Hochschule Niederrhein
University of Applied Sciences

Elektrotechnik
und Informatik
Faculty of Electrical Engineering
and Computer Science

Construction of Confidence Intervals

Christoph Dalitz

Technischer Bericht Nr. 2017-01

IMPRESSUM

Technische Berichte des Fachbereichs Elektrotechnik und Informatik,
Hochschule Niederrhein

ISSN 2199-031X

HERAUSGEBER

Christoph Dalitz und Steffen Goebbels
Fachbereich Elektrotechnik und Informatik

ANSCHRIFT

Hochschule Niederrhein
Reinarzstr. 49
47805 Krefeld

http://www.hsnr.de/fb03/technische-berichte/

C. Dalitz: „Construction of confidence intervals.“ Technical Report No. 2017-01, pp. 15-28,
Hochschule Niederrhein, Fachbereich Elektrotechnik und Informatik, 2017

Technical Report No. 2017-01, pp. 15-28, Hochschule Niederrhein, Fachbereich Elektrotechnik & Informatik (2017)

Construction of Confidence Intervals*

Christoph Dalitz
Institute for Pattern Recognition
Niederrhein University of Applied Sciences
Reinarzstr. 49, 47805 Krefeld, Germany
christoph.dalitz@hsnr.de

Abstract

Introductory texts on statistics typically only cover the classical "two sigma" confidence interval for the mean value and do not describe methods to obtain confidence intervals for other estimators. The present technical report fills this gap by first defining different methods for the construction of confidence intervals, and then by their application to a binomial proportion, the mean value, and to arbitrary estimators. Beside the frequentist approach, the likelihood ratio and the highest posterior density approach are explained. Two methods to estimate the variance of general maximum likelihood estimators are described (Hessian, Jackknife), and for arbitrary estimators the bootstrap is suggested. For three examples, the different methods are evaluated by means of Monte Carlo simulations with respect to their coverage probability and interval length. R code is given for all methods, and the practitioner obtains a guideline which method should be used in which cases.

1 Introduction

When an unknown model parameter is estimated from experimental data, the estimation always yields a value, be the sample size large or small. We would, however, expect a more accurate value from a larger sample. A *confidence interval* measures this "accuracy" in some way. As "accuracy" can be defined in different ways, there are different approaches to the construction of confidence intervals.

The most common approach is the *frequentist approach*, which is based on the *coverage probability* and is taught in introductory texts on statistics [1]. It assumes the unknown parameter to be known and then chooses an interval around the estimator that includes the parameter with a given probability (typically 95%). The *evidence based approach* utilizes the *likelihood ratio* and chooses an interval wherein the likelihood function is greater than a given threshold (typically $1/8$ of its maximum value) [2]. The *Bayesian approach* treats the unknown parameter as a random variable and estimates its distribution from the observation. This leads to the *highest posterior density* interval [3].

Both for binomial proportions and for mean values, simple formulas or algorithms to compute confidence intervals can be given. A possible evaluation criteria for the obtained intervals is the coverage probability. One should think that this criterion favors the frequentist approach, but even for this approach, the coverage probability may vary considerably, depending on the true parameter value. For non-symmetric intervals, another evaluation criterion is the interval length because, from two intervals with the same coverage probability, the shorter one is preferable.

Beyond the binomial proportion and the mean value, there is no standard formula for computing a confidence interval. For maximum likelihood estimators, it is however known that they are asymptotically normal, provided the likelihood function is sufficiently smooth [4]. In these cases, the confidence interval for the mean value can be used. This requires an estimation of the estimator variance, which can be done in two ways: the diagonal elements of the inverted *Hessian matrix* of the log-likelihood function, or the *Jackknife* variance.

For non-smooth likelihood functions or for arbitrary estimators, only the *bootstrap* method is universally applicable. This method generates new data from the observations by random sampling with replacement and estimates the confidence interval from the sampled data. In principle, the bootstrap method is always applicable, even in cases when the other methods work, but in the experiments described in this report, the bootstrap method had a poorer coverage probability than the classic confidence interval, and it should therefore only be used when other methods cannot be applied.

*Für eine deutsche Version dieses Artikels siehe pp. 1-14 dieses Berichts.

This report is organized as follows: section 2 defines the basic terms estimator, coverage probability, likelihood ratio, and posterior density. In sections 3 and 4, the different approaches are applied to the binomial proportion and to the mean value. Sections 5 and 6 describe construction methods for confidence intervals for maximum likelihood estimators and for arbitrary estimators. Section 7 presents Monte Carlo experiments that evaluate the coverage probability of the different confidence intervals. The final section makes recommendations which confidence interval should be used in which case.

2 Basic terms

The probability distribution of a random variable X be known except for the value of some parameter θ. In other words: the shape of the probability density $f_\theta(x)$ be known, but not the value of the parameters θ. In the most general case, θ is a vector and represents several parameter values. If X is normal distributed, for instance, then θ represents two parameters: $\theta = (\mu, \sigma^2)$. An *estimator* is a function to estimate the unknown parameter from independent observations $x_1, \ldots, x_n$ of the random variable X. The particular estimated value is denoted with $\hat{\theta}$:

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) \tag{1}$$

Simple examples are the relative frequency as an estimator for a binomial proportion, or the statistical average as an estimator for the parameter μ of the normal distribution.

2.1 Maximum likelihood (ML)

The *maximum likelihood principle* is a general method to obtain estimators [4]. It chooses the parameter θ in such a way that the *likelihood function* L or[1] the *log-likelihood function* ℓ is maximized:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) \tag{2a}$$

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i) \tag{2b}$$

Loosely speaking, $L(\theta)$ is a measure for the probability of the observation $x_1, \ldots, x_n$ under the assumption that the true parameter value is θ. If $\theta = (\theta_1, \ldots, \theta_t)$ and $\ell(\theta)$ is differentiable, the maximum likelihood principle yields t equations for the determination of the t parameters $\theta_1, \ldots, \theta_t$:

$$\frac{\partial}{\partial \theta_i} \ell(\theta) = 0 \quad \text{for } i = 1 \ldots, t \tag{3}$$

Maximum likelihood estimators have a number of attractive properties like asymptotic normality under quite general conditions. This will play a role in section 5. In many cases, the equations (3) cannot be solved in closed form, thereby making a numerical maximization of the log-likelihood function necessary. If this is not possible, one might try other methods that possibly yield estimators in a simpler way, like the *method of moments* or its generalization [5].

2.2 Coverage probability

An estimation function (1) yields only a single value and is therefore called a *point estimator*. A *confidence interval*, on the contrary, gives a region $[\theta_l, \theta_u]$ wherein the parameter falls with high probability. The boundaries $\theta_{l,u}$ of the interval depend on the observed data $x_1, \ldots, x_n$ and are thus random variables. The *frequentist approach* is based on the following consideration: if θ is the true parameter value, then it ideally should fall into the confidence interval with a predefined *coverage probability* $(1 - \alpha)$:

$$P_{cov}(\theta) = P(\theta \in [\theta_l, \theta_u]) = 1 - \alpha \tag{4}$$

Unfortunately, Eq. (4) cannot be used to determine θ_l and θ_u, because the unknown θ is part of the equation. This dilemma can be resolved when the problem is reinterpreted as a hypothesis testing problem: under the hypothesis $\theta \notin [\theta_l, \theta_u]$, the probability that the estimator deviates from θ more than the observed value $\hat{\theta}$ is less than α. Or, in hypothesis testing lingo: if θ were one of the interval boundaries, then everything beyond $\hat{\theta}$ would fall into the rejection region. When the probability α is distributed evenly among small and large deviations, the formal definition of the *frequentist confidence interval* becomes[2]:

$$P_{\theta=\theta_l}(\hat{\theta} \geq \theta_0) = \alpha/2 \quad \text{and} \tag{5a}$$

[1] Note that $L(\theta)$ and $\log L(\theta)$ have their maximum at the same argument, because the logarithm is a monotonic function.

[2] This definition reads slightly different from the definition given by DiCiccio & Efron [6]: Eq. (5b) is identical, but in Eq. (5a) they write "$>$" instead of "$\geq$". This makes no difference for continuous random variables, but it would treat the two boundaries differently for discrete random variables.

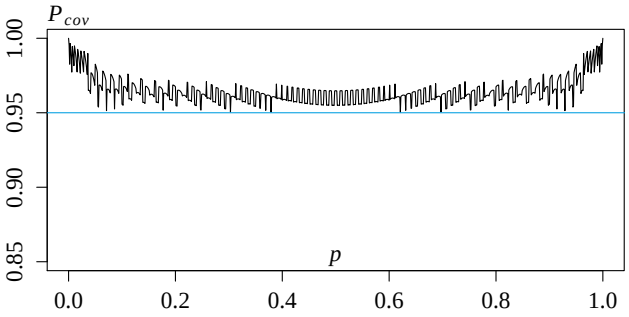


Figure 1: Coverage probability P_{cov} of the "exact" confidence interval for a binomial proportion after Eq. (5) as a function of the true parameter p for $n = 100$ and $\alpha = 0.05$.

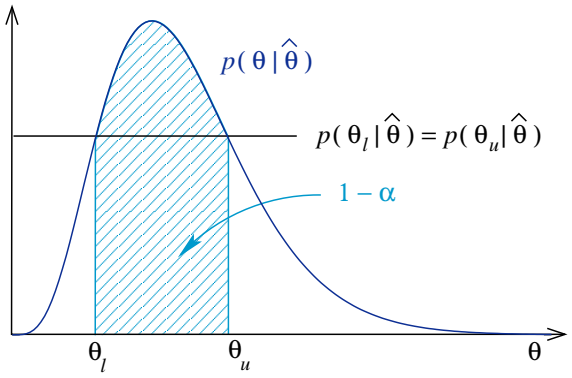


Figure 2: Determination of the highest posterior density interval $[\theta_l, \theta_u]$ according to Eq. (8).

$$P_{\theta=\theta_u}(\hat{\theta} \leq \theta_0) \quad = \quad \alpha/2 \tag{5b}$$

where θ_0 is the observed value for the estimator and $P_{\theta=\theta_{l,u}}$ is the probability under the assumption that the true parameter value is the lower or upper boundary, respectively.

Although the confidence interval obtained by solving Eq. (5) for θ_l and θ_u is guaranteed to have have at least $1-\alpha$ coverage probability independent from θ, there are two hitches: the example in Fig. 1 shows that even an "exact" confidence interval directly computed with Eq. (5) can have coverage probability that is too large for most values of θ, which means that the interval is too wide. Moreover, the probability is often known only approximately, or Eq. (5) can only be solved asymptotically, which leads to an approximate confidence interval, which can have $P_{cov}(\theta)$ less than $1-\alpha$.

2.3 Likelihood ratio

A different approach to obtain a confidence interval is based on the likelihood function (2a). The ML estimator $\hat{\theta}$ chooses θ such that it maximizes the probability of the observed data. However, other values of θ lead to a high probability of the observation, too. It is thus natural to define an interval wherein the ratio $L(\hat{\theta})/L(\theta)$ is greater than some threshold. To distinguish this interval from the frequentist confidence interval, it is called the *likelihood ratio support interval* $[\theta_l, \theta_u]$:

$$\frac{L(\theta)}{L(\hat{\theta})} \geq \frac{1}{K} \quad \text{for all } \theta \in [\theta_l, \theta_u] \tag{6}$$

where $\hat{\theta}$ is the ML estimator for θ. A common choice for K is $K = 8$ because, in the case of mean values, it leads to intervals very close to the frequentist interval for $\alpha = 0.05$ (see section 4.2).

2.4 Posterior density

A third approach to confidence interval construction tries to estimate a probability density for θ on basis of the observation $\hat{\theta}$. The true parameter θ is here considered as a random variable, and $p_\theta(\hat{\theta})$ is a conditional probability density[3] $p(\hat{\theta}|\theta)$ that can be computed with Bayes' formula:

$$p(\theta|\hat{\theta}) = \frac{p(\hat{\theta}|\theta) \cdot p(\theta)}{\int_{\mathbb{R}} p(\hat{\theta}|\tau) \cdot p(\tau)\, d\tau} \tag{7}$$

Based on this density, the *highest posterior density (HPD) interval* is defined as the region $[\theta_l, \theta_u]$ with highest probability density values and a total probability of $(1-\alpha)$. Formally, this definition leads to the coupled equations (see Fig. 2)

$$1-\alpha \quad = \quad \int_{\theta_l}^{\theta_u} p(\theta|\hat{\theta})\, d\theta \qquad \text{and} \tag{8a}$$

$$p(\theta_l|\hat{\theta}) \quad = \quad p(\theta_u|\hat{\theta}) \tag{8b}$$

Apart from the nuisance that this system of equations can only be solved numerically, the HPD interval has a fundamental deficiency: to compute $p(\theta|\hat{\theta})$ with Eq.. (7), it is necessary to make an assumption about the "a priori distribution" $p(\theta)$ of the unknown parameter θ, and this assumption is arbitrary. Typically, $p(\theta)$ is chosen to be constant which implies that nothing is known about the approximate location of θ.

[3] Note that θ and $\hat{\theta}$ are continuous variables, so that their probability distribution is described by a density, here denoted with the lower case letter p.

Although this assumption is rarely realistic in practical situations, this does not necessarily mean that the HPD interval is bad. As we will see in the next section, it can even have a good coverage probability.

3 Relative frequencies

The relative frequency $\hat{p}$ is a ubiquitous estimator for a probability, or a binomial proportion p. The probability distribution of $\hat{p}$ is exactly given by the binomial distribution. When an event has probability p, the probability that it occurs k times in n independent trials is

$$P_p(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{9}$$

The relative frequency $\hat{p} = k/n$ then has the probability

$$P_p(\hat{p} = p_0) = \binom{n}{np_0} p^{np_0} (1-p)^{n(1-p_0)} \tag{10}$$

Eq. (10) is the starting point for all confidence intervals of the relative frequency.

3.1 Frequentist interval for $\hat{p}$

Insertion of (10) into Eq. (5) yields the following equations to determine boundaries p_l and p_u:

$$1 - \text{pbinom}\left((k-1)/n, n, p_l\right) = \alpha/2 \tag{11a}$$

$$\text{and} \qquad \text{pbinom}\left(k/n, n, p_u\right) = \alpha/2 \tag{11b}$$

where $k/n = \hat{p}$ is the observed relative frequency, and *pbinom* is the R function for the cumulative distribution function (CDF) of the binomial distribution. In the special cases $k = 0$ or $k = 0$, one of the equations (11) does not have a solution because p_l and p_u are restricted to the interval $[0, 1]$. In these cases, let $p_l = 0$ ($k = 0$) or $p_u = 1$ ($k = n$), respectively. The other boundary can be found analytically in these cases as

$$k = 0 \quad \Rightarrow \quad [p_l, p_u] = [0, 1 - \sqrt[n]{\alpha/2}] \tag{12a}$$

$$k = n \quad \Rightarrow \quad [p_l, p_u] = [\sqrt[n]{\alpha/2}, 1] \tag{12b}$$

In all other cases, Eq. (11) must be solved numerically, e.g., with the R function *uniroot*[4]. The corresponding R code is given in listing 1. This interval is known as the *Clopper-Pearson* interval [8], which is also implemented in the R function *binom.confint* from the package *binom*, with the option *method='exact'*.

```
ci.binom <- function(n, k, alpha) {
  if (k == 0) {
    p1 <- 0.0
    p2 <- 1 - (alpha/2)**(1/n)
  }
  else if (k == n) {
    p1 <- (alpha/2)**(1/n)
    p2 <- 1.0
  }
  else {
    helper <- function(p, k, n, val) {
      return (pbinom(k, n, p) - val)
    }
    r <- uniroot(helper, k=(k-1),
                 n=n, val=1-alpha/2,
                 interval=c(0,1))
    p1 <- r$root
    r <- uniroot(helper, k=k,
                 n=n, val=alpha/2,
                 interval=c(0,1))
    p2 <- r$root
  }
  return (data.frame(p1=p1, p2=p2))
}
```

Listing 1: R implementation of the exact Clopper-Pearson confidence interval for the relative frequency after Eqs. (11) & (12).

An approximate confidence interval is obtained when the binomial distribution is replaced by the normal distribution, which is justified by the central limit theorem. For large n, $\hat{p}$ is approximately normally distributed with $\mu = p$ and $\sigma^2 = p(1-p)/n$. With this approximation, Eq. (5a) becomes

$$1 - \text{pnorm}\left(\hat{p}, p_l, \sqrt{p_l(1-p_l)/n}\right) = \alpha/2$$

$$\Leftrightarrow \quad \text{pnorm}\left(\frac{\hat{p} - p_l}{\sqrt{p_l(1-p_l)/n}}, 0, 1\right) = 1 - \alpha/2$$

$$\Leftrightarrow \quad \frac{\hat{p} - p_l}{\sqrt{p_l(1-p_l)/n}} = z_{1-\alpha/2} \tag{13}$$

where *pnorm* is the R function for the CDF of the normal distribution, and $z_{1-\alpha/2} = \text{qnorm}(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. The quadratic equation (13) and its analogous version for p_u can be solved analytically, thereby yielding the

[4] It would also be possible to use the inverse of the incomplete beta function, because 1-*pbinom* can be expressed through this function (see [7] Eq. 26.5.7). The inverse of the incomplete beta function, however, must be computed numerically either.

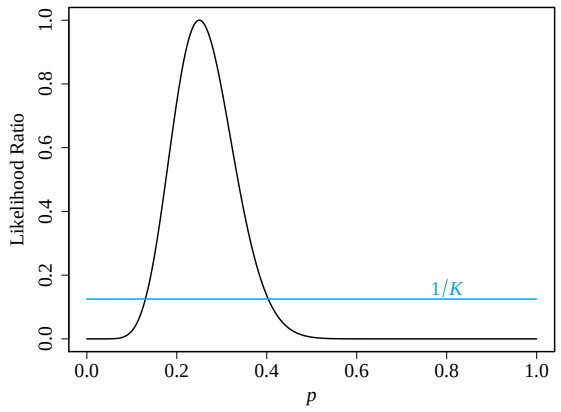


Figure 3: Likelihood ratio $L(p)/L(\hat{p})$ of the binomial distribution for $n = 40$ and $k = 10$.

Wilson interval:

$$\frac{1}{1+z^2/n}\left[\hat{p}+\frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2}{4n^2}}\right] \quad (14)$$

where $z = z_{1-\alpha/2}$, for the sake of brevity. In the comparative study [9], Brown et al. recommended the Wilson interval due to its coverage probability. For large n, Eq. (14) asymptotically transforms into the classical *Wald interval* that is taught in introductory text books:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \quad (15)$$

3.2 Likelihood ratio for $\hat{p}$

When the event of interest occurs k times in n trials, the likelihood function is

$$L(p) = p^k(1-p)^{n-k} \quad (16)$$

The relative frequency $\hat{p} = k/n$ is the ML estimator for p. The likelihood ratio support interval therefore encompasses all p with

$$\frac{L(p)}{L(\hat{p})} = \frac{p^k(1-p)^{n-k}}{\hat{p}^k(1-\hat{p})^{n-k}} \geq \frac{1}{K} \quad (17)$$

A plot of the function on the left hand side is shown in Fig. 3. Eq. (17) must be solved numerically, e.g., with the R function *uniroot*. A possible implementation is given in listing 2.

```
lr.binom <- function(n, k, K) {
  helper <- function(p, n, k, K) {
    return (p**k * (1-p)**(n-k) /
            ((k/n)**k * (1-k/n)**(n-k))
            - 1/K)
  }
  p1 <- rep(0,length(k))
  p2 <- p1
  if (k==0) {
    p1 <- 0
  } else {
    r <- uniroot(helper,n=n,k=k,K=K,
                 interval=c(0,k/n))
    p1 <- r$root
  }
  if (k==n) {
    p2 <- 1
  } else {
    r <- uniroot(helper,n=n,k=k,K=K,
                 interval=c(k/n,1))
    p2 <- r$root
  }
  return (data.frame(p1=p1, p2=p2))
}
```

Listing 2: R code that computes the likelihood ratio support interval for the relative frequency according to Eq. (17).

3.3 Highest posterior density for $\hat{p}$

The R package *HDInterval* provides the function *hdi* for computation of HPD intervals. *hdi* requires as one function argument a function that computes the inverse of $\int_{-\infty}^{\theta} p(\tau|\hat{\theta})\, d\tau$. This means that *hdi* is only applicable in cases where this inverse function can be readily computed. The binomial distribution is such a case.

Insertion of the binomial distribution (10) into Eq. (7) yields with the assumption of a constant "a priori" density $p(\theta) = const.$:

$$\begin{aligned} p(p|k) &= \frac{\binom{n}{k}p^k(1-p)^{n-k}}{\int_0^1 \binom{n}{k}q^k(1-q)^{n-k}\,dq} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,p^{a-1}(1-p)^{b-1} \\ &= \text{dbeta}(p,a,b) \end{aligned} \quad (18)$$

where $a = k+1$ and $b = n-k+1$, and *dbeta* is the R function for the probability density of the beta distribution. The inverse CDF of the beta distribution in provided by R as the function *qbeta*, so that the HPD interval can be computed with the code in listing 3.

```
library(HDInterval)
ci <- hdi(qbeta, 1-alpha,
          shape1=(k+1),
          shape2=(n-k+1))
p1 <- ci[1]; p2 <- ci[2]
```

Listing 3: R code that computes the $(1-\alpha)$ HPD interval for the relative frequency.

4 Mean values

Another ubiquitous estimator is the statistical average $\overline{x}$ as an estimator for the expectation value $\mu = E(X)$. For the statistical average $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, it is possible to construct a quantity that only depends on the unknown μ and has a known distribution, albeit only in the special case that the variable X is normally distributed. In this case, the random variable

$$Z = \frac{\overline{x} - \mu}{\sqrt{s^2/n}} \quad \text{with } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad (19)$$

is t distributed with $(n-1)$ degrees of freedom[5]. If X is not normally distributed, it is at least known from the central limit theorem that the quantity (19) is approximately standard normally distributed[6] [1]. In general, it is not known whether X is normally distributed, which means that confidence intervals for the mean value can alternatively be based on the t distribution or the normal distribution.

4.1 Frequentist interval for μ

Let μ_0 be the observed value for $\overline{x}$. Then the Eq. (5a) specifying μ_l reads, with utilization of the t distribution:

$$\begin{aligned} & P_{\mu=\mu_l}(\overline{x} \geq \mu_0) = \alpha/2 \quad (20)\\ \Leftrightarrow\quad & P\left(Z \geq (\mu_0 - \mu_l)/\sqrt{s^2/n}\right) = \alpha/2 \\ \Leftrightarrow\quad & 1 - \mathrm{pt}\left((\mu_0 - \mu_l)/\sqrt{s^2/n}, n-1\right) = \alpha/2 \\ \Leftrightarrow\quad & (\mu_0 - \mu_l)/\sqrt{s^2/n} = \mathrm{qt}(1-\alpha/2, n-1) \\ \Leftrightarrow\quad & \mu_l = \mu_0 - \mathrm{qt}(1-\alpha/2, n-1)\cdot\sqrt{s^2/n} \end{aligned}$$

where *pt* is the CDF of the t distribution, and *qt* its inverse. In the same way, Eq. (5b) can be solved for μ_u. With utilization of the symmetry property $\mathrm{qt}(t) = -\mathrm{qt}(1-t)$, the confidence interval based upon the t distribution becomes:

$$\overline{x} \pm t_{1-\alpha/2}(n-1)\cdot\sqrt{s^2/n} \quad (21)$$

where $t_{1-\alpha/2}(n-1)$ denotes the $(1-\alpha/2)$ quantile of the t distribution, which can be computed with the R function *qt*.

Based on the normal distribution, the same calculation method yields the confidence interval

$$\overline{x} \pm z_{1-\alpha/2}\cdot\sqrt{s^2/n} \quad (22)$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution, which can be computed with the R function *qnorm*.

It seems paradoxical that we obtain the different confidence intervals (21) or (22), depending on a condition (the underlying distribution) that we do not know about. This is no contradiction, however. Although

$$t_{1-\alpha/2}(n-1) > z_{1-\alpha/2} \quad \text{for all } n \quad (23)$$

and the interval (21) is therefore always slightly larger, for large n both intervals become asymptotically similar because of

$$\lim_{n\to\infty} t_{1-\alpha/2}(n-1) = z_{1-\alpha/2} \quad (24)$$

For $\alpha = 0.05$, both values are close to two, which leads for both of the above confidence intervals to the rule of thumb "two times sigma", with $\sigma = \sqrt{s^2/n}$.

4.2 Likelihood ratio for μ

On basis of the t distribution the specifying equation (6) for the likelihood ratio support interval reads

$$\frac{L(\mu)}{L(\hat{\mu})} = \left(1 + \frac{n(\overline{x}-\mu)^2}{s^2(n-1)}\right)^{-n/2} \geq \frac{1}{K} \quad (25)$$

This equation can readily be solved for μ, which yields the support interval

$$\overline{x} \pm \sqrt{(K^{2/n}-1)s^2\frac{n-1}{n}} \quad (26)$$

On basis of the normal distribution the specifying equation reads

$$\frac{L(\mu)}{L(\hat{\mu})} = \exp\left(-\frac{n(\overline{x}-\mu)^2}{2s^2}\right) \geq \frac{1}{K} \quad (27)$$

[5] The esoterically sounding term "degrees of freedom" is just the parameter of the t distribution.

[6] The "standard" normal distribution is the normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$.

This can again be solved elementary for μ, too, which yields the support interval

$$\overline{x} \pm \sqrt{\frac{2s^2}{n} \ln K} \tag{28}$$

It seems as though (26) and (28) were completely different intervals, but in fact they are very similar: for large n, both intervals are asymptotically equal because of[7]

$$\ln x = \lim_{n\to\infty} n(x^{1/n} - 1) \tag{29}$$

The numerical evaluation of the right hand side of Eq. (26) becomes inaccurate, however, for large n due to extinction of the most leading digit from from similar floating point numbers. Therefore, Eq. (28) is preferable for large n even in the case of the t distribution.

When we compare the support interval (28) with the confidence interval (21), we see the reason for the choice $K = 8$: it is $\sqrt{2\ln 8} \approx 2.0393$, which means that the frequentist interval for $\alpha = 0.05$ and the LR support interval roughly coincide. For $K = 7$, it is even with good accuracy $\sqrt{2\ln K} \approx z_{1-\alpha/2}$, but, as we will see in section 7, the frequentist interval based on $z_{1-\alpha/2}$ is generally to small, so that $K = 8$ is a safer choice.

4.3 Highest posterior density for μ

On basis of the t distribution, Eq. (7) becomes with the assumption of a constant "a priori" distribution $p(\mu) = const.$:

$$\begin{aligned} p(\mu|\overline{x}) &= \frac{\sqrt{n}\Gamma(\frac{n}{2})}{s\sqrt{\pi(n-1)}\Gamma(\frac{n-1}{2})} \left(1 + \frac{(\overline{x}-\mu)^2 n}{s^2(n-1)}\right)^{-\frac{n}{2}} \\ &= \sqrt{\frac{n}{s^2}} \cdot \mathrm{dt}\left(\frac{(\overline{x}-\mu)\sqrt{n}}{s}, n-1\right) \end{aligned} \tag{30}$$

where *dt* is the R function for the probability density of the t distribution. On basis of the normal distribution, we obtain under the analogous assumption $p(\mu) = const.$:

$$\begin{aligned} p(\mu|\overline{x}) &= \sqrt{\frac{n}{2\pi s^2}} \cdot \exp\left(-\frac{(\overline{x}-\mu)^2 n}{2s^2}\right) \\ &= \mathrm{dnorm}(\mu, \overline{x}, s^2/n) \end{aligned} \tag{31}$$

where *dnorm* is the R function for the probability density of the normal distribution. The resulting densities are thus identical to the symmetric densities used for the frequentist interval, which has the effect that the specifying equation (20) for the HPD interval has the same solution as the specifying equation for the frequentist interval. The HPD interval for the mean value is therefore exactly identical to the frequentist interval (21) or (21), respectively.

This is no coincidence, but a consequence of the fact that μ is a "location parameter", i.e., that $p(\overline{x}|\mu) = f(\overline{x}-\mu)$. When this functional relationship holds, frequentist interval and HPD interval are always identical [10].

5 Maximum likelihood estimators

To obtain a confidence interval for different estimators, it is necessary to know the probability distribution of the estimated value $\hat{\theta}$. Unfortunately, this is almost impossible in other cases than the aforementioned two examples. There is however a large category of estimators for which the asymptotic distribution is known: maximum likelihood (ML) estimators are asymptotically normally distributed around the true value θ for "regular" log-likelihood functions[8] $\ell(\theta)$ (see Eq. (2b)) for large n. In other words, the asymptotic probability density of $\hat{\theta}$ is given by

$$p(\hat{\theta}) = \frac{\exp\left(-\frac{1}{2}\langle\hat{\theta}-\theta, \Sigma^{-1}(\hat{\theta}-\theta)\rangle\right)}{\sqrt{(2\pi)^t \det(\Sigma)}} \tag{32}$$

where t is the number of parameters $\theta = (\theta_1, \ldots, \theta_t)$, Σ is the covariance matrix, and the exponent "-1" denotes matrix inversion.

If it is thus possible to determine the covariance matrix $(\sigma_{ij}) = \Sigma$, then its diagonal elements $\sigma_{ii} = Var(\theta_i)$ can be used to construct confidence intervals based on the normal distribution as in section 4:

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\sigma_{ii}} \tag{33}$$

Alternatively, it would also be sufficient to have a direct estimator for the variances σ_{ii} of the parameters. This leads to two possible approaches for an estimation of the variance of maximum likelihood estimators:

[7] This limiting value follows from inversion of [7] Eq. 4.2.21.

[8] The precise requirements are as follows: the log-likelihood function $\ell(\theta)$ must be three times continuously differentiable, the expectation values of all first and second derivatives exist, and the third derivations must be bounded by a function with finite expectation value [4].

- estimation of the covariance matrix via inversion of the Hessian matrix of the log-likelihood function
- jackknife estimator for the variance

The first method has the advantage that it can yield closed formulas for the variance in cases that allow for an analytic calculation of the Hessian matrix. The second method has the advantage that it requires no analytic or numeric calculation of derivatives at all, but that it provides an elementary and fast algorithm for computing the variance.

When the requirements listed in footnote 8 do not hold, the Hessian matrix cannot be computed, and the first method is ruled out. Although the jackknife variance can nevertheless be computed even in this case, it is of little use, because neither is guaranteed that the estimator is normally distributed, nor that the jackknife variance is a good estimator for the true variance (see [11] for a counterexample). In such a situation, it is therefore necessary to resort to the bootstrap method which is described in section 6.

5.1 Hessian matrix

When the preconditions mentioned in footnote 8 hold, the covariance matrix in Eq. (32) can be estimated through [4]

$$(\sigma_{ij}) = \left(- \left. \frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j} \right|_{\theta=\hat{\theta}} \right)^{-1} \tag{34}$$

where $\ell(\theta)$ is the log-likelihood function form Eq. (2b), and the exponent "-1" denotes matrix inversion.

In many cases, neither the equation (3) specifying the ML estimator $\hat{\theta}$ can be solved in closed form, nor can the inverse of the Hessian matrix (34) be computed analytically. This does not mean, however, that this method must be ruled out in this case, because a numerical solution is often viable. The R function *optim* even offers an argument *hessian=TRUE* which asks for an additional estimation of the Hessian matrix during optimization. An example implementation utilizing this function is given in listing 4.

```
lnL <- function(theta1, theta2, ...) {
  # definition of the negative (!)
  # log-likelihood function
  ...
}

# starting values for the optimization
theta0 <- c(start1, start2, ...)

# optimization
p <- optim(theta0, lnL, hessian=TRUE)
if (p$convergence == 0) {
  theta <- p$par
  covmat <- solve(p$hessian)
  sigma <- sqrt(diag(covmat))
}
```

Listing 4: R code for the numerical calculation of an ML estimator for $\theta = (\theta_1, \ldots, \theta_t)$ in combination with a variance estimation for the estimated values. The log-likelihood function must be defined negatively, because *optim* seeks the minimum instead of the maximum.

5.2 Jackknife

The jackknife method is based on the idea to compute the estimator $\hat{\theta}(x_1, \ldots, x_n)$ many times, but each time with the omission of one value x_i. The variance of $\hat{\theta}$ is then estimated form the distribution of these "delete-one" estimators. Let $\theta_{(i)}$ be the estimator computed without the i-th data point x_i. Then the jackknife estimator for the variance of $\hat{\theta}$ is:

$$\sigma_{JK}(\hat{\theta}) = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} (\theta_{(i)} - \theta_{(.)})^2} \tag{35}$$

$$\text{with} \quad \theta_{(.)} = \frac{1}{n} \sum_{i=1}^{n} \theta_{(i)}$$

When θ is a vector with several components, it is also possible to estimate the entire covariance matrix $Sigma$ with the jackknife. This is of little use however, because the confidence intervals (33) only need the diagonal elements σ_{ii} of Σ. Hence it is sufficient to apply Eq. (35) to each component of θ. For asymptotically normally distributed ML estimators, σ_{JK} is an asymptotically unbiased and consistent estimator for their variance [12]. An implementation of formula (35) is given in listing 5.

6 Bootstrap

Similar to the jackknife method, the bootstrap method is based on the generation of new data sets from the original data $x_1, \ldots, x_n$, In the bootstrap, this is however not done deterministically via cyclic omission,

```
theta.hat <- function(x) {
  # implementation of the estimator
  ...
}
theta.jk <- rep(0, n)
for (i in 1:n) {
  theta.jk[i] <- theta.hat(x[-i])
}
theta.dot <- mean(theta.jk)
sigma.jk <- sqrt((n-1) *
        mean((theta.jk-theta.dot)^2))
```

Listing 5: Calculation of the jackknife variance of an estimator $\hat{\theta}(x_1, \ldots, x_n)$ in R.

but in a random way. This can either be be done by n-fold drawing with replacement (*non-parametric bootstrap*), or by n-fold generation of random numbers distributed according to the density estimated with the estimator (*parametric bootstrap*). The non-parametric bootstrap thus considers all observed data, while the parametric bootstrap only considers the point estimator $\hat{\theta}$ computed from the data.

When we repeat the drawing of new data sets R times, we obtain a Monte-Carlo simulation of the distribution of the estimator $\hat{\theta}$. From this distribution, confidence intervals can be estimated.[9,10]. There is a bewildering variety of methods for estimating a confidence interval from the simulated distribution, which are summarized together with their asymptotic coverage probability in [15]. Their theoretical background is explained in [6]. The most important methods are:

Percentile & Basic. The *Percentile bootstrap* was the original method proposed by Efron. It simply takes the percentiles of the simulated distribution $\hat{\theta}_1, \ldots, \hat{\theta}_n$ of $\hat{\theta}$. The *Basic bootstrap interval* flips the percentile bootstrap at $\hat{\theta}$. Venables & Ripley recommend the Basic bootstrap over the Percentile bootstrap [16], but the experiments in section 7 lead to the opposite conclusion.

Bias corrected accelerated (BC_a). This method tries to estimate transformation parameters that make the distribution symmetric. This is the method recommended by Efron.

[9]It is also possible to estimate the variance from this Monte Carlo simulation [13], but a confidence interval based on this variance would again make the assumption of a normally distributed $\hat{\theta}$.

[10]You could think that instead of the bootstrap random samples, one could alternatively estimate the confidence interval from the distribution of the n jackknife "delete one" estimators $\theta_{(i)}$. This does not work, however, because even in the case of regular ML estimators, the distribution of the $\theta_{(i)}$ is not normal and therefore not representative for the distribution of $\hat{\theta}$ [14].

```
# estimator function; indices are
# for boot() to select data points
schaetzer <- function(x, indices) {
  x.auswahl <- x[indices]
  ... # compute estimator from x.auswahl
  return(theta.hat)
}

# bootstrap confidence intervals
boot.out <- boot(data=x,
        statistic=schaetzer, R=1000)
ci <- boot.ci(boot.out,
        conf=0.95, type="all")

# percentile interval:
theta1 <- ci$perc[4]
theta2 <- ci$perc[5]

# basic interval:
theta1 <- ci$basic[4]
theta2 <- ci$basic[5]

# BCa interval:
theta1 <- ci$bca[4]
theta2 <- ci$bca[5]
```

Listing 6: Calculation of bootstrap confidence intervals with the R library *boot*.

It can be shown that the BC_a interval has a coverage probability that converges asymptotically for large n to the nominal value $1-\alpha$ with a rate $o(n^{-1})$ [6]. This is a faster convergence than for the classical $z_{1-\alpha/2}\sigma$ interval, which has a convergence rate of $o(n^{-1/2})$. DiCiccio & Efron concluded form this observation that the bootstrap method is generally preferable (comments to [6], p. 228):

> "If the standard intervals were invented today, they might not be publishable."

This is somewhat misleading, however, because the difference between the different confidence intervals is marginal for large n anyway, and the convergence rate for large n is therefore of merely theoretical interest. Of practical relevance is instead the behavior for small n, where the bootstrap intervals perform poorer than the classic intervals (see section 7). In defense of the comparatively poor performance of the bootstrap for

small n in a specific study, its inventor, Bradley Efron, wrote [17]:

> "Bootstrap methods are intended to supplement rather than replace parametric analysis, particularly when parametric methods can't be used because of modeling uncertainties or theoretical intractability."

The function *boot.ci* from the R library *boot* can compute a number of bootstrap confidence intervals, including the three aforementioned. According to Efron & Tibshirani [18], the minimum value for the number r of bootstrap replications should be $R = 1000$. Usage of the R function *boot.ci* is shown in listing 6.

Apart from the confusion about the most appropriate bootstrap interval in a particular situation, the bootstrap method has another drawback: as it is based on Monte Carlo simulations, its results are not deterministic and not reproducible. This means that two different researchers will end with different confidence intervals for the same data. Leon Jay Gleser sees therein a violation of a rule that he calls the "first law of applied statistics" (comments to [6], p. 220):

> "Two individuals using the same statistical method on the same data should arrive at the same conclusion."

It should be noted that the differences are small, though. Nevertheless they are noticeable and the bootstrap method might therefore leave some users with a slightly uneasy feeling.

7 Performance in examples

This section provides a comparative evaluation of the different confidence intervals with respect to examples for all three of the aforementioned cases. Apart from the coverage probability P_{cov}, the relative size of the confidence intervals is of interest, too.

For fixed n, the relative frequency can only be one of $n + 1$ discrete values, so that $P_{cov}(p)$ can be computed exactly. As an example for the mean value, I have chosen an asymmetric distribution with density $f(x) = 3x^2$, such that Monte Carlo simulations might show whether the bootstrap provides any advantages over the classic intervals that assume symmetry. As an example for an ML estimator, I have chosen the parameter λ of the exponential distribution. In this example, even the inverse of the Hesse matrix can be calculated analytically in closed form, which allows for a comparison of all methods by means of a Monte Carlo simulation. From the bootstrap methods, I have only tested the non-parametric bootstrap because the parametric method is not universally applicable, but must be tailored to each particular use case, which might be too much of an effort for an end user only interested in confidence intervals[11].

7.1 P_{cov} for the relative frequency

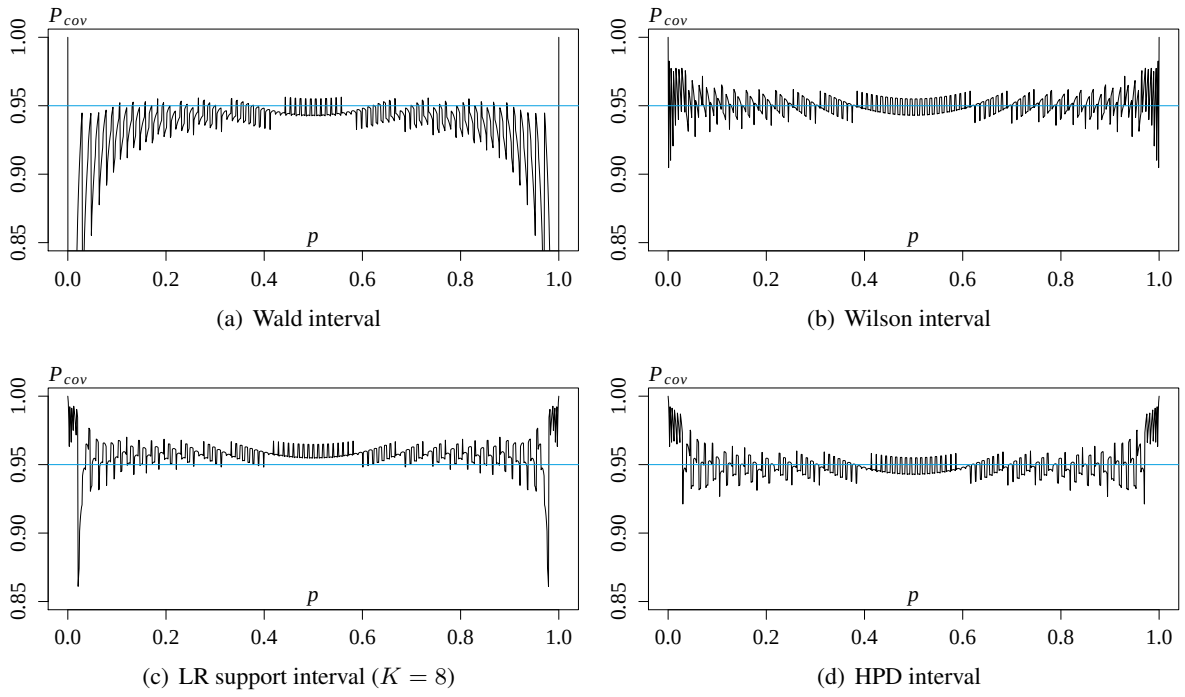
The coverage probability of different confidence intervals for a binomial proportion was already investigated by Brown et al. [9]. Based on their results, they recommended the Wilson interval. As they did not include the LR support interval or the HPD interval in their study, the corresponding behavior of $P_{cov}(p)$ as a function of p is shown in Fig. 4. The corresponding behavior of the "exact" (Clopper-Pearson) interval is shown in Fig. 1. The curves have been computed as follows:

- for every $0 \leq k \leq n$, the confidence interval was calculated
- for every sampled value $p \in [0, 1]$, the probabilities of all k were added for which p fell into the confidence interval

As already noted by Brown et al., the classical Wald interval taught in statistics text books has a way too low coverage probability almost over the entire range of p values. P_{cov} even approaches zero for small or large p. The Wilson interval, on the contrary, fluctuates around the nominal value, albeit with greater deviations towards the boundaries of the p-range. Interestingly, the HPD interval has an even better better coverage probability than the Wilson interval because it mostly shows a similar behavior, but has no too small values at the boundaries. The behavior of the LR support interval for $K = 8$ is similar to that of the exact Clopper-Pearson interval, but there are instances where P_{cov} falls considerably below the nominal value.

Another evaluation criterion is the interval length, which should be minimal for comparable coverage probability. The maximum length of all intervals occurs for $p = 1/2$ and is plotted as a function of n in Fig. 5. The widest interval is the exact interval, which

[11] Apart from an understanding of probability theory, it also requires knowledge about the generation of random numbers (transformation method, rejection method [19]).



(a) Wald interval

(b) Wilson interval

(c) LR support interval ($K = 8$)

(d) HPD interval

Figure 4: Coverage probability $P_{cov}(p)$ of the confidence intervals for a binomial proportion as a function of the true parameter value p for $n = 100$ and $1 - \alpha = 0.95$.

is inevitable prize for the guarantee of $P_{cov}(p) \geq 1-\alpha$ with a greater P_{cov} more often than not. Curiously enough, the maximum length of the Wald interval is greater than that of the Wilson or HPD interval, although its coverage probability is smaller. This apparent contradiction is resolved when the interval lengths for varying $\hat{p}$ with fixed n are compared (see Fig. 6). It can be seen that the classical Wald interval is unnecessarily wide for $\hat{p} \approx 0.5$, but too short for $\hat{p} \approx 0$ or $\hat{p} \approx 1$.

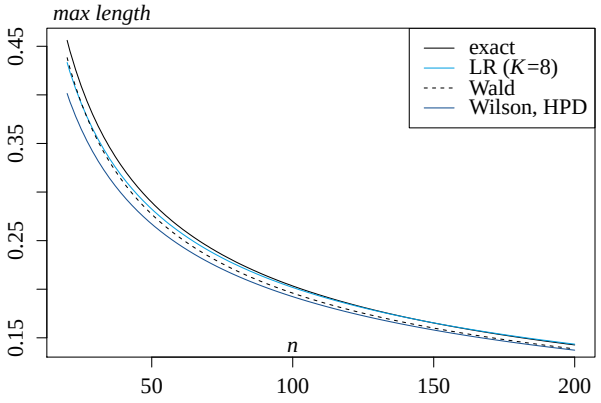


Figure 5: Maximum length of the confidence intervals for the relative frequency as a function of n for $1 - \alpha = 0.95$. The maximum length of the HPD and Wilson interval are nearly identical.

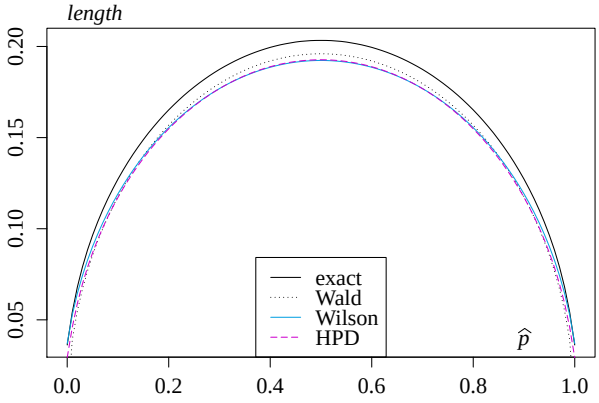


Figure 6: Confidence interval length for the relative frequency as a function of $\hat{p}$ for $1 - \alpha = 0.95$ and $n = 100$.

It is interesting to note that the HPD interval for $\hat{p} \approx 0$ or $\hat{p} \approx 1$ is shorter than the Wilson interval, even though it has a considerably higher coverage probability in this region. With respect to the criteria coverage probability and length, the HPD interval has the best properties. It has the drawback, though, that it can be computed only numerically (see listing 3). If a closed formula is required, the Wilson interval (see Eq. (14))

can be used as an alternative, provided $\hat{p}$ is not too close to zero or one.

7.2 P_{cov} for the mean value

To compare the classical confidence intervals for the mean value with the bootstrap intervals, I have chosen a random variable with the probability density

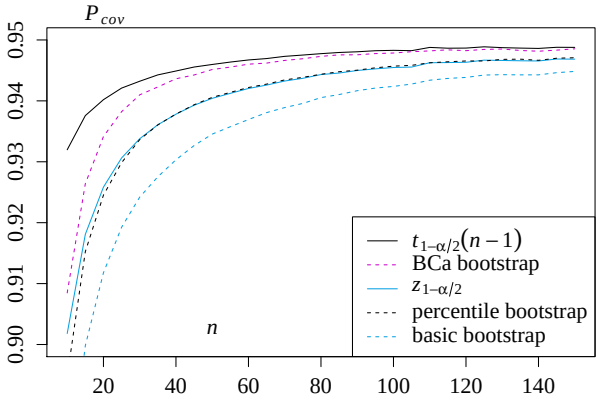
$$f(x) = \begin{cases} 3x^2 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

The expectation value of this distribution is $3/4$, and random numbers drawn from this distribution can be generated by means of the transformation method [19] with

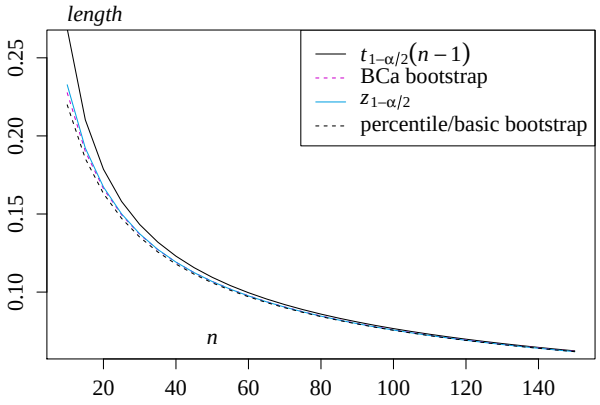
```
runif(N, min=0, max=1)**(1/3)
```

The number of simulated mean value measurements was set to $N = 10^6$, which means that the coverage probability can be estimated with an accuracy ± 0.0004 for $\alpha = 0.05$.

The behavior of P_{cov} and the length of the different confidence intervals as a function of the number n of observed data points is shown in Fig. 7. Surprisingly, the classic confidence interval based on the t distribution has the best coverage probability throughout, although the distribution of $\overline{x}$ is asymmetric for small n. The weaknesses of the bootstrap method for small n are thus not compensated in this case by its ability to take asymmetries into consideration. The best bootstrap interval in this case is the BC_a interval. It has a length that is is similar to the classic $z_{1-\alpha/2}$ interval, but with a greater P_{cov}. Venables' & Ripley's recommendation for the basic over the percentile bootstrap cannot be confirmed, but, on the contrary, the basic bootstrap interval has a clearly too low P_{cov} in this case, whilst the percentile interval has a coverage probability that is comparable to the classic $z_{1-\alpha/2}$ interval.



(a) Coverage probability



(b) Average length

Figure 7: Coverage probability and average length of the different confidence intervals for the mean value of n random variables distributed according to Eq. (36).

7.3 P_{cov} for ML estimators

Let us consider the exponential distribution as a test case for comparing the different confidence intervals for maximum likelihood estimators. The exponential distribution has the single parameter λ and the probability density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

The log-likelihood function obtained from this density is

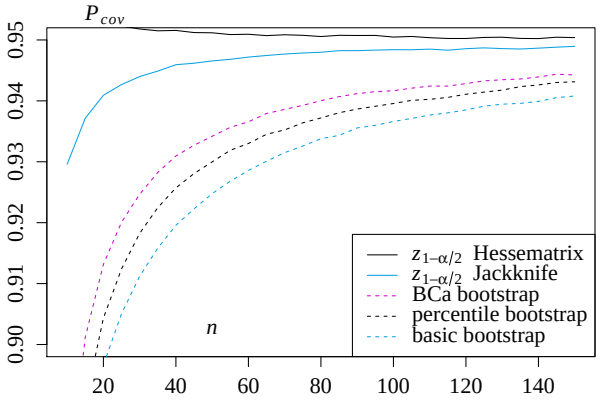
$$\ell(\lambda) = n \log(\lambda) + \lambda \sum_{i=1}^{n} x_i \tag{38}$$

The ML estimator for λ is obtained by solving the equation (3) for λ:

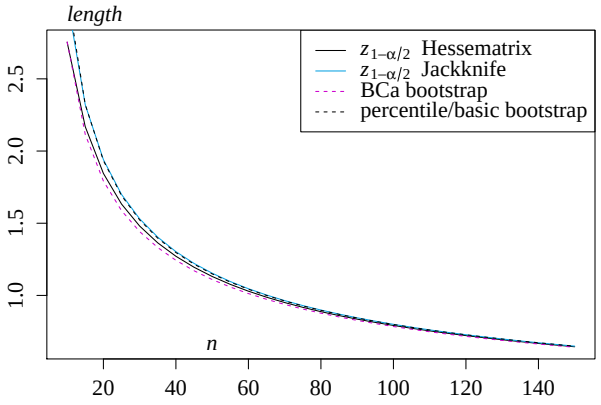
$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{x}} \tag{39}$$

As the exponential distribution only has a single parameter, the Hessian matrix is of dimension 1×1, ergo a scalar. It can be readily computed as

$$H(\lambda) = \left(\frac{\partial^2}{\partial \lambda^2} \ell \right) = \left(-\frac{n}{\lambda^2} \right) \tag{40}$$



(a) Coverage probability



(b) Average length

Figure 8: Coverage probability and average length of the different confidence intervals for the ML estimator of the parameter λ of the exponential distribution.

When the variance of $\hat{\lambda}$ is estimated form the Hessian with the method of section 5.2, it reads

$$\hat{\sigma}_{HM} = \sqrt{\Big(- H(\hat{\lambda})\Big)^{-1}} = \frac{\hat{\lambda}}{\sqrt{n}} \qquad (41)$$

Again, I have generated $N = 10^6$ times n exponentially distributed random numbers with $\lambda = 2$ in order to simulate the distribution of $\hat{\lambda}$ and to compare P_{cov} and average length of the different confidence intervals. The results are shown in Fig. 8. The classical interval with $\hat{\sigma}_{HM}$ has the best coverage probability, followed by the classical interval with $\hat{\sigma}_{JK}$. Among the bootstrap intervals, the BC_a interval has the highest coverage probability, and again the percentile bootstrap performs better than the basic bootstrap. Venables' & Ripley's recommendation in favor of the basic bootstrap must therefore be rejected. Overall, the bootstrap intervals show a coverage probability that is clearly below the nominal value $1 - \alpha$.

It is surprising that the confidence interval based on the jackknife variance is wider, but has a smaller coverage probability than the interval based on the Hessian. A closer look at the simulated distribution of $\hat{\lambda}$ reveals that in this case the ML estimator is biased and is on average too large[12]. As $\hat{\sigma}_{HM}$ is proportional to $\hat{\lambda}$ according to Eq. (41), the confidence interval is wider when the estimated value is too large, which compensates the bias of the ML estimator in this case. This lead to a correlation of $|\hat{\lambda} - \lambda|$ with respect to $\hat{\sigma}_{HM}$ of about 0.60 in the Monte-Carlo simulations for $n = 20$, but of only about 0.40 with respect to $\hat{\sigma}_{JK}$. This explains why P_{cov} can be smaller for the wider interval.

8 Conclusions

For the practitioner, the comparative evaluation of the different confidence intervals leads to the following recommendations:

1) For a relative frequency, the HPD interval (listing 3) or the Wilson interval (Eq. (14)) should be used. The Wilson interval has the advantage of a closed formula, but it has a smaller coverage probability than the HPD interval for p values close to zero or one.

2) For mean values, the classical confidence interval based on the t distribution should be used (Eq. (21)).

3) For ML estimators with a smooth log-likelihood function, the confidence interval $z_{1-\alpha/2} \cdot \hat{\sigma}$ should be used. The variance $\hat{\sigma}$ can be estimated either from the Hessian matrix or, in a simpler way, by means of the jackknife (listing 5).

4) In the remaining cases, the BC_a bootstrap interval should be used.

The results of this technical report thus confirm the already cited remark by Efron [17]:

> "Bootstrap methods are intended to supplement rather than replace parametric analysis, particularly when parametric methods can't be used because of modeling uncertainties or theoretical intractability."

[12] ML estimators are only guaranteed to be *asymptotically* unbiased for large n.

References

[1] L. Fahrmeir, R. Künstler, I. Pigeot, and G. Tutz, *Statistik*. Berlin: Springer, 5 ed., 2004.

[2] J. D. Blume, "Likelihood methods for measuring statistical evidence," *Statistics in medicine*, vol. 21, no. 17, pp. 2563–2599, 2002.

[3] N. Turkkan and T. Pham-Gia, "Computation of the highest posterior density interval in Bayesian analysis," *Journal of statistical computation and simulation*, vol. 44, no. 3-4, pp. 243–250, 1993.

[4] W. H. Greene, *Econometric Analysis*. New Jersey: Prentice Hall, 4 ed., 2000.

[5] P. Zsohar, "Short introduction to the generalized method of moments," *Hungarian Statistical Review*, vol. 16, pp. 150–170, 2010.

[6] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statistical science*, pp. 189–228, 1996.

[7] M. Abramowitz, I. Stegun, M. Danos, and J. Rafelski, *Pocketbook of mathematical functions*. Frankfurt: Harri Deutsch, 1984.

[8] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[9] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Statistical science*, vol. 16, no. 2, pp. 101–117, 2001.

[10] D. Karlen, "Credibility of confidence intervals," in *Advanced Statistical Techniques in Particle Physics, Proceedings*, (Durham), pp. 53–57, 2002.

[11] R. G. Miller, "A trustworthy jackknife," *The Annals of Mathematical Statistics*, vol. 35, no. 4, pp. 1594–1605, 1964.

[12] J. A. Reeds, "Jackknifing maximum likelihood estimates," *The Annals of Statistics*, vol. 6, no. 4, pp. 727–739, 1978.

[13] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, vol. 37, no. 1, pp. 36–48, 1983.

[14] C.-F. J. Wu, "Jackknife, bootstrap and other resampling methods in regression analysis," *the Annals of Statistics*, vol. 14, no. 4, pp. 1261–1295, 1986.

[15] J. Carpenter and J. Bithell, "Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians," *Statistics in medicine*, vol. 19, no. 9, pp. 1141–1164, 2000.

[16] W. Venables and B. Ripley, *Modern Applied Statistics with S*. New York: Springer, 4 ed., 2002.

[17] B. Efron, "Bootstrap confidence intervals: Good or bad?," Technical Report 116, Stanford University, Division of Biostatistics, March 1987.

[18] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical science*, vol. 1, no. 1, pp. 54–75, 1986.

[19] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical recipes in Pascal: the art of scientific computing*. Cambridge University Press, 1989.