that in the long run, the proportion of intervals covering the fixed population parameter equals $1 - \alpha$. Therefore, in this framework of repeated sampling under the same conditions, a probability statement can be made, saying that the probability that the stochastic interval will cover the unknown fixed population value equals $1 - \alpha$. This probability statement is about intervals (sample statistics), *not* about the population parameter, which remains fixed when different samples are taken.

The lower and upper bounds of a two-sided confidence interval are random (they may change from sample to sample). In a given sample, however, they

are known numbers. On the other hand, the population parameter θ is a fixed but unknown number. It is this contraposition ‘stochastic but unknown’ versus ‘fixed but unknown’ that makes the interpretation of a confidence interval so difficult, because of the mind’s tendency to think that the unknown quantity θ has a probability distribution. But as long as the concept of probability refers to the frequentist point of view — what happens if the sampling experiment is repeated — that is incorrect thinking. Only in Bayesian statistics, not in classical statistics, a parameter can have a probability distribution.

The idea of constructing confidence intervals in the framework of repeated sampling is illustrated in Figure 1. It is the probabilistic game of throwing sticks of different lengths at the fixed target θ . The intervals (the random lengths of the sticks) are constructed in such a way that, in the long run, the target is missed in $100\alpha\%$ of the cases.

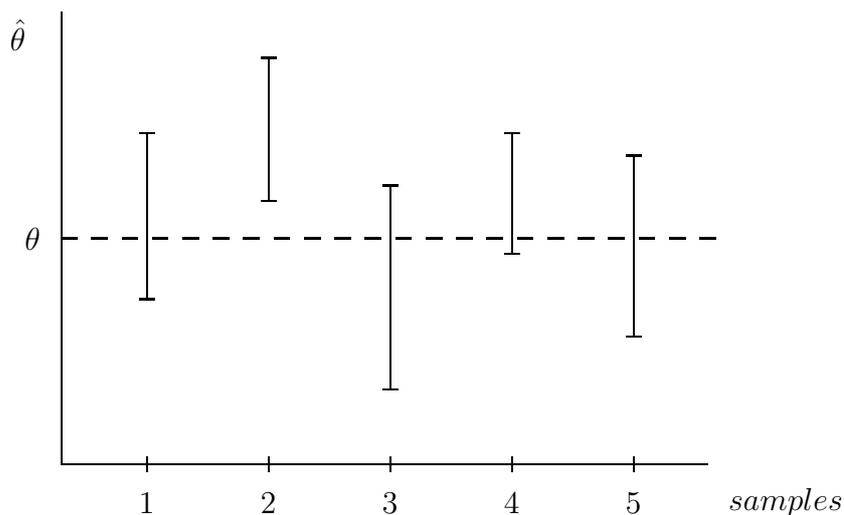


Figure 1. Confidence intervals for a fixed population parameter θ in five samples from the same population.

If the interest is in a lower bound θ_{2l} with reliability $1 - \alpha$, it is chosen in such a way that for each number $\theta_0 \geq \theta_{2l}$ the null hypothesis $H_0 : \theta = \theta_0$ in a one-sided test at a significance level α against the alternative $H_1 : \theta < \theta_0$ is not rejected; values smaller than θ_{2l} are for the given sample of observations very unlikely. One-sided upper limits θ_{2u} are defined by switching the inequality signs.

Along with the p -value of a test statistic – reflecting significance at a quantitative level – the confidence interval gives sufficient information to convey significances as dichotomous decisions, as well as the practical relevance of the sample results. Any value outside the interval is rejected as a null hypothesis (*significance*), and of course the parameter estimate itself (*relevance*) is always reported along with the safety limits of the interval (cf. Van den Ende & Verhoef, 1973, Section 8.8).

If a 95% confidence interval for a population parameter has been computed by the computer program CONFIN (Boomsma & Popping, 1977), and if the sample size was n , what is the meaning of the resulting confidence interval?

For all options in CONFIN, *two-sided* confidence intervals are computed with a confidence coefficient $1 - \alpha$. If a *one-sided* interval is needed, the value of α must be adjusted properly. For example, if a one-sided 99% confidence interval has to be computed, the confidence coefficient must have the input value of 0.98. The users themselves can adjust the fixed lower or upper limit of a one-sided interval. For example, for a correlation coefficient -1.0 for a right-sided interval and 1.0 for a left-sided interval).

Confidence intervals for a binomial proportion

Assumptions

1. n independent Bernoulli trials with constant success probability π .
2. The measurement scale of random variable X is at least nominal.

Description

Let x be the number of successes in a random sample of size n . A success is observed if X_i , $i = 1, 2, \dots, n$, has a specific characteristic; a failure is observed if X_i does not have that characteristic. The proportion of successes in the sample is denoted as $\hat{\pi} = x/n$, and the proportion in the population as π .

Computation

Four types of confidence intervals can be distinguished: Wilson's score interval (Wilson, 1927), the Wald interval (Wald & Walfowitz, 1939), the adjusted Wald interval (Agresti & Coull, 1998), and the 'exact' Clopper-Pearson interval (Clopper & Pearson, 1934). Each of these intervals are defined now, starting with the latter.

a. The Clopper-Pearson interval

To avoid normal theory approximations some textbooks recommend the Clopper-Pearson ‘exact’ confidence interval for π (Clopper & Pearson, 1934). This interval is based on inverting equal-tailed binomial tests of the null hypothesis $H_0 : \pi = \pi_0$ against the alternative hypothesis $H_1 : \pi \neq \pi_0$. Although this interval is often depicted as an *exact* interval (e.g., Sachs, 1974, p. 258), we quote Neyman (1935) on this matter: “exact probability statements are impossible in the case of the binomial distribution” (p. 111). The reason is that the observed number of successes x is always an integer, and not real-valued. In the general case of discontinuous distributions, exact probability statements regarding the problem of confidence intervals are impossible. Although Neyman claimed that “the system of Clopper and Pearson could not be bettered” (p. 111), later it was shown that improved confidence intervals are possible; see, for example, Agresti and Coull (1998), Agresti and Caffo (2000) and Brown, Cai and DasGupta (2001, 2002). Notice, however, that Neyman was probably referring to the lower bound of the Pearson-Clopper interval: its confidence level would always be at least $1 - \alpha$.

The program CONFIN (Boomsma & Popping, 1977) computes π_l and π_u of the interval $\pi_l \leq \pi \leq \pi_u$, with a confidence coefficient of at least $1 - \alpha$.

For different values of x , the lower and upper bounds of the interval are defined as follows.

For $0 < x < n$, the solution for π_l and π_u is found via quantiles of the $F_{\nu_2}^{\nu_1}$ -distribution. That is,

$$\pi_l = \frac{x}{x + (n - x + 1)F} \quad , \quad (1)$$

where F is the $F_{\nu_2, 1-\alpha/2}^{\nu_1}$ quantile with $\nu_1 = 2n - 2x + 2$ and $\nu_2 = 2x$ degrees of freedom, and

$$\pi_u = \frac{(x + 1)F}{n - x + (x + 1)F} \quad , \quad (2)$$

where F is the $F_{\nu_2, 1-\alpha/2}^{\nu_1}$ quantile with $\nu_1 = 2x + 2$ and $\nu_2 = 2n - 2x$ degrees of freedom.

For $x = 0$, the lower limit $\pi_l = 0$, because the upper limit π_u satisfies the equality $(1 - \pi_u)^n = \alpha/2$, from which it follows that $\pi_u = 1 - (\alpha/2)^{1/n}$.

For $x = n$, the upper limit $\pi_u = 1$, because the lower limit satisfies $\pi_l^n = \alpha/2$, which makes $\pi_l = (\alpha/2)^{1/n}$.

b. The Wald interval

The *normal theory approximation* of a confidence interval for a proportion is known as the Wald interval, defined as $\hat{\pi} \pm z\sqrt{\hat{\pi}(1 - \hat{\pi})/n}$, where z is the $z_{1-\alpha/2}$ quantile of the standard normal distribution (Wald & Wolfowitz, 1939). However, it has been known for some time that the Wald interval performs poorly, unless n is quite large; see, for example, Ghosh (1979) and Blyth and Still (1983). The interval procedure is *conservative* due to the discreteness of the binomial distribution; conservative means that the empirical value of the confidence coefficient is larger than the nominal level $1 - \alpha$.

In Hogg and Tanis (2001) this interval is discussed in Section 7.5, and defined by Equation 7.5-2.

c. The adjusted Wald interval

Agresti and Coull (1998) introduced a slight modification of the Wald interval by adding two successes and two failures. Thus the point estimator

$$\hat{\pi}_w = \frac{x + 2}{n + 4} \tag{3}$$

is used, which Agresti and Coull call the *Wilson point estimator* of π because it was Wilson (1927, p. 211) who first referred to the usefulness of this estimator.

The variance of the estimated proportion $\hat{\pi}_w$ is defined as

$$s_w^2 = \frac{\hat{\pi}_w(1 - \hat{\pi}_w)}{n + 4} \tag{4}$$

The program CONFIN computes π_l and π_u of the interval $\pi_l \leq \pi \leq \pi_u$, with an approximate confidence coefficient of $1 - \alpha$.

The lower and upper bounds of the adjusted Wald interval are defined as

$$\pi_l = \hat{\pi}_w - z s_w \tag{5}$$

and

$$\pi_u = \hat{\pi}_w + z s_w \quad , \quad (6)$$

where π_w and s_w^2 are defined by (3) and (4), respectively, and z is the $z_{1-\alpha/2}$ quantile of the standard normal distribution.

d. The Wilson score interval

The program CONFIN computes π_l and π_u of the interval $\pi_l \leq \pi \leq \pi_u$, with an approximate confidence coefficient of $1 - \alpha$.

The lower and upper bounds of the interval are defined as

$$\pi_l = \frac{\hat{\pi} + \frac{z^2}{2n} - z \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \quad , \quad (7)$$

and

$$\pi_u = \frac{\hat{\pi} + \frac{z^2}{2n} + z \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \quad , \quad (8)$$

where z is the $z_{1-\alpha/2}$ quantile of the standard normal distribution.

In Hogg and Tanis (2001) this interval is nicely derived in Section 7.5, and defined by Equation 7.5-4.

To clarify on the term score test, it should be noted that the Wilson's score confidence interval is the inversion of what is called the score test for π . Whereas Wald tests are based on the log-likelihood at the maximum likelihood estimate ($\hat{\pi}$ in our example), score tests are based on the log-likelihood at the null-hypothesis value of the parameter (in our case at π_0 under $H_0 : \pi = \pi_0$). See Buse (1982) for a further discussion of the different types of statistical tests.

Which interval to choose?

The results of Agresti and Coull (1998) and Agresti and Caffo (2000) indicate

that there is much in favor of the adjusted Wald and the Wilson score intervals, relative to the ‘exact’ Clopper-Pearson and the Wald intervals. “The Clopper-Pearson interval has coverage probabilities bounded below by the nominal confidence level, but the typical coverage probability is much higher than that level. The score and adjusted Wald intervals can have coverage probabilities lower than the nominal confidence level, yet the typical coverage probability is close to that level” (Agresti & Coull, 1998, p. 125). Agresti and Caffo (2000) re-emphasize that the poor performance of the Wald interval is unfortunate, since it is the simplest approach to present in elementary statistics courses. They strongly recommend that instructors present the score interval instead (p. 122). Hogg & Tanis (2001, p. 380f.) illustrate how to do so in an insightful manner.

References

- Agresti, A., & Coull, B.A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, **52**, 119–126.
- Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, **54**, 280–288.
- Blyth, C.R., & Still, H.A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association*, **78**, 108–116.
- Boomsma, A., & Popping, R. (1977). *CONFIN: A description of an interactive program for the calculation of confidence intervals for certain population parameters* (Heymans Bulletin, HB-77-261-RP). University of Groningen, Department of Statistics & Measurement Theory. [A new version of the program is in progress, and will be available shortly.]
- Brown, L.D., Cai, T.T., & DasGupta, A. (2001). Confidence intervals for a binomial proportion (with discussion). *Statistical Science*, **16**, 101–133.
- Brown, L.D., Cai, T.T., & DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, **30**, 160–201.

-
- Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, **36**, 153–157.
- Clopper, C.J., & Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.
- Ghosh, B.K. (1979). A comparison of some approximate confidence intervals for the binomial parameter. *Journal of the American Statistical Association*, **74**, 894–900.
- Hogg, R.V., & Tanis, E.A. (2001). *Probability and statistical inference* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Laplace, P.S. (1814). *Théorie analytique des probabilités* (2nd ed.). Paris: Courcier.
- Neyman, J. (1934). On two different aspects of the representative method: The method of stratified sampling and the method of purposive sampling. *Journal of the Royal Statistical Society*, **97**, 558–625.
- Neyman, J. (1935). On the problem of confidence limits. *The Annals of Mathematical Statistics*, **6**, 111–116.
- Sachs, L. (1974). *Angewandte Statistik. Planung und Auswertung Methoden und Modelle* (4. Auflage). Berlin: Springer.
- Van den Ende, H., & Verhoef, M. (1973). *Inductieve statistiek voor gedragswetenschappen. Een kritische inleiding*. Amsterdam: Agon Elsevier.
- Wald, A., & Wolfowitz, J. (1939). Confidence limits for continuous distribution functions. *The Annals of Mathematical Statistics*, **10**, 105–118.
- Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209–212.

Confidence intervals for a binomial proportion

Copyright © 2006 by Anne Boomsma, Department of Statistics & Measurement Theory, University of Groningen, The Netherlands

Alle rechten voorbehouden. Niets in deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, en/of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotocopie, microfilm of op enige andere manier, zonder voorafgaande schriftelijke toestemming van de auteur.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the author.