Interval Score for Comparison of Confidence Intervals

Master Thesis in Biostatistics (STA495)

by

Lisa Hofer
lisa.hofer@uzh.ch
11-717-709

supervised by

Prof. Dr. Leonhard Held
Department of Biostatistics
University of Zurich

University of Zurich[UZH]

Zurich, January 2022

Abstract

In many application areas, different ways to construct confidence intervals exist and methods to compare them are necessary to decide which should be used in practice. Especially for the binomial proportion, there are more than 20 confidence intervals to choose from. Gneiting and Raftery (2007) suggested that proper scoring rules, as used for (central) prediction intervals, could also be useful to compare confidence intervals. The proposed interval score is a loss function that combines the coverage as a measure of calibration and the interval width as a measure of sharpness. We evaluated eleven confidence intervals for the binomial proportion regarding the expected interval score. By using a summary measure which can take into account different weighting of the true proportions, we obtained a clear ranking of the confidence intervals. In general, this ranking recommends the Wilson confidence interval or Bayesian equal-tailed or HPD intervals with a uniform prior. If the performance in scenarios with rare cases is important, as for example in the estimation of sensitivity and specificity, it recommends Bayesian intervals with Jeffreys' prior. While the use of proper scoring rules in estimation problems yet needs to be put on a theoretically more solid basis, the results suggest that the interval score is a useful evaluation method as opposed to coverage probability or expected interval width alone. This novel approach for comparison of confidence intervals could also be used in other application areas such as confidence intervals for meta-analysis.

Acknowledgments

First, I would like to thank my supervisor Leonhard Held for his support and plenty of good advice. Being included in weekly research meetings with Charlotte Micheloud and Samuel Pawel has been incredibly helpful. I enjoyed the scientific exchange and most of all that I experienced real enthusiasm about statistical methodology. I would also like to thank Tilmann Gneiting and Johannes Bracher for their expert comments about my project. Furthermore, I am grateful to the lecturers from the Master Program in Biostatistics: Torsten Hothorn, Reinhard Furrer, Leonhard Held, Ulrike Held and Eva Furrer. I have learned a lot from each of you and want to thank especially Eva Furrer for being such a committed study coordinator and helping the students in every possible way. What this program offers, both from a social and a professional aspect, is truly unique. I am also grateful to my fellow students who accompanied me during my studies. Especially to Lucas Kook who has become a great friend and whose thoughts about statistics I value highly. Finally, I would like to thank Charlotte, Samuel and Lucas for their great reviews and valuable suggestions.

Lisa Hofer
January 5, 2022

Contents

Chapter 1

Introduction

Estimation of unknown parameters such as treatment effects is a key task in biostatistics. Based on the interpretation of these estimates, medical decisions are made. Today, it is strongly recommended to report point effect estimates together with confidence intervals (Altman *et al.*, 2000). That is, an interval estimate should be indicated as a range of plausible values for the unknown parameter. In contrast to a point estimate or a p-value, the confidence interval provides information about the magnitude and the precision of the effect estimate at once (Rothman, 1986).

Although researchers agree that point estimates should be reported together with confidence intervals, it is not clear which confidence interval to choose. For many parameters, there exist different interval estimators depending, for example, on different ways of approximation. In particular for the binomial proportion, one can choose from numerous different methods to compute confidence intervals. In 1998, seven methods have been compared in Newcombe (1998) and new methods are still being developed; see Gillibert *et al.* (2021) for a recent systematic review and Pires and Amado (2008) for a comparison of 20 methods. The goal of these comparisons is to recommend the method with the best properties for the practical use.

The quality of a confidence interval is commonly assessed by two properties: the actual coverage probability and the expected interval width. The confidence interval should cover the true parameter with a high probability and should be precise. That means that the actual coverage probability should be close to a nominal confidence level (usually 95%) and the expected interval width should be small. In terms of these evaluation criteria, usually the Wilson confidence interval (score method) is recommended for binomial proportions (Newcombe, 1998; Held and Sabanés Bové, 2020). However, there is still no widely recognized consent. Still, often the Wald confidence interval is used in practice although it is known for its coverage bias (Brown *et al.*, 2001; Gillibert *et al.*, 2021). The reason is mainly that it has a simple form, hence it is easy to use and simple to communicate. Since there are multiple evaluation criteria, the recommendations depend on what evaluation criteria are used and how they are assessed.

One issue is the trade-off between coverage and width. The best interval would be as small as possible while still respecting the correct coverage. This is a trade-off because decreasing the interval width decreases the coverage and increasing the coverage increases the interval width. This relation is referred to as *sharpness subject to calibration*, where the coverage calibrates the interval while the width determines its sharpness (Gneiting *et al.*, 2007). A way to assess calibration and sharpness simultaneously are scoring rules. Such a scoring rule is the interval score that combines coverage and width in a loss function. The interval score has been developed for prediction intervals in Gneiting and Raftery (2007) where the authors suggest that it could also be used for interval estimates. It is intended to compare intervals for the same nominal coverage that have equal lower and upper exceedance probabilities which is called central (Gneiting and

Raftery, 2007, p. 18). Only in this case, the interval score is a *proper* scoring rule, that is, a scoring rule such that the optimal interval estimate minimizes the expected score.

Another issue is that the comparisons depend on the chosen confidence level, the value of the true unknown parameter and possibly on other (known) parameters. In the binomial case, for example, these parameters are the (known) sample size n, the (chosen) confidence level γ and the (unknown) true success probability π. Graphical representations can be used where the coverage probability or the expected width are represented as a function of π for fixed n and γ as for example in Held and Sabanés Bové (2020, p. 117–119). However, it is hard to make a recommendation only based on graphical representations since the performance of a method might vary substantially depending on the parameter setting. It is a well-known problem that the actual coverage probability of the Wald confidence interval is poor for extreme cases where π is near the boundaries 0 and 1 (Brown *et al.*, 2001; Held and Sabanés Bové, 2020). Moreover, the coverage probability oscillates in an erratic way as a function of π. Less known is that these oscillations lead to poor coverage probability also if π is not near the boundaries (Vollset, 1993; Agresti and Coull, 1998; Newcombe, 1998; Brown *et al.*, 2001). A summary measure that summarizes over different parameter values such as the possible values for π is needed in order to make recommendations. In the literature, minimum and mean coverage probabilities are used (Newcombe, 1998; Pires and Amado, 2008).

In this master thesis, a new technique for the comparison of confidence intervals for the binomial proportion is investigated. The main objective is to address the trade-off between coverage and width by using the expected interval score as a measure to compare confidence intervals. Binomial proportions have been chosen as an example with many different confidence intervals of which the following eleven have been selected: Clopper-Pearson, Wilson, Wald, Rindskopf (logit Wald with adjustment), variance-stabilized Wald, Agresti-Coull, likelihood ratio and the Bayesian equal-tailed and HPD intervals with uniform and Jeffreys priors. Moreover, also the issue of suitable summary measures, which is to some extent application-specific, is addressed. First, the integral of the expected interval score over all possible true proportions π is used as a summary measure summarizing for different values of π. Secondly, the integral of the expected interval score after the variance-stabilizing transformation of π is used as a summary measure that attributes more weight to the extreme cases of π at the boundaries. Thirdly, the weighted interval score is used to summarize over different confidence levels. Since not all confidence intervals are central, a *generalized* interval score for non-central intervals is developed.

Using these two novel approaches to evaluate confidence intervals, a clear ranking of the confidence intervals is obtained. Regarding the integral of the expected interval score on the scale of π, the best confidence intervals are the uniform equal-tailed followed by the Wilson. Regarding the integral on the variance-stabilized scale, the best intervals are the Jeffreys equal-tailed followed by the uniform HPD. The Wald confidence interval is the worst and should not be used. The same results hold when different confidence levels are combined using the weighted interval score. In terms of the generalized interval score for non-central intervals, the HPD intervals outperform the equal-tailed intervals.

The structure of this master thesis is as follows: The confidence intervals, the interval score as well as the summary measures are described in Chapter 2, followed by the presentation of the results in Chapter 3 and a discussion in Chapter 4.

Chapter 2

Methods

This chapter introduces notation and summarizes the statistical methods that are used in this master thesis. Unless otherwise stated, the presentation of the methods including notation follows Held and Sabanés Bové (2020).

2.1 Binomial proportion

The parameter of interest is the unknown success probability $\pi \in (0, 1)$ of a binomial sample $X \sim \text{Bin}(n, \pi)$, where X denotes the number of successes and n is the known number of trials, also called the sample size. A realization of the random variable X is denoted by x. The binomial distribution is a discrete distribution, $x \in \{0, 1, \ldots, n\}$, with probability mass function

$$f(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}.$$

For the boundary values $\pi = 0$ or $\pi = 1$, this distribution would be degenerate in the sense that it would be deterministic: identical to 0 for $\pi = 0$ or identical to n for $\pi = 1$.

The probability π of a certain event is often referred to as the binomial proportion of that event. If π is the underlying probability of the event in a population of size n, then $n\pi$ is the expected number of events, hence π is the (expected) proportion of events in that population.

2.2 Likelihood inference for a proportion

In a binomial experiment, the *maximum likelihood estimate* (MLE) of the unknown parameter π is the observed proportion

$$\widehat{\pi}_{\text{ML}}(x) = \frac{x}{n}.$$

The MLE is obtained by maximizing the (log-)likelihood function in π. For an observation x from $X \sim \text{Bin}(n, \pi)$, the likelihood function and related quantities from likelihood inference (see Held and Sabanés Bové (2020) for definitions) are the following:

$$\begin{aligned}
\text{likelihood function} \quad & L(\pi; x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \propto \pi^x (1 - \pi)^{n-x}, \\
\text{log-likelihood function} \quad & l(\pi; x) = x \log(\pi) + (n - x) \log(1 - \pi), \\
\text{score function} \quad & S(\pi; x) = \frac{x}{\pi} - \frac{n - x}{1 - \pi}, \\
\text{Fisher information} \quad & I(\pi; x) = \frac{x}{\pi^2} + \frac{n - x}{(1 - \pi)^2},
\end{aligned}$$

$$\text{expected Fisher information} \quad J(\pi) = \frac{n}{\pi(1-\pi)},$$
$$\text{observed expected Fisher information} \quad J\{\widehat{\pi}_{\text{ML}}(x)\} = \frac{n}{\widehat{\pi}_{\text{ML}}(x)\{1-\widehat{\pi}_{\text{ML}}(x)\}}.$$

These quantities are *estimates* as a function of the realization x and *estimators* as a function of the random variable X. As a function of X, they can be used to find an *approximate pivot*, *i. e.* a statistic whose distribution is asymptotically independent of the true parameter π. Approximate pivots are important for the construction of confidence intervals. The following three are the most popular:

$$\text{Wald statistic} \quad (\widehat{\pi}_{\text{ML}}(X) - \pi)\sqrt{J(\widehat{\pi}_{\text{ML}}(X))} = \frac{\frac{X}{n} - \pi}{\sqrt{\frac{X/n(1-X/n)}{n}}} \overset{a}{\sim} \text{N}(0,1),$$
$$\text{score statistic} \quad \frac{S(\pi; X)}{\sqrt{J(\pi)}} = \frac{\frac{X}{n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \overset{a}{\sim} \text{N}(0,1),$$
$$\text{likelihood ratio statistic} \quad -2\log\left\{\frac{L(\pi)}{L(\widehat{\pi}_{\text{ML}}(X))}\right\} \overset{a}{\sim} \chi^2(1),$$

where $\text{N}(0,1)$ denotes the standard normal distribution, $\chi^2(1)$ denotes the chi-squared distribution with one degree of freedom and the symbol $\overset{a}{\sim}$ means "is asymptotically distributed as" in the sense of convergence in distribution for $n \to \infty$. The denominator of the Wald statistic is equal to the standard error

$$\text{se}(\widehat{\pi}_{\text{ML}}(X)) = \sqrt{\frac{\widehat{\pi}_{\text{ML}}(X)(1-\widehat{\pi}_{\text{ML}}(X))}{n}}.$$

The Wald, Wilson and likelihood ratio confidence intervals are based on these statistics (see Section 2.3).

The approximate distribution of the likelihood ratio statistic is derived in Wilks' theorem using the asymptotical characteristic function (Wilks, 1938). The approximate distribution of the score statistic is derived using the central limit theorem. Now, the Wald statistic is an approximation of the score statistic that uses the standard error as a consistent estimator of the standard deviation $\sqrt{J^{-1}(\pi)}$ of the MLE (Held and Sabanés Bové, 2020, p. 98–99). It can also be viewed as an approximation of the (square root of the) likelihood ratio statistic that uses a second-order Taylor expansion of the log-likelihood function around the MLE (Held and Sabanés Bové, 2020, p. 109–110). Consequently, the Wilson and the likelihood ratio confidence intervals are more accurate than the Wald confidence interval.

Subsequently, $\widehat{\pi}_{\text{ML}}(x)$ and $\widehat{\pi}_{\text{ML}}(X)$ will both be abbreviated by $\widehat{\pi}_{\text{ML}}$ to simplify notation.

2.3 Frequentist intervals

The frequentist framework assumes that the proportion π is fixed but unknown. More precisely, it assumes a binomial distribution $\text{Bin}(n, \pi)$ for the number of successes x in a population of size n with an underlying true success probability π. An interval estimate (or estimator) for the parameter π in this setting is called *confidence interval* (CI). A $\gamma \cdot 100\%$ confidence interval for π, where $\gamma \in (0,1)$ is the *confidence level*, is defined as an interval $[L, U]$ that fulfills

$$\mathsf{Pr}(L \leq \pi \leq U) = \gamma.$$

Since π is fixed, no probability statement is attached to π but to the limits L and U. This is why the interpretation of a frequentist interval is slightly intricate: For repeated random samples $X \sim \text{Bin}(n, \pi)$, a $\gamma \cdot 100\%$ confidence interval will cover the parameter π in $\gamma \cdot 100\%$ of all cases.

The perfect confidence interval has a coverage probability of $\gamma \cdot 100\%$. However, since the binomial distribution is discrete, all confidence intervals will only approximately have the intended coverage probability (Held and Sabanés Bové, 2020; Gillibert *et al.*, 2021).

Let $\alpha = 1 - \gamma$ be the associated non-coverage and q be the $(1 + \gamma)/2$ quantile of $\text{N}(0, 1)$.

2.3.1 Clopper-Pearson

The (discrete) realizations x will lie between some x_1 and x_2 with a probability of *at least* γ. The $\gamma \cdot 100\%$ Clopper-Pearson confidence interval (Clopper and Pearson, 1934) inverts the determining inequalities for this interval for x to obtain an interval for the (continuous) parameter π. The limits L and U are derived from the two equations (Pires and Amado, 2008)

$$\sum_{j=x}^{n} \binom{n}{j} L^j (1-L)^{n-j} = \frac{\alpha}{2} \quad \text{and} \quad \sum_{j=0}^{x} \binom{n}{j} U^j (1-U)^{n-j} = \frac{\alpha}{2}. \tag{2.1}$$

The quantities in (2.1) are interpreted as $\mathsf{Pr}(X \geq x)$ for $X \sim \text{Bin}(n, L)$ and $\mathsf{Pr}(X \leq x)$ for $X \sim \text{Bin}(n, U)$. The relation

$$\sum_{j=x}^{n} \binom{n}{j} \pi^j (1-\pi)^{n-j} = \int_0^{\pi} f_b(t)dt$$

to the beta density function f_b of $\text{Be}(x, n - x + 1)$ is used to solve (2.1) for the limits

$$L = b_{(1-\gamma)/2}(x, n - x + 1) \quad \text{and} \quad U = b_{(1+\gamma)/2}(x + 1, n - x) \quad \text{for} \quad 0 < x < n,$$

where $b_\gamma(\alpha, \beta)$ is the γ quantile of $\text{Be}(\alpha, \beta)$. For $x = 0$ and $x = n$, the lower, respectively upper, limit is improper (one parameter is 0). In these cases, the solutions are calculated directly:

$$\begin{aligned} L = 0 \quad \text{and} \quad U = 1 - (\alpha/2)^{1/n} \quad \text{for} \quad x = 0, \\ L = (\alpha/2)^{1/n} \quad \text{and} \quad U = 1 \quad \text{for} \quad x = n. \end{aligned}$$

The Clopper-Pearson interval is known as an "exact" interval because (2.1) uses the exact distribution $X \sim \text{Bin}(n, \pi)$. However, it does not have exact coverage probability equal to γ. On the contrary, the minimum coverage probability (for any true proportion π) is *at least* γ. Hence, it is conservative.

2.3.2 Wilson

Based on the standard normal approximate pivot of the score statistic, the $\gamma \cdot 100\%$ Wilson confidence interval (Wilson, 1927) is the set of all parameter values π that satisfy

$$\pi^2 \left(n^2 + nq^2\right) + \pi \left(-2nx - nq^2\right) + x^2 = 0.$$

Solving this quadratic equation yields the limits

$$\frac{x + q^2/2}{n + q^2} \pm \frac{q\sqrt{n}}{n + q^2} \sqrt{\widehat{\pi}_{\text{ML}}(1 - \widehat{\pi}_{\text{ML}}) + \frac{q^2}{4n}}.$$

The center (or midpoint) of the Wilson interval is the relative proportion of successes after adding $q^2/2$ successes and $q^2/2$ non-successes to the sample. It is called a *shrinkage estimator* (Newcombe, 2013).

2.3.3 Wald

Based on the standard normal approximate pivot of the Wald statistic, the limits of the $\gamma \cdot 100\%$ Wald confidence interval (Wald and Wolfowitz, 1939) have the simple form

$$\widehat{\pi}_{\text{ML}} \pm q \cdot \text{se}(\widehat{\pi}_{\text{ML}}) \quad \text{with} \quad \text{se}(\widehat{\pi}_{\text{ML}}) = \sqrt{\frac{\widehat{\pi}_{\text{ML}}(1 - \widehat{\pi}_{\text{ML}})}{n}}.$$

Due to the normal approximation, the Wald interval may fall outside the range $(0, 1)$ for π. This problem is referred to as *overshoot* or *boundary violation* and happens for small or large x. For a 95% interval, overshoot occurs whenever $x = 1$ or $x = 2$, and also when $x = 3$ except when $n < 14$ (Newcombe, 1998). It happens more often for large confidence levels since then, the confidence interval is wider. Any overshoot is truncated to $(0, 1)$, as it is usually done in the literature. Truncation cannot affect coverage properties but limits 0 or 1 are unsatisfactory since they are uninterpretable if $0 < x < n$ (Newcombe, 2013).

Since the standard error is 0 for the extreme cases $x = 0$ ($\widehat{\pi}_{\text{ML}} = 0$) and $x = n$ ($\widehat{\pi}_{\text{ML}} = 1$), the Wald interval is a *degenerate* or *zero width interval* in these cases (for any confidence level). Some modifications to avoid this problem have been proposed in the literature:

1. Use "exact" Clopper-Pearson limits for $x = 0$ and $x = n$ (Vollset, 1993; Pires and Amado, 2008).

2. Calculate a one-sided interval for $x = 0$ and $x = n$ using a standard error that is computed for $x = 0 + 0.5$ and $x = n - 0.5$, respectively (Held and Sabanés Bové, 2020).

3. Add 0.5 successes and non-successes when $x = 0$ or $x = n$ (Rindskopf, 2000). This adjustment increases the sample size by 1 (only) for $x = 0$ and $x = n$.

4. Always add 0.5 successes and non-successes (Rubin and Schenker, 1987; Rindskopf, 2000). This adjustment increases the sample size by 1 for any x. It can be interpreted as a Bayesian interval with Jeffreys' prior (Rubin and Schenker, 1987).

5. Use a shrinkage estimator obtained by adding a pseudo-frequency $\psi > 0$ to the number of successes and to the number of non-successes (Newcombe, 2013). This modification shrinks the MLE towards the center 0.5. The modified interval can be interpreted as a Bayesian interval with a $\text{Be}(\psi, \psi)$ prior. The previous modification chooses $\phi = 0.5$ and the Agresti-Coull interval is obtained for $\phi = 2$.

6. Use a continuity correction by subtracting $1/(2n)$ from the lower limit and adding $1/(2n)$ to the upper limit (Newcombe, 2013; Pires and Amado, 2008). This correction increases the coverage probability but also increases the expected width (by $1/n$) which in turn produces more boundary violation, *e. g.* for $x = 0$ and $x = n$ (Newcombe, 2013).

No modification is really satisfactory (Newcombe, 2011). The first three modifications only adjust the two problematic outcomes which is inconsistent. For example, rather the Clopper-Pearson method should be recommended directly for extreme cases instead of imposing its limits for the Wald interval. The last three modifications are at least consistent for any outcome. Two of them are considered in the present comparison (Rindskopf logit Wald and Agresti-Coull).

Besides truncation, the Wald interval will not be adjusted. The non-adjusted version is called *imputed non-coverage* in Vollset (1993) referring to its degeneracy for $x = 0$ and $x = n$.

2.3.4 Rindskopf (logit Wald with adjustment)

A $\gamma \cdot 100\%$ Wald confidence interval is calculated for the logit transformed parameter

$$\phi = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

with

$$\widehat{\phi}_{\text{ML}} = \log\left(\frac{x+0.5}{n-x+0.5}\right) \quad \text{and} \quad \text{se}(\widehat{\phi}_{\text{ML}}) = \sqrt{\frac{1}{x+0.5} + \frac{1}{n-x+0.5}}.$$

The adjustment of adding 0.5 successes and non-successes ensures that also for the cases $x = 0$ and $x = n$ (with otherwise infinite MLE and standard error) an interval can be computed. Since the scale of ϕ is $(-\infty, \infty)$, this interval is boundary respecting. Back-transformation to the scale of π with the inverse logit function

$$\pi = \text{expit}(\phi) = \frac{\exp(\phi)}{1+\exp(\phi)}$$

yields the Rindskopf confidence interval. It is called Rindskopf here because the used adjustment was suggested in Rindskopf (2000).

2.3.5 Variance-stabilized Wald

A $\gamma \cdot 100\%$ Wald confidence interval is calculated for the variance-stabilizing transformation

$$\phi = \arcsin\left(\sqrt{\pi}\right)$$

with

$$\widehat{\phi}_{\text{ML}} = \arcsin\left(\sqrt{\widehat{\pi}_{\text{ML}}}\right) \quad \text{and} \quad \text{se}(\widehat{\phi}_{\text{ML}}) \approx \frac{1}{\sqrt{4n}}.$$

It is called variance-stabilizing since the variance of $\widehat{\phi}_{\text{ML}}$ is asymptotically independent of the parameter ϕ. Since the scale of ϕ is $(0, \pi/2)$, where π for once means the mathematical constant, this interval may overshoot (*e. g.* for $x = 0$ and $x = n$). In these cases, the interval is truncated to $(0, \pi/2)$. Back-transformation to the scale of π with the inverse function

$$\pi = \sin^2(\phi)$$

yields the variance-stabilized Wald confidence interval.

2.3.6 Agresti-Coull

The limits of the $\gamma \cdot 100\%$ Agresti-Coull confidence interval are

$$\tilde{\pi} \pm q \cdot \sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{n+4}} \quad \text{with} \quad \tilde{\pi} = \frac{x+2}{n+4}.$$

The Agresti-Coull confidence interval was called the "add two successes and two failures" adjusted Wald confidence interval in Agresti and Coull (1998). It was motivated by the finding that for $\gamma = 0.95$, where $q^2 \approx 4$, the midpoint $\tilde{\pi}$ is approximately equal to the midpoint of the Wilson interval. Although this is only true for $\gamma = 0.95$, the same adjustment is used for any confidence level. Moreover, the midpoint $\tilde{\pi}$ is identical to the Bayes estimate (mean of the posterior distribution) for a $\text{Be}(2, 2)$ prior.

2.3.7 Likelihood ratio

The $\gamma \cdot 100\%$ likelihood ratio confidence interval uses the right tail of the approximate pivot for the likelihood ratio statistic to derive the condition

$$-2\log\left\{\frac{L(\pi)}{L(\widehat{\pi}_{\mathrm{ML}})}\right\} \leq \chi^2_\gamma(1), \tag{2.2}$$

where $\chi^2_\gamma(1)$ is the γ quantile of the $\chi^2(1)$ distribution. The confidence interval consists of all parameter values π that satisfy (2.2). It is calculated numerically using `R` function `uniroot`. Only one solution is obtained for the cases $x = 0$ and $x = n$ where the second limit is set to 0, respectively 1.

2.4 Bayesian intervals

In the Bayesian framework, the unknown parameter π is not fixed but assumed to be a random variable with a *prior distribution* with density function $f(\pi)$. Having observed realization x of random variable X with density $f(x \,|\, \pi)$, the density $f(\pi \,|\, x)$ of the *posterior distribution* of π is calculated using Bayes' theorem

$$f(\pi \,|\, x) = \frac{f(x \,|\, \pi) f(\pi)}{f(x)} = \frac{f(x \,|\, \pi) f(\pi)}{\int f(x \,|\, \pi) f(\pi) d\pi}.$$

Beta priors have the appropriate support $(0, 1)$ and are *conjugate* for the binomial proportion. Conjugate means that also the posterior is a beta distribution: The $\mathrm{Be}(\alpha, \beta)$ prior leads to the $\mathrm{Be}(\alpha + x, \beta + n - x)$ posterior.

The posterior distribution can be used to make probability statements about the parameter π. A Bayesian interval estimate for the parameter π is called *credible interval* (abbreviated also by CI with small abuse of notation). A credible interval $[l, u]$ for π with *credibility level* $\gamma \in (0, 1)$ is defined by two quantiles l and u of the posterior distribution that fulfill

$$\int_l^u f(\pi \,|\, x) d\pi = \gamma.$$

Under the assumed prior distribution, the random variable $\pi \,|\, x$ is contained in a $\gamma \cdot 100\%$ credible interval with probability γ (in contrast to frequentist inference). Bayesian intervals would have exact mean coverage probability equal to γ under the specified prior distribution (Pires and Amado, 2008). However, the frequentist setting does not assume a prior distribution but intervals using the following two priors are known to have favourable properties when viewed as frequentist intervals (Newcombe, 2013, p. 20).

1. Using the *uniform* prior $\pi \sim \mathrm{U}(0, 1) = \mathrm{Be}(1, 1)$, the posterior distribution is

   $$\pi \,|\, x \sim \mathrm{Be}(1 + x, 1 + n - x).$$

   This prior is non-informative (on the scale of π) since any value between 0 and 1 is equally likely. However, it would not be uniform anymore for nonlinear transformations of π.

2. Using the *Jeffreys* prior $\pi \sim \mathrm{Be}(1/2, 1/2)$, the posterior distribution is

   $$\pi \,|\, x \sim \mathrm{Be}(1/2 + x, 1/2 + n - x).$$

   The Jeffreys prior is defined as the prior that is proportional to $\sqrt{J(\pi)}$ and it is invariant under reparametrization. That means the prior of a transformed parameter is still a

Jeffreys prior. Held and Sabanés Bové (2020, p. 186–187) provides an argument why this is a non-informative prior: In a frequentist setting, the average information about π in the data is measured by $J(\pi)$. In this sense, non-informative would mean that $J(\pi)$ is independent of π and one could choose a uniform prior for π. However, if $J(\pi)$ does depend on π, one should first remove the dependence by applying the variance-stabilizing transformation $\phi = \arcsin(\sqrt{\pi})$. The Jeffreys prior for π is the uniform prior for ϕ.

2.4.1 Equal-tailed

The limits of the $\gamma \cdot 100\%$ equal-tailed credible interval are the $(1-\gamma)/2$ and $(1+\gamma)/2$ quantiles of the posterior distribution. A probability mass of $\alpha/2$ is cut off from both tails of the posterior distribution. For the two priors, the limits are:

1. Uniform prior: $b_{(1-\gamma)/2}(1+x, 1+n-x)$ and $b_{(1+\gamma)/2}(1+x, 1+n-x)$.
2. Jeffreys prior: $b_{(1-\gamma)/2}(1/2+x, 1/2+n-x)$ and $b_{(1+\gamma)/2}(1/2+x, 1/2+n-x)$.

The Clopper-Pearson confidence interval limits can be interpreted in a Bayesian way with two different priors. The lower limit corresponds to the lower limit of an equal-tailed interval with an improper $\text{Be}(0, 1)$ prior, while the upper limit corresponds to the upper limit of an equal-tailed interval with an improper $\text{Be}(1, 0)$ prior.

2.4.2 Highest posterior density

The highest posterior density (HPD) interval is the unique (for the chosen prior) $\gamma \cdot 100\%$ credible interval $[l, u]$ that fulfills the condition

$$f(\pi \,|\, x) \geq f(\tilde{\pi} \,|\, x)$$

for all $\pi \in [l, u]$ and all $\tilde{\pi} \notin [l, u]$. It consists of all the parameter values with the highest posterior density until they reach a probability mass of γ. This is the smallest interval that has mean coverage probability γ under the specified prior.

The HPD interval is calculated numerically except for the two extreme cases. Since the posterior density is monotone decreasing for $x = 0$, the lower limit is 0 and the upper limit is the γ quantile of the posterior distribution. Since the posterior density is monotone increasing for $x = n$, the upper limit is 1 and the lower limit is the $1-\gamma$ quantile of the posterior distribution.

2.5 Interval evaluation methods

The most commonly used evaluation methods for confidence intervals are the coverage probability and the expected width as for example in Vollset (1993); Brown *et al.* (2001); Held and Sabanés Bové (2020); Pires and Amado (2008). Common summary measures are the minimum or mean coverage probabilities used in Newcombe (1998) and Pires and Amado (2008). The (expected) interval score that has been used for prediction intervals is suggested in Gneiting and Raftery (2007) as a new method to evaluate (central) confidence intervals.

2.5.1 Coverage probability

The *coverage* of a confidence interval $[l, u]$ for a true proportion π is

$$\text{C}(l, u, \pi) = \mathbb{1}(l \leq \pi \leq u),$$

where $\mathbb{1}$ denotes the indicator function. So, the coverage is 1 if the confidence interval contains the true proportion and 0 otherwise.

The *coverage probability* is the probability with which a confidence interval contains the true proportion. The confidence level γ that is associated with the confidence interval is called the *nominal* coverage probability. It is the coverage probability that the interval should theoretically attain. However, empirically, the *actual* coverage probability often deviates from the nominal coverage probability. Since a binomial sample $X \sim \text{Bin}(n, \pi)$ only has a finite number of possible outcomes, the actual coverage probability can be calculated analytically (Held and Sabanés Bové, 2020; Pires and Amado, 2008):

$$\begin{aligned} \text{CP}(\pi) &= \sum_{x=0}^{n} \mathsf{Pr}(X = x)\, \text{C}\{l(x), u(x), \pi\} \\ &= \sum_{x=0}^{n} \binom{n}{x} \pi^x (1-\pi)^{n-x}\, \text{C}\{l(x), u(x), \pi\}, \end{aligned} \tag{2.3}$$

where the limits l and u depend on a realization x from $X \sim \text{Bin}(n, \pi)$.

Note that the the limits depend on n and γ as well. This dependence does not need to be made explicit in (2.3) as long as both are assumed to be known and fixed. For $\gamma = 0.95$ and $n = 50$, Figure 2.1 shows the (actual) coverage probabilities of the discussed confidence and credible intervals (only for the Jeffreys prior) as a function of the true proportion π. It also shows smoothed coverage probabilities (in black) that are computed using a specific kernel function as described in Bayarri and Berger (2004) for a smoothing parameter $\varepsilon = 0.025$. Only the first half of the values for π is displayed since coverage probabilities (also expected widths and interval scores) are symmetric around 0.5 due to the equivariance of lower and upper bounds of the confidence intervals (Gillibert *et al.*, 2021). This figure reproduces Figures 4.9 and 6.7 in Held and Sabanés Bové (2020), where it should be mentioned that the boundary cases $x = 0$ and $x = n$ for the Wald type confidence intervals are handled differently.

2.5.2 Expected width

The *width* of a confidence interval $[l, u]$ for a true proportion π is

$$\text{W}(l, u) = u - l$$

and the *expected width* (expectation w.r.t. the distribution of the data X) is (Pires and Amado, 2008)

$$\text{EW}(\pi) = \sum_{x=0}^{n} \binom{n}{x} \pi^x (1-\pi)^{n-x}\, \text{W}\{l(x), u(x)\}. \tag{2.4}$$

The limits l and u depend on a realization x, while n and γ are assumed to be known and fixed. For $\gamma = 0.95$ and $n = 50$, Figure 2.1 shows the expected widths of the discussed confidence and credible intervals (only for the Jeffreys prior) as a function of the true proportion. Figure 4.10 in Held and Sabanés Bové (2020) displays the widths of the frequentist intervals except Agresti-Coull as a function of x.

2.5.3 Expected interval score

The *interval score* is introduced in Gneiting and Raftery (2007, Section 6, p. 12) as a special case of a more general proper scoring rule for predictive quantiles. This subsection summarizes how it is developed. The same notation is used, except that the variable x is replaced by y since x already denotes the number of successes in a binomial experiment.

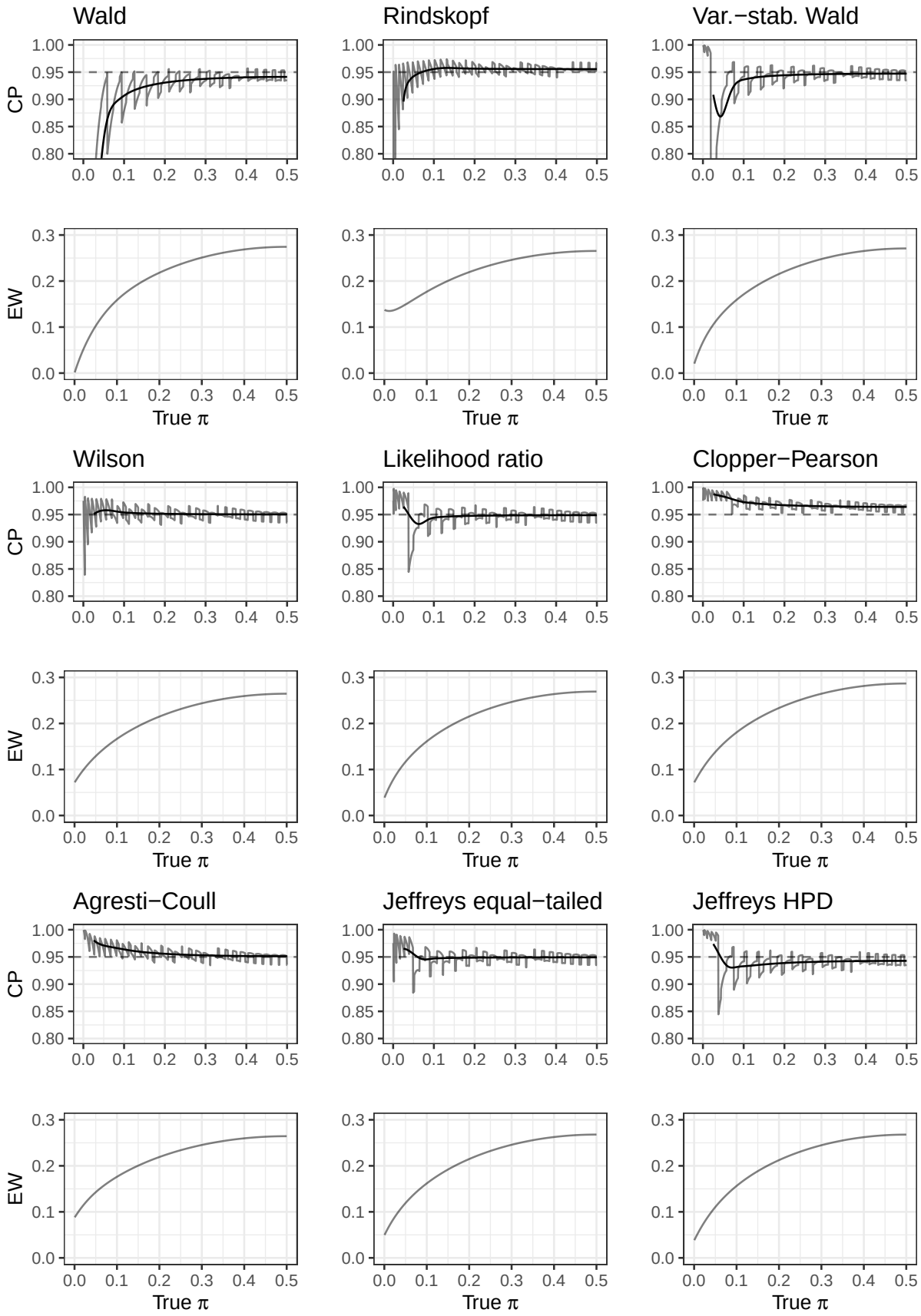

Figure 2.1: Coverage probability and expected width for each confidence interval as a function of π for $n = 50$ and $\gamma = 0.95$. The locally smoothed coverage probability is added in black.

Let $\alpha_1, \ldots, \alpha_k$ be the levels of the sought quantiles $q_1, \ldots, q_k$, let $r_1, \ldots, r_k$ be the predicted quantiles and let y be the true outcome. The rewarded score is denoted by $S(r_1, \ldots, r_k; y)$ and the expected score under a probability measure P is defined by

$$S(r_1, \ldots, r_k; P) = \int S(r_1, \ldots, r_k; y) dP(y).$$

Let $q_1, \ldots, q_k$ be the true quantiles for P at levels $\alpha_1, \ldots, \alpha_k$. Following Cervera and Muñoz (1996), the scoring rule S is said to be *proper* if for any real numbers $r_1, \ldots, r_k$ and for any probability measure P,

$$S(q_1, \ldots, q_k; P) \geq S(r_1, \ldots, r_k; P).$$

This means that the expected score of a proper scoring rule is maximized under the true quantiles. Hence, predictions with larger scores are better.

First, scoring rules that are proper for the prediction of a single quantile are presented (the proof presented here is essentially the same but contains more steps).

Theorem 2.1. If s is non-decreasing and h is arbitrary, then the scoring rule

$$S(r; y) = \alpha s(r) + \{s(y) - s(r)\} \mathbb{1}(y \leq r) + h(y)$$

is proper for predicting the quantile at level $\alpha \in (0, 1)$.

Proof. Let P be any probability measure, F_P be the associated distribution function and q be the true quantile for P at level α, *i. e.* $F_P(q) = \alpha$. Without loss of generality, assume that $r < q$ (the case $r > q$ is analogous). Then,

$$\begin{aligned}
S(q; P) - S(r; P) &= \int_{-\infty}^{\infty} \Big[\alpha s(q) + \{s(y) - s(q)\} \mathbb{1}(y \leq q) + h(y)\Big] dP(y) \\
&\quad - \int_{-\infty}^{\infty} \Big[\alpha s(r) + \{s(y) - s(r)\} \mathbb{1}(y \leq r) + h(y)\Big] dP(y) \\
&= \alpha s(q) \cdot 1 + \int_{-\infty}^{q} s(y) dP(y) - s(q) \cdot F_P(q) + \int_{-\infty}^{\infty} h(y) dP(y) \\
&\quad - \left[\alpha s(r) \cdot 1 + \int_{-\infty}^{r} s(y) dP(y) - s(r) \cdot F_P(r) + \int_{-\infty}^{\infty} h(y) dP(y)\right] \\
&= \alpha s(q) + \int_{-\infty}^{q} s(y) dP(y) - s(q)\alpha \\
&\quad - \alpha s(r) - \int_{-\infty}^{r} s(y) dP(y) + s(r) F_P(r) \\
&= \int_{r}^{q} s(y) dP(y) - \alpha s(r) + s(r) F_P(r) \\
&\geq s(r)\{F_P(q) - F_P(r)\} + s(r)\{F_P(r) - \alpha\} \\
&= s(r)\{\alpha - F_P(r)\} + s(r)\{F_P(r) - \alpha\} \\
&= 0,
\end{aligned}$$

where for the inequality, it is needed that s is non-decreasing. Hence, $S(q; P) \geq S(r; P)$ for any r and P and therefore S is a proper scoring rule. □

Then, from Theorem 2.1, proper scoring rules for the prediction of multiple quantiles are deduced.

Corollary 2.1. If s_i is non-decreasing for $i = 1, \ldots, k$ and h is arbitrary, then the scoring rule

$$S(r_1, \ldots, r_k; y) = \sum_{i=1}^{k} \Big[\alpha_i s_i(r_i) + \{ s_i(y) - s_i(r_i) \} \mathbb{1}(y \leq r_i) \Big] + h(y) \tag{2.5}$$

is proper for predicting the quantiles at levels $\alpha_1, \ldots, \alpha_k \in (0, 1)$.

The scoring rule from Corollary 2.1 can be viewed as the sum of k scoring rules for a single quantile. Since the sum of two proper scoring rules is again a proper scoring rule, Corollary 2.1 follows directly from Theorem 2.1 (this explanation is not provided in Gneiting and Raftery (2007)).

Now, interval forecasts are a special case of quantile forecasts. A *central* $(1 - \alpha) \cdot 100\%$ prediction interval is defined such that its lower and upper limits are the predictive quantiles at level $\alpha/2$ and $1 - \alpha/2$. It is called *central* because the two tail probabilities are equal to $\alpha/2$. Let (l, u) be a central $(1 - \alpha) \cdot 100\%$ prediction interval. A proper scoring rule of the form (2.5) for predicting this interval fixes $\alpha_1 = \alpha/2$ and $\alpha_2 = 1 - \alpha/2$. By choosing $s_1(y) = s_2(y) = 2y/\alpha$ and $h(y) = -2y/\alpha$, Equation (2.5) yields

$$\begin{aligned} S(l, u; y) &= \frac{\alpha}{2} \frac{2}{\alpha} l + \frac{2}{\alpha}(y - l) \mathbb{1}(y \leq l) + \left(1 - \frac{\alpha}{2}\right) \frac{2}{\alpha} u + \frac{2}{\alpha}(y - u) \mathbb{1}(y \leq u) - \frac{2}{\alpha} y \\ &= l + \frac{2}{\alpha}(y - l) \mathbb{1}(y \leq l) + \frac{2}{\alpha} u - u + \frac{2}{\alpha}(y - u) \mathbb{1}(y \leq u) - \frac{2}{\alpha} y \\ &= l - u + \frac{2}{\alpha}(y - l) \mathbb{1}(y \leq l) + \frac{2}{\alpha}(u - y) \mathbb{1}(y > u). \end{aligned}$$

The *interval score* is obtained for this choice of s_1 and s_2 by additionally reverting the sign:

$$\begin{aligned} \mathrm{IS}_\alpha(l, u, y) &= (u - l) + \frac{2}{\alpha}(l - y) \mathbb{1}(y < l) + \frac{2}{\alpha}(y - u) \mathbb{1}(y > u) \\ &= \mathrm{W}(l, u) + \frac{2}{\alpha} \min(|y - l|, |y - u|) \Big[1 - \mathrm{C}(l, u, y) \Big]. \end{aligned} \tag{2.6}$$

Note that the probability of the event $y = l$ is almost surely 0, which is why the two quantities $\mathbb{1}(y < l)$ and $\mathbb{1}(y \leq l)$ are interchangeable. The second line of (2.6) expresses the interval score in terms of width and coverage. Width and coverage are combined into a negatively oriented score, meaning that lower scores are better. It consists of two penalties: the width where a larger interval is a larger penalty and non-coverage weighted by the minimal distance of the true observation to the interval and by $2/\alpha$.

In a binomial experiment $X \sim \mathrm{Bin}(n, \pi)$ (now $y = \pi$ is the true proportion), the *expected interval score* of a $\gamma \cdot 100\%$ confidence interval is

$$\mathrm{EIS}_\alpha(\pi) = \mathrm{EW}(\pi) + \frac{2}{\alpha} \sum_{x=0}^{n} \binom{n}{x} \pi^x (1 - \pi)^{n-x} \min\{|\pi - l(x)|, |\pi - u(x)|\} \Big[1 - \mathrm{C}\{l(x), u(x), \pi\} \Big].$$

The interval score is proposed as a loss function for optimum score interval estimation. It is a proper score in the sense that the score is proper in the setting of probabilistic predictions (more details in Chapter 4). However, this only applies to central interval estimators meaning $\gamma \cdot 100\%$ confidence intervals with lower and upper *exceedance probability* $\alpha/2$ (Gneiting and Raftery, 2007, Subsection 9.3). The exceedance probability is called left and right *non-coverage probability* in Newcombe (1998). For non-central intervals, the interval score is not a proper scoring rule (Brehmer and Gneiting, 2021).

2.5.4 Generalized interval score

Since a confidence interval is not necessarily central, it would be useful to have a (proper) score that incorporates the two possibly non-central tails. As part of this master thesis, a *generalized interval score* has been developed that simplifies to the interval score in the special case of central intervals. As the interval score, it is derived in the prediction setting.

Let α_1 and α_2 be the levels of two quantiles such that $\alpha_1 + (1 - \alpha_2) = \alpha$. Note that α_1 and $1-\alpha_2$ are the two tail probabilities. They define a (possibly non-central) $(1-\alpha)\cdot 100\%$ prediction interval. Choosing the two non-decreasing functions $s_1(y) = y/\alpha_1$ and $s_2(y) = y/(1-\alpha_2)$ and $h(y) = -s_2(y)$, Equation (2.5) yields

$$\begin{aligned} S(l,u;y) &= \alpha_1 \frac{1}{\alpha_1} l + \frac{1}{\alpha_1}(y-l)\,\mathbb{1}(y \le l) + \alpha_2 \frac{1}{1-\alpha_2} u + \frac{1}{1-\alpha_2}(y-u)\,\mathbb{1}(y \le u) - \frac{1}{1-\alpha_2} y \\ &= l + \frac{1}{\alpha_1}(y-l)\,\mathbb{1}(y \le l) + \frac{1-(1-\alpha_2)}{1-\alpha_2} u + \frac{1}{1-\alpha_2}(y-u)\,\mathbb{1}(y \le u) - \frac{1}{1-\alpha_2} y \\ &= l - u + \frac{1}{\alpha_1}(y-l)\,\mathbb{1}(y \le l) + \frac{1}{1-\alpha_2}(u-y)\,\mathbb{1}(y > u). \end{aligned}$$

Reverting the sign yields the *generalized interval score*:

$$\mathrm{GIS}_{\alpha_1,\alpha_2}(l,u,y) = (u-l) + \frac{1}{\alpha_1}(l-y)\,\mathbb{1}(y < l) + \frac{1}{1-\alpha_2}(y-u)\,\mathbb{1}(y > u). \tag{2.7}$$

For a central prediction interval with $\alpha_1 = \alpha/2$ and $\alpha_2 = 1 - \alpha/2$, expression (2.7) is equal to the interval score (2.6). So, in theory it is possible to extend the interval score to non-central intervals. However, the definition of the score would depend on the tail probabilities. Consequently, prediction intervals with different tail probabilities would be assessed with different scores, hence the score would not be proper.

2.5.5 Weighted interval score

The *weighted interval score* defined in Bracher *et al.* (2021) assesses central prediction intervals for multiple levels in terms of the weighted sum of the interval scores:

$$\mathrm{WIS}_{\alpha_{0:K}}(l_{1:K}, u_{1:K}, y) = \frac{1}{K + 1/2}\left\{ w_0\,|y - m| + \sum_{k=1}^{K} w_k\,\mathrm{IS}_{\alpha_k}(l_k, u_k, y) \right\},$$

for levels $1-\alpha_0, \ldots, 1-\alpha_K$, non-negative weights $w_0, \ldots, w_K$ (such that it is proper) and predictive median m (corresponding to $\alpha_0 = 1$). For $w_k = \alpha_k/2$, large K and equally spaced values $\alpha_1, \ldots, \alpha_K$ (covering the whole unit interval), it is approximately equal to the continuous ranked probability score (CRPS). The reason for the factor $1/(K + 1/2)$ is probably that for this particular choice of weights

$$\sum_{k=1}^{K} w_k \frac{2}{\alpha_k} + w_0 = \sum_{k=1}^{K} \frac{\alpha_k}{2}\frac{2}{\alpha_k} + \frac{1}{2} = K + \frac{1}{2},$$

which are the summed up weights for the non-coverage part of the interval scores where $2/\alpha_k$ is integrated in the weights as well.

The intuition behind the weighted interval score is to describe the predictive distribution using many central intervals. The same could be done for the posterior distribution in the Bayesian setting. However, for fixed parameters in the frequentist setting, only large confidence levels are useful. In practice, often levels 0.9, 0.95 or 0.99 are used ($\alpha_1 = 0.1$, $\alpha_2 = 0.05$ and

$\alpha_3 = 0.01$). Since the nice connection to the CRPS (probably) does not hold considering only large levels, a simpler version of the weighted interval score, with all weights equal to 1 and without the constant factor $1/(K + 1/2)$, will be used for those levels:

$$\mathrm{WIS}_{\alpha_{1:3}}(l_{1:3}, u_{1:3}, y) = \sum_{k=1}^{3} \mathrm{IS}_{\alpha_k}(l_k, u_k, y).$$

2.5.6 Summary measures

The coverage probability, expected width and expected interval score can only be computed for a fixed true proportion π (see *e. g.* Figure 2.1). One way to summarize these measures over different values for π is to integrate the functions over π. Like this, for any n and γ, one number can be compared between different methods. The integral over all possible true proportions can be regarded as a global average as compared to local averages for the coverage probability in Bayarri and Berger (2004). It is also closely related to the concept of *integrated risk* in Bayesian decision theory (Robert, 2007, p. 62–63), which integrates the frequentist risk (*e. g.* the expected interval score) over π with respect to the prior distribution of π. Note that the integral considered here would be the integrated risk for the uniform prior.

Another issue is the poor performance of some methods for observations x near the boundaries. An idea how to give more weight to rare cases is to integrate on the variance-stabilized scale. That means instead of integrating a function $f(\pi)$ from 0 to 1, the function $f(\sin^2(\phi))$ is integrated from 0 to $\pi/2$ (here, π means the mathematical constant), where $\phi = \arcsin\sqrt{\pi}$ is the transformed parameter on the variance-stabilized scale. It holds that

$$f(\pi) = f\{\sin^2(\arcsin\sqrt{\pi})\} = f\{\sin^2(\phi)\}.$$

Figure 2.2 illustrates how the integral on the usual scale without transformation compares to the integral on the scale after a variance-stabilizing transformation. It is illustrated exemplary for the expected width of a 95% Wald confidence interval with $n = 50$. Note that the x-axis for the transformed integral is shifted to the left.

The variance-stabilizing transformation is symmetric around 0.5 and transforms the unit interval of true proportions π in a way that stretches the boundaries compared to the middle part, as can be seen in Figure 2.3. Hence, integration over π on the variance-stabilized scale gives indeed more weight to the boundaries. This is even more evident looking at the weight function of this transformation which will now be derived using an arbitrary example.

The variance-stabilizing transformation maps the interval $[0.01, 0.02]$ to the larger interval $[0.1, 0.14]$. This transformed interval is weighted by its relative length compared to the original interval:

$$\frac{\arcsin\left(\sqrt{0.02}\right) - \arcsin\left(\sqrt{0.01}\right)}{0.02 - 0.01} \approx 4.17.$$

Letting the length of the interval become smaller corresponds to the derivative of the transformation function. So, the weight function, which gives a weight to each point in the unit interval, is the first derivative of the transformation function,

$$\frac{d}{d\pi} \arcsin\left(\sqrt{\pi}\right) = \frac{1}{2\sqrt{\pi(1-\pi)}},$$

which is valid for $0 < \pi < 1$. Up to a constant factor $2/\pi$ (where π means the mathematical constant), this is the density of the Jeffreys prior (Be$(1/2, 1/2)$). This relation was already touched upon in the discussion of the Jeffreys prior in Section 2.4. The weight density is displayed in Figure 2.3. Consequently, the integral on the variance-stabilized scale is the integrated risk for the Jeffreys prior (up to the constant factor $2/\pi$).

Figure 2.2: Integral of expected width of the 95% Wald confidence interval for $n = 50$ on the scale of $\pi \in (0, 1)$ (blue) and on the variance-stabilized scale of $\phi = \arcsin(\sqrt{\pi}) \in (0, \pi/2)$ (red).

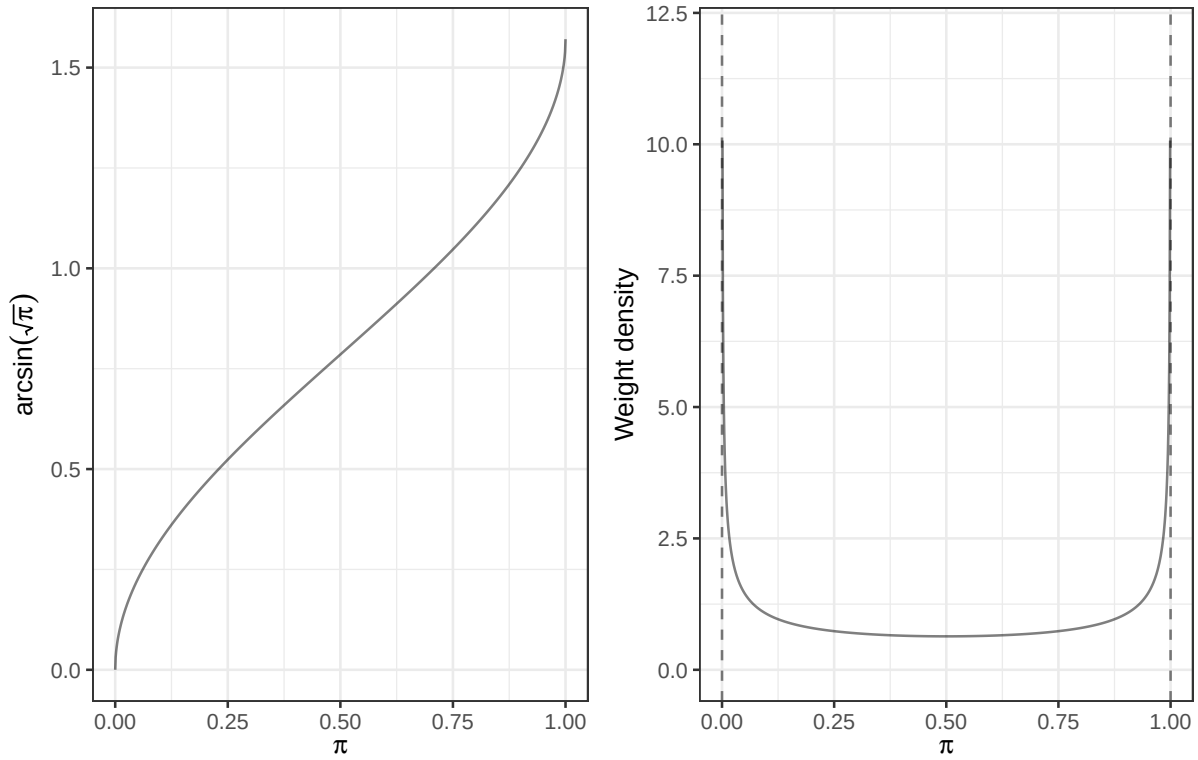

Figure 2.3: Variance-stabilizing transformation (left) and corresponding weight density (right).

2.6 Asymptotics

The coverage probability is asymptotically equal to the nominal coverage probability γ. Similarly, asymptotical references (under normality) can be derived for the expected width and the expected interval score. They are derived because they will be used as a reference in Chapter 3.

Using the normal approximation of the binomial distribution $X \sim \text{Bin}(n, \pi)$ by the central limit theorem, $X/n \overset{a}{\sim} \text{N}(\pi, \pi(1-\pi)/n)$, the asymptotical reference is the expected interval score of the $\gamma \cdot 100\%$ confidence interval with limits

$$L = \widehat{\pi}_{\text{ML}} - q \cdot \sigma \quad \text{and} \quad U = \widehat{\pi}_{\text{ML}} + q \cdot \sigma$$

under normality of X/n with known standard deviation $\sigma = \sqrt{\pi(1-\pi)/n}$. The interval score of this confidence interval is

$$\text{IS}_\alpha(L, U, \pi) = 2q\sigma + \frac{2}{\alpha}(L - \pi)\,\mathbb{1}(L > \pi) + \frac{2}{\alpha}(\pi - U)\,\mathbb{1}(U < \pi)$$

and, by linearity, the expected interval score is

$$\mathsf{E}\{\text{IS}_\alpha(L, U, \pi)\} = 2q\sigma + \frac{2}{\alpha}\,\mathsf{E}\{(L - \pi)\,\mathbb{1}(L - \pi > 0)\} + \frac{2}{\alpha}\,\mathsf{E}\{(\pi - U)\,\mathbb{1}(\pi - U > 0)\}.$$

Under the (asymptotical) normal distribution of X/n, both $L - \pi$ and $\pi - U$ have a normal distribution $\text{N}(-q\sigma, \sigma^2)$. Defining $Y \sim \text{N}(-q\sigma, \sigma^2)$, the asymptotical value of the expected interval score is

$$\begin{aligned}
\mathsf{E}\{\text{IS}_\alpha(L, U, \pi)\} &= 2q\sigma + 2\frac{2}{\alpha}\,\mathsf{E}\{Y\,\mathbb{1}(Y > 0)\} \\
&= 2q\sigma + 2\frac{2}{\alpha}\,\mathsf{E}\{Y \mid Y > 0\}\,\mathsf{Pr}(Y > 0) \\
&= 2q\sigma + 2\,\mathsf{E}\{Y \mid Y > 0\} \\
&= 2q\sigma + 2\left(-q\sigma + \frac{\sigma\phi(q)}{1 - \Phi(q)}\right) \\
&= \frac{2\sigma\phi(q)}{1 - \Phi(q)},
\end{aligned}$$

where ϕ and Φ are the density and distribution functions of a standard normal. The derivation uses conditional expectations and that $\mathsf{Pr}(Y > 0)$ is equal to $\alpha/2$. The expectation of a truncated normal random variable $Y \mid Y > 0$ is computed according to Johnson *et al.* (1994). In particular, the asymptotical value of the expected width is

$$\mathsf{E}\{\text{W}(L, U)\} = 2q\sigma.$$

The asymptotical references of the expected interval score and the expected width are concave curves as a function of the true proportion π, as can be seen in Figure 2.4 in an example with $n = 50$ and $\gamma = 0.95\%$.

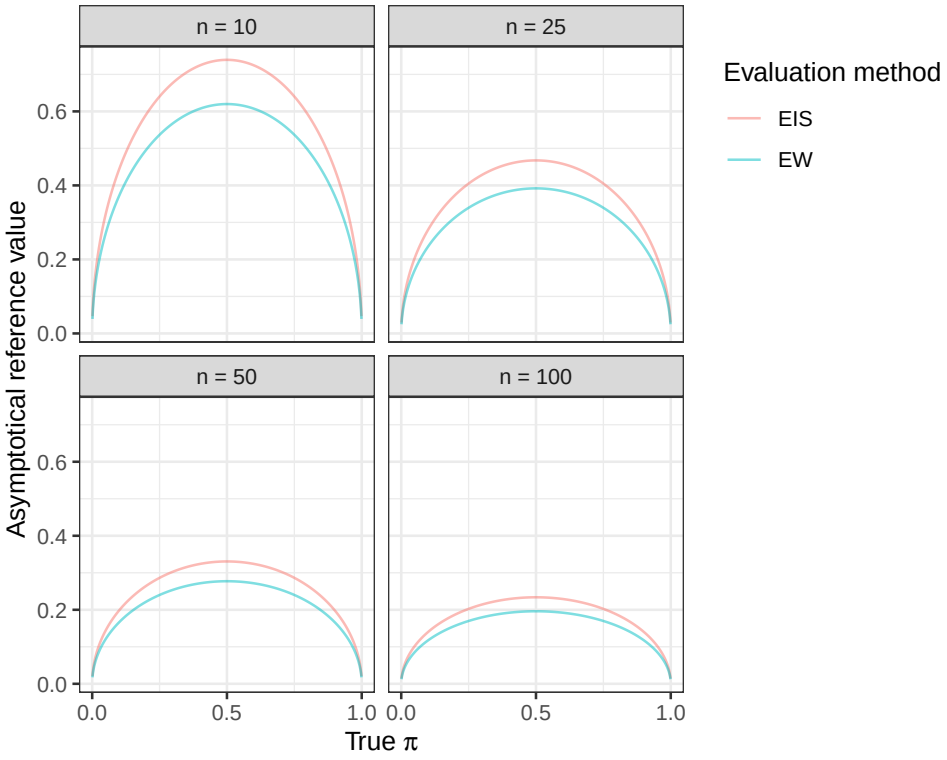

Figure 2.4: Asymptotical expected interval score and asymptotical expected width of a confidence interval under asymptotical normality of a binomial sample as a function of π for $\gamma = 0.95$ and $n \in \{10, 25, 50, 100\}$.

Chapter 3

Results

In this chapter, confidence intervals are compared using the coverage probability, the expected width and the expected interval score (also weighted and generalized versions). Two of them are in fact credible intervals (using non-informative priors) but are viewed as frequentist intervals and also assessed with these frequentist measures.

3.1 Central intervals

A confidence interval is central if the lower and upper exceedance probabilities are equal. For the Bayesian intervals, the equal-tailed interval is clearly central and the HPD interval is not. Now, confidence intervals based on normal approximation are only asymptotically central. Those are the Wilson, Wald, Rindskopf, variance-stabilized Wald and Agresti-Coull intervals. Such confidence intervals are also considered as central in Pires and Amado (2008). The likelihood ratio confidence interval is also asymptotically central. While it is constructed only from one tail of a $\chi^2(1)$ distribution, it is equivalent to the signed likelihood ratio statistic which is asymptotically standard normal (Held and Sabanés Bové, 2020, p. 106). For the conservative Clopper-Pearson confidence interval, the two exceedance probabilities are not both equal to $\alpha/2$ but are both at least $\alpha/2$. In some sense, this is a "central" constraint and asymptotically, it should be less conservative. So, all considered frequentist intervals are intended to be central but it only holds asymptotically.

It is necessary that the confidence intervals are central for the interval score to be a proper scoring rule. As proposed in Gneiting and Raftery (2007), only confidence intervals with the same nominal coverage are compared. Moreover, all intervals except the HPD interval are intended to be central and interpreted as such. However, a caveat should be kept in mind that all intervals except the equal-tailed interval are not central in a rigorous way.

3.2 Coverage

Figure 3.1 compares the smoothed coverage probabilities (as in Bayarri and Berger (2004) with smoothing parameter $\varepsilon = 0.025$) of 95% confidence intervals of binomial samples for different sample sizes n as a function of the true proportion π. Since they are symmetric around 0.5, only the first half of the values for π are shown. The nominal coverage is marked as a dashed line. One can see that for increasing n, the coverage probabilities converge to the nominal coverage. The coverage probability of the Wald CI decreases a lot more at the boundaries $\pi = 0$ and $\pi = 1$ than is visible in the plots (below 30% for $n = 10$ and below 65% for $n = 50$). In order to see the differences of the confidence intervals better, these parts of the Wald CI have been cut off.

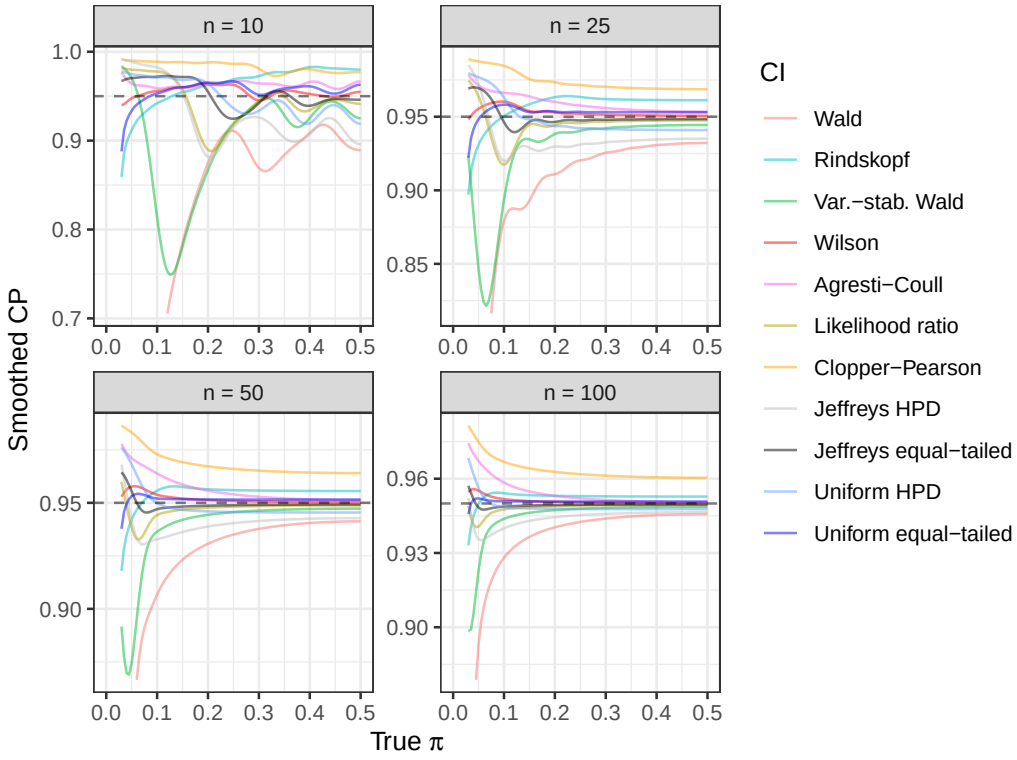

Figure 3.1: Smoothed coverage probability for each method as a function of π for $\gamma = 0.95$ and $n \in \{10, 25, 50, 100\}$.

The Wald CI can be identified as the one with the worst coverage probability. It is practically always smaller than all the other CIs. In particular, it is always smaller than the nominal coverage (it approaches the nominal coverage from below). Also the variance-stabilized Wald CI drops dramatically near the boundary. Clearly, the Clopper-Pearson CI is the most conservative. Also the Agresti-Coull CI is conservative. The Rindskopf CI is conservative except for values of π near the boundary. The Wilson and the uniform equal-tailed CIs perform well. Only at the boundaries, other methods like the uniform HPD, the Jefreys HPD and equal-tailed and the likelihood ratio CIs peform better (in decreasing order).

Figure 3.2 compares smoothed coverage probabilities for fixed sample size $n = 50$ but different confidence levels γ. For increasing confidence level, there are less differences between the methods. Overall, the comparison between the confidence intervals does essentially not change for different confidence levels, except for the ranking of the Agresti-Coull CI.

3.3 Width

Figure 3.3 compares the expected widths of 95% confidence intervals of binomial samples for different sample sizes n as a function of the true proportion π (symmetric around $\pi = 0.5$). What is shown is the difference between the expected width and the asymptotical reference value of the expected width (under normality). This is done because otherwise the differences between the CIs would not be visible due to the curvature of the expected widths as a function of π (see Figure 2.1). By taking the difference to the asymptotical reference curve, the curvature can be removed. For a fixed π, the differences between the CIs become smaller for increasing n. Otherwise, the comparison is similar for all n.

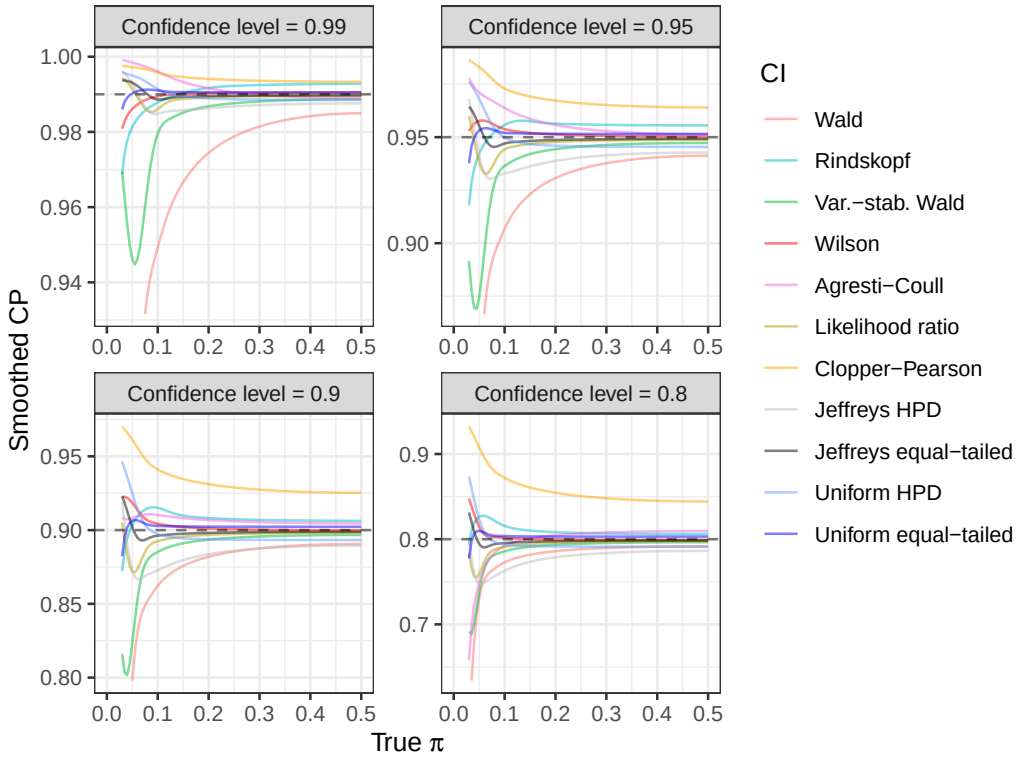

Figure 3.2: Smoothed coverage probability for each method as a function of π for $n = 50$ and $\gamma \in \{0.99, 0.95, 0.9, 0.8\}$.

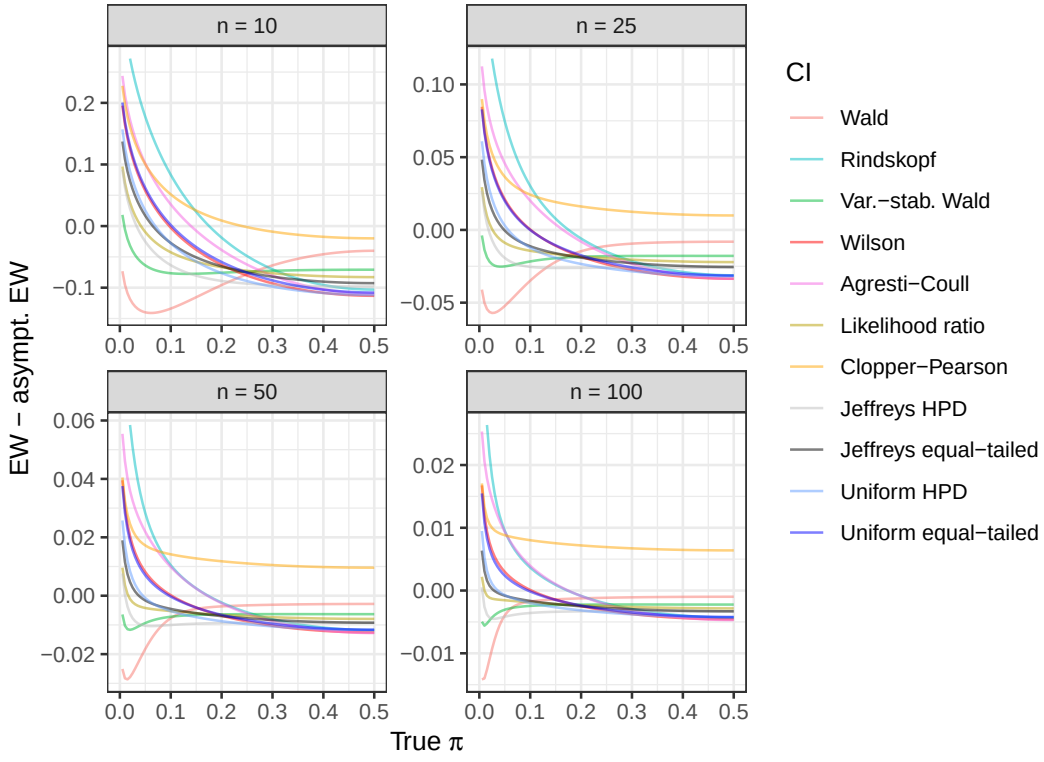

Figure 3.3: Difference between expected width and asymptotical EW for each method as a function of π for $\gamma = 0.95$ and $n \in \{10, 25, 50, 100\}$.

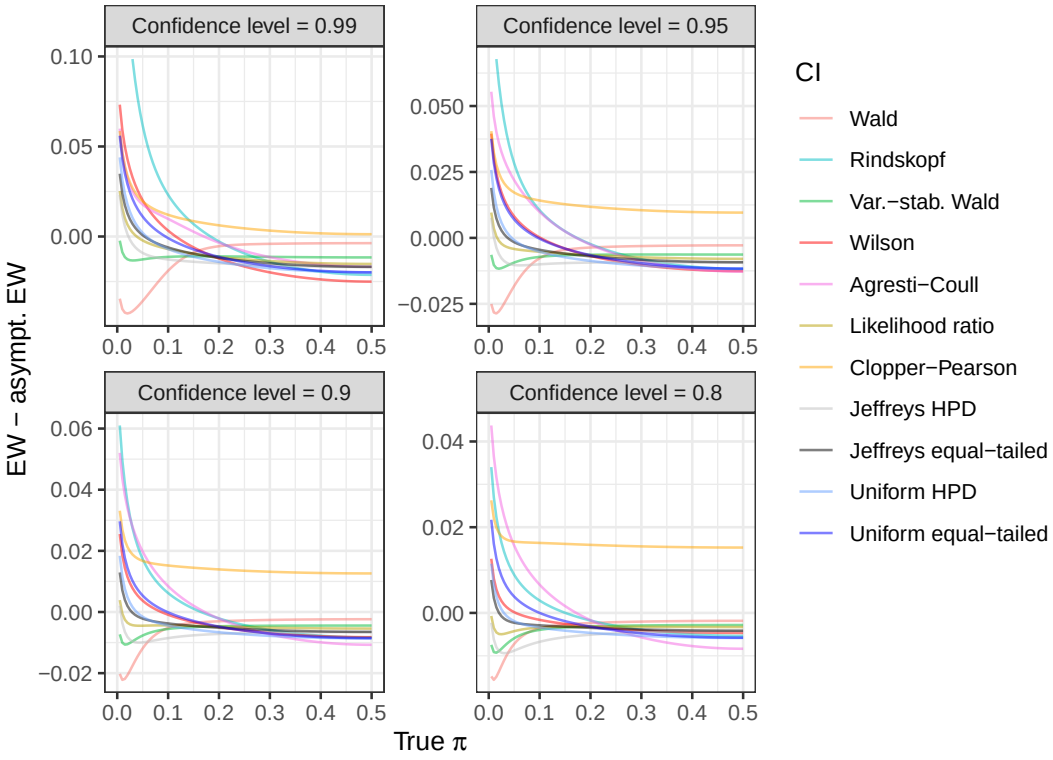

Figure 3.4: Difference between expected width and asymptotical EW for each method as a function of π for $n = 50$ and $\gamma \in \{0.99, 0.95, 0.9, 0.8\}$.

The Wald and the variance-stabilized Wald CIs are the smallest if π is near the boundaries. However, as seen in the previous section, this comes at the cost of poor coverage probability. As expected, the conservative Clopper-Pearson CI is the largest. The HPD intervals are always smaller than the equal-tailed intervals for the same prior. Conversely, the HPD intervals have too low coverage compared to the equal-tailed intervals. The Agresti-Coull CI is generally smaller than the Rindskopf CI and the Wilson CI is generally smaller than those two. The likelihood ratio interval performs better if π is near the boundary but worse otherwise.

For different confidence levels with $n = 50$, the overall results do not change except for the ranking of the Agresti-Coull CI, as can be seen in Figure 3.4. For increasing confidence level, the CIs and also the differences between the CIs become larger for a fixed π.

3.4 Interval score

Figure 3.5 compares the expected interval scores of 95% confidence intervals of binomial samples for different sample sizes n as a function of the true proportion π (symmetric around 0.5). For the same reason as for the expected width, the difference to the asymptotical reference is used to compare the CIs. Lower scores are better. The values for the Wald CI would be even larger but to see the differences between the CIs better, these parts have been cut off in the plots.

The Wald CI is also the worst CI in terms of the expected interval score which combines width and coverage. The variance-stabilized Wald CI performs equally bad except for the boundaries of π where it has the best score. It is the other way around for the Rindskopf and the Agresti-Coull CIs which are the best for $\pi = 0.5$ but the worst or second worst for π near the boundaries. Similarly, the HPD intervals are better than the equal-tailed intervals for $\pi = 0.5$ but worse at

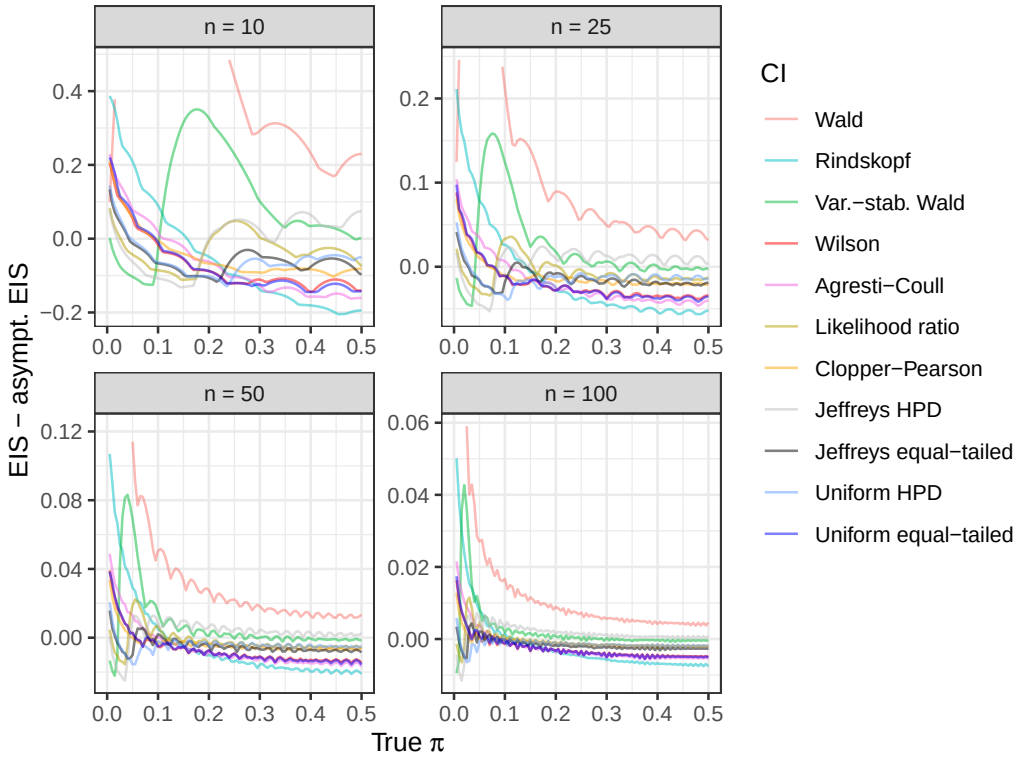

Figure 3.5: Difference between expected interval score and asymptotical EIS for each method as a function of π for $\gamma = 0.95$ and $n \in \{10, 25, 50, 100\}$.

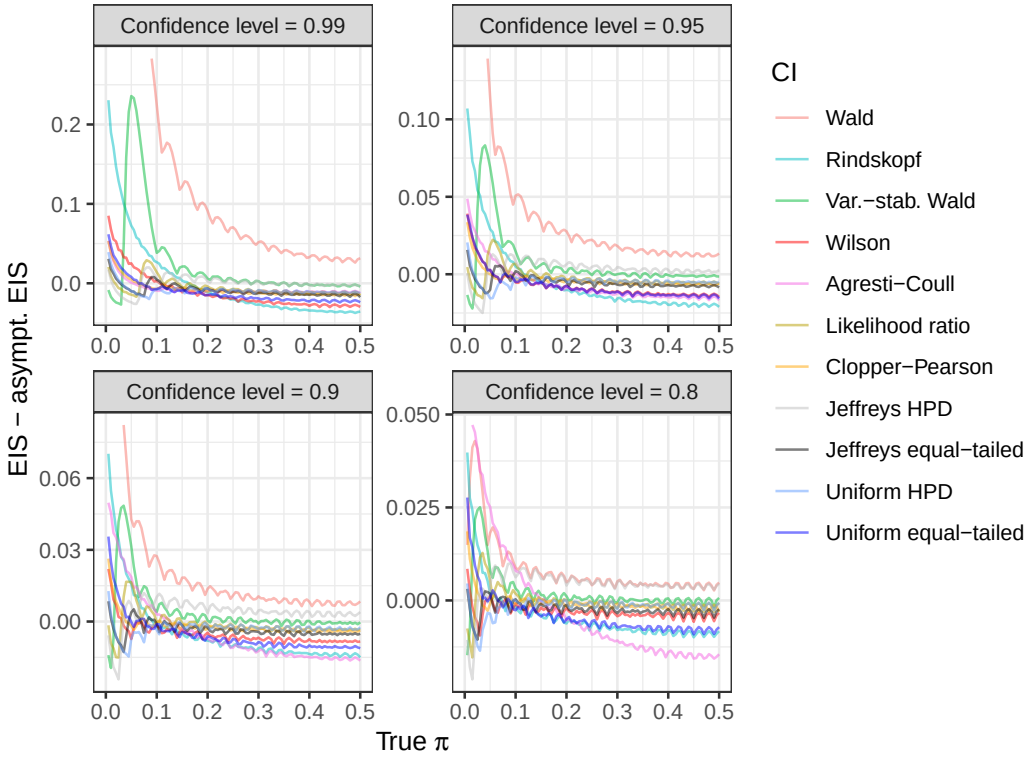

Figure 3.6: Difference between expected interval score and asymptotical EIS for each method as a function of π for $n = 50$ and $\gamma \in \{0.99, 0.95, 0.9, 0.8\}$.

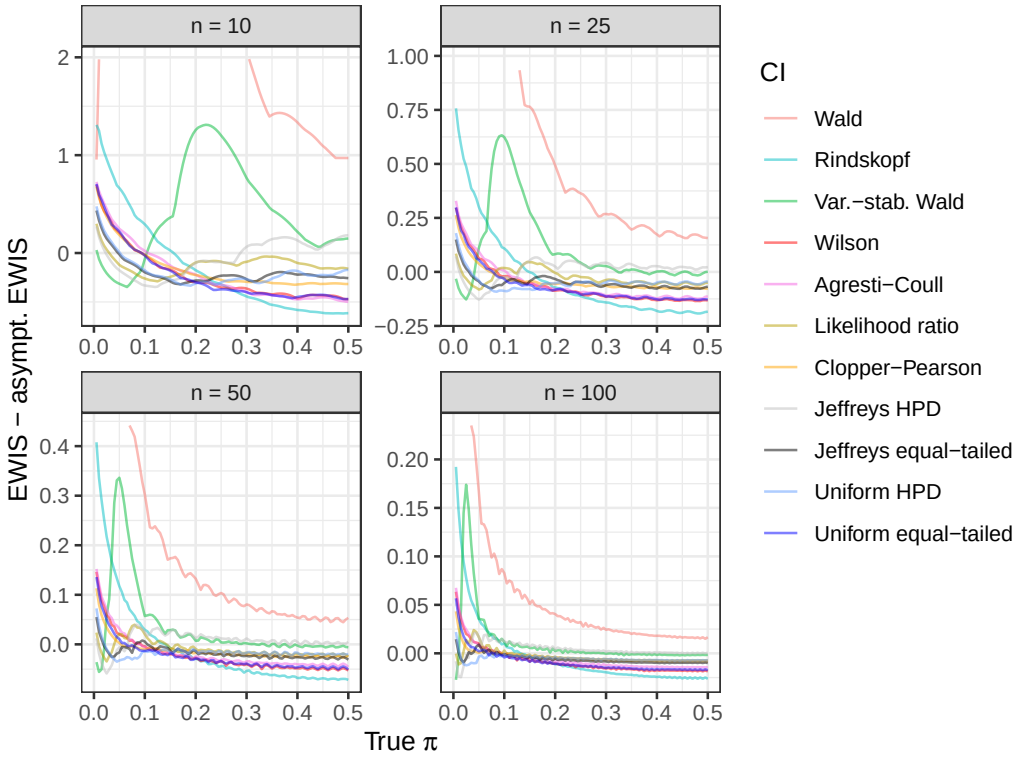

Figure 3.7: Difference between expected weighted interval score and asymptotical EWIS for each method as a function of π for $n \in \{10, 25, 50, 100\}$ combining confidence levels $\gamma \in \{0.9, 0.95, 0.99\}$.

the boundaries. The Wilson and the uniform equal-tailed CIs are equivalent (for all n). The likelihood ratio CI is worse than these two except at the boundaries.

Figure 3.6 compares the CIs for fixed $n = 50$ but varying confidence levels. Overall, different confidence levels mostly affect the ranking of the Agresti-Coull CI. Also, for confidence levels lower than 0.95, the Wilson CI is better than the uniform equal-tailed at the boundaries but worse near $\pi = 0.5$. For larger confidence levels, it is the other way around.

3.4.1 Weighted interval score

Figure 3.7 considers the three confidence levels 0.9, 0.95 and 0.99 at once by comparing the expected *weighted* interval scores for different sample sizes n. What changes compared to the expected interval score for a level of 0.95, is that the Agresti-Coull CI performs equally well as the Wilson and the uniform equal-tailed CIs. The Agresti-Coull CI was the only one that changed its ranking for different levels.

3.4.2 Integral as summary measure

In this subsection, the integral over all possible true proportions $\pi \in (0, 1)$ is used as a summary measure. The function that is integrated is the difference between the expected interval score and the asymptotical reference of the expected interval score. For each sample size n, now the methods can be compared with one number by these integrals.

Figure 3.8 yields a clear ranking of the CIs that virtually holds for every n between 10 and 100. The legend in Figure 3.8 is ranked for $n = 10$ where the best CI is at the top. The three best

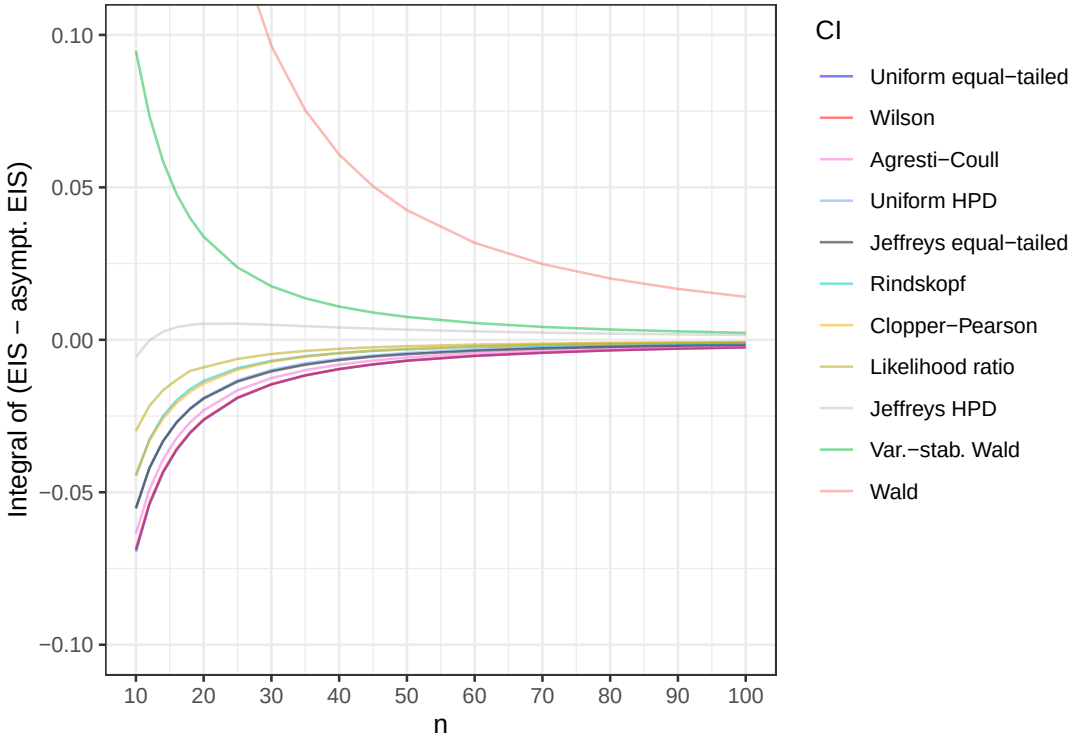

Figure 3.8: Integral of difference between expected interval score and asymptotical EIS for each method (ranked for $n = 10$) as a function of n for $\gamma = 0.95$, integrated over $\pi \in (0, 1)$.

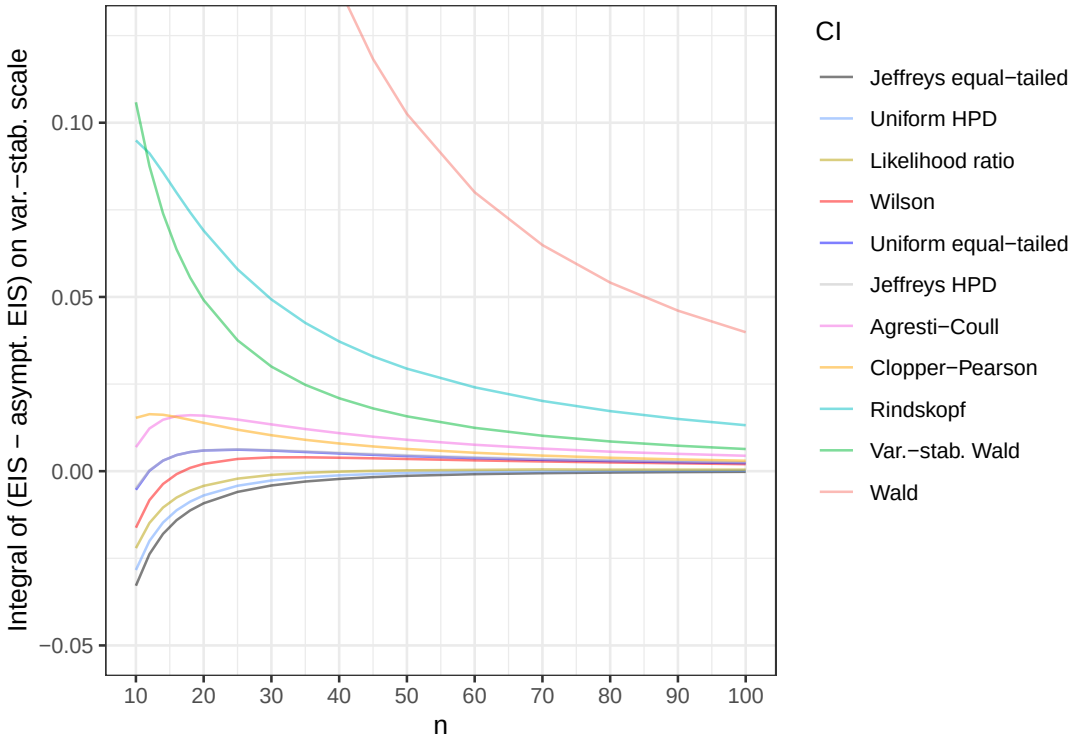

Figure 3.9: Integral of difference between expected interval score and asymptotical EIS for each method (ranked for $n = 10$) as a function of n for $\gamma = 0.95$, integrated over the variance-stabilized $\phi = \arcsin(\sqrt{\pi}) \in (0, \pi/2)$.

CIs are the uniform equal-tailed, the Wilson and the Agresti-Coull (in decreasing order). The three worst CIs are the Jeffreys HPD, the variance-stabilized Wald and the Wald (in decreasing order). All CIs converge to the same value for increasing n but not all at the same speed. The Wald CI coverges slower. In this case, they converge to 0 since the integral is computed for the difference to the asymptotical reference value.

Figure 3.9 is a comparison of the CIs by the integral of the expected interval score (difference to the asymptotical reference) after the variance-stabilizing transformation of the true proportions π. This corresponds to weighting the integral where more weight is attributed to rare cases near the two boundaries 0 and 1. Poor performance for boundary values is a common problem in CIs for the binomial proportion. So, this comparison of the CIs is like a sensitivity analysis with special attention to the performance in extreme cases.

The ranking of the CIs does indeed change compared to the usual integral. Now, for $n = 10$, the Jeffreys equal-tailed, the uniform HPD and the Likelihood ratio CIs are at the top of the list (in decreasing order), followed by the Wilson CI. The variance-stabilized Wald and Wald CIs are still at the bottom of the list. For small n, the ranking changes between the Rindskopf and the variance-stabilized Wald CIs and between the Agresti-Coull and the Clopper-Pearson CIs. The good performance of the Jeffreys equal-tailed CI can be explained by the connection between the Jeffreys prior and the variance-stabilizing transformation (see Subsection 2.5.6).

Essentially the same results hold for the integrals of the weighted expected interval scores. The only differences are: For the integral without transformation, the Rindskopf CI worsens towards the likelihood ratio CI and the Agresti-Coull CI worsens towards the Jeffreys equal-tailed CI. For the integral after the variance-stabilizing transformation, the Wilson CI is worse than the uniform equal-tailed CI for $n > 15$.

3.5 Generalized interval score

Figure 3.10 compares the expected interval score to the expected generalized interval score for 95% Jeffreys HPD intervals for different sample sizes n as a function of the true proportion π (symmetric around 0.5). If π is not near the boundaries, the Jeffreys HPD interval has the better ranking with the generalized interval score. Assuming the true π lies on the left of the CI, then the generalized interval score is smaller than the interval score if and only if

$$\alpha_1 > \frac{\alpha}{2}.$$

The same holds for α_2, assuming that the true π lies on the right of the CI. Near the boundaries of π, the expected generalized interval score is penalized more than the expected interval score because then extreme observations are more likely for which the lower (or upper) exceedance probability is actually (under the posterior) smaller than $\alpha/2$.

Overall, looking at the generalized interval score leads to better rankings for the HPD intervals compared to the equal-tailed intervals, see Figure 3.11. The same range of the y-axis is chosen such that it can be compared to Figure 3.8. This means that in terms of the expected generalized interval score (using the integral summary measure), the uniform HPD interval is at the top of the list, followed by the Jeffreys HPD interval. Note that for the other CIs, the generalized interval score is the same as the interval score since they are considered to be central.

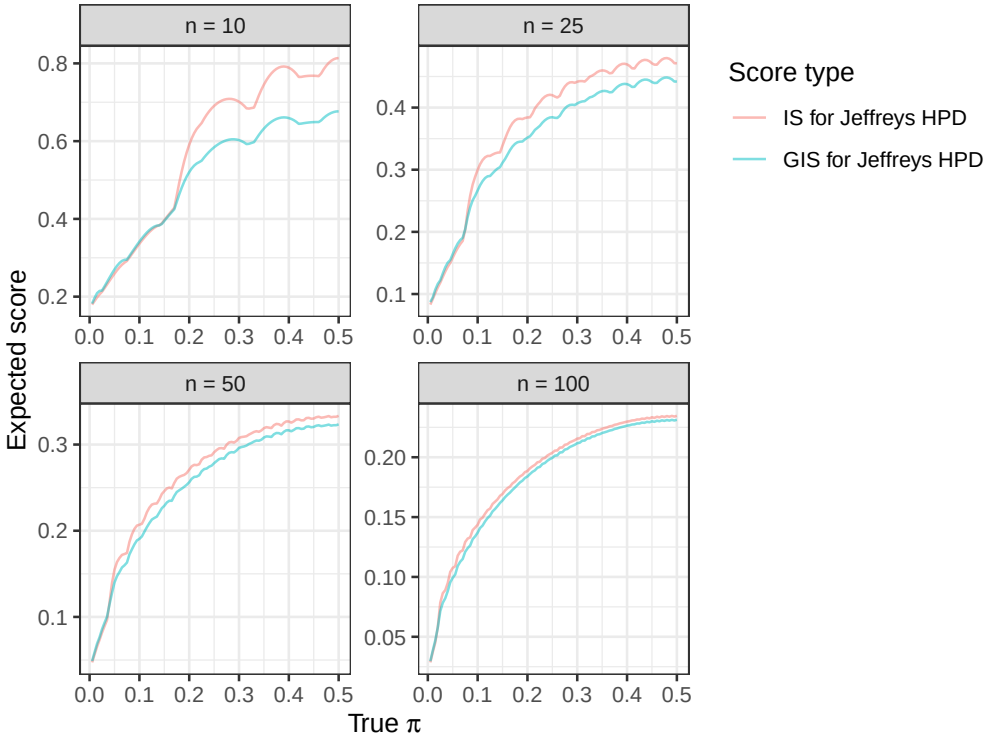

Figure 3.10: Expected interval score and expected generalized interval score for the Jeffreys HPD interval as a function of π for $\gamma = 0.95$ and $n \in \{10, 25, 50, 100\}$.

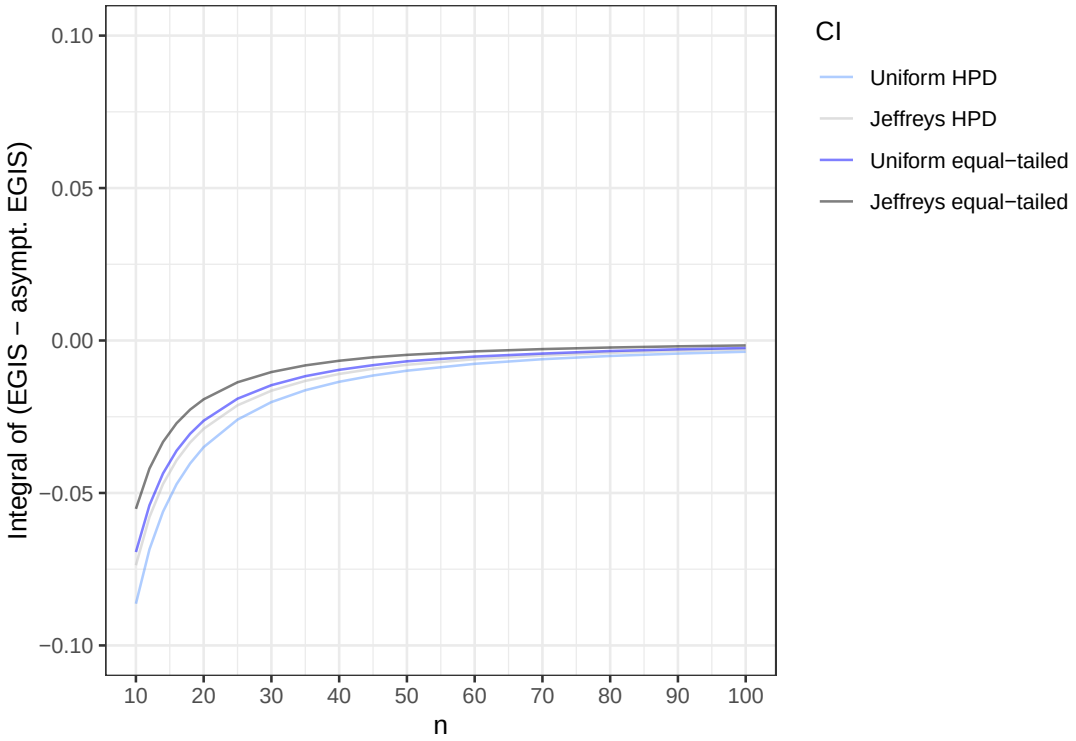

Figure 3.11: Integral of difference between expected generalized interval score and asymptotical EGIS for each Bayesian method (ranked for $n = 10$) as a function of n for $\gamma = 0.95$, integrated over $\pi \in (0, 1)$.

Chapter 4

Discussion

The interval score is a measure to evaluate confidence intervals that combines width and coverage. It is intuitively appealing since it allows for a decomposition into a measure of sharpness and penalties for over- and underestimation (Bracher *et al.*, 2021). The measure of sharpness is simply the interval width. The size of the penalties are the minimal distance of the true parameter to the confidence interval times $2/\alpha$, where α is the complement of the confidence level. So, the interval score does not only address non-coverage per se but the *amount* of non-coverage. Non-coverage is penalized more if the true parameter is far away from the confidence interval and if the confidence level is large. So far, the interval score has only been used for prediction intervals but not for confidence intervals.

In this master thesis, eleven confidence intervals for the binomial proportion have been compared using the expected interval score. The frequentist intervals were the Clopper-Pearson, Wilson, Wald, Rindskopf (logit Wald with adjustment), variance-stabilized Wald, Agresti-Coull and likelihood ratio confidence intervals. Also Bayesian equal-tailed and HPD intervals for uniform and Jeffreys priors were included in the analysis. These intervals are interpreted as frequentist intervals in the comparison.

Both in terms of coverage probability and in terms of the expected interval score, the Wald CI is the worst of the compared methods. Its good performance in terms of width for true proportions near the boundaries influences the interval score less than the bad coverage. Also in the literature, the consensus is that the Wald CI should generally not be used (Brown *et al.*, 2001; Held and Sabanés Bové, 2020; Gillibert *et al.*, 2021).

The confidence intervals with the overall best performance, again both in terms of coverage probability and the expected interval score, are the Wilson CI and the uniform equal-tailed CI. This is also the result of a clear ranking by the integral summary measure when the expected interval score is integrated over the true proportions from 0 to 1. However, the Jeffreys equal-tailed CI or the uniform HPD CI should be chosen when rare cases are of special interest. These are the best methods in the ranking by the integral summary measure when the variance-stabilizing transformation is applied to the set of true proportions. Integrating on the transformed scale results in a higher weight for boundary values. Recommendations are based on whether rare cases are of interest since usually different methods perform well for rare cases and non-rare cases.

By looking at coverage probabilities and expected widths as functions of π, the Wilson CI and the Jeffreys equal-tailed CI are also generally recommended in Held and Sabanés Bové (2020) and for small sample sizes $n \leq 40$ in Brown *et al.* (2001). For $n > 40$, the Agresti-Coull CI is recommended because it is simpler. Gillibert *et al.* (2021) prefers equal-tailed (central) CIs because of the balance of one-sided errors. The Clopper-Pearson mid-P CI or a modified Jeffreys equal-tailed CI are recommended (but the modification is criticized) based on *local average*

coverage computed by randomly drawing the true proportion π and the outcome x. Other summary measures are the minimum or mean coverage probabilities. If a nominal minimum coverage probability is desired, Newcombe (1998) recommends the Clopper-Pearson CI. The Wilson CI or the mid-P Clopper-Pearson CI are recommended if a nominal mean coverage probability is desired. Pires and Amado (2008) recommends the Agresti-Coull CI considering nominal mean coverage probability because it is simpler.

The integral summary measure, closely related to the integrated risk in Bayesian decision theory, is also a way of averaging the considered measure. To the best knowledge of the author, it has not been used elsewhere to evaluate CIs for the binomial proportion. The advantage of this technique is that more weight can be given to rare cases if this is a concern. Poor performance in scenarios with rare cases is a problem for many methods. This should be taken into account since according to Tuyl (2007, p. 17), "the most important property of a method is that it produces sensible intervals for any possible data outcome", and according to Jaynes (1976, p. 178), "the merits of any statistical method are determined by the results it gives when applied to specific problems" (quoted in Newcombe (2013, Subsection Boundary Anomalies)). Indeed, in practice, low or zero counts can happen if the sample size is small or for rare diseases. Large counts nearly equal to n can happen *e. g.* when estimating the sensitivity or the specificity of a diagnostic test (Newcombe, 2013). Also in these cases, a confidence interval method should produce a sensible interval.

The major advantage of the interval score as an evaluation method is that it is a proper scoring rule. According to Gneiting and Raftery (2007), it solves that "the question of measuring optimality (either frequentist or Bayesian) of a set estimator against a loss criterion combining size and coverage does not yet have a satisfactory answer", pointed out by Casella *et al.* (1993, p. 141). The interval score indeed combines width and coverage. However, it is only a proper scoring rule for interval estimators that are central. This is a limitation for the use of the interval score for the comparison of confidence intervals since many confidence intervals are not central. In this comparison of confidence intervals for the binomial proportion, only the equal-tailed CIs are central while the HPD CIs are not and all the others are only asymptotically central.

A generalization of the interval score for non-central CIs has been defined in this master thesis. It is not a proper scoring rule anymore since the generalized interval score depends on the two unequal exceedance probabilities which vary for different CIs. The advantage of the generalized interval score is that the amount of non-coverage is correctly assessed by taking into account the correct exceedance probability. For the HPD CI, the generalized interval score is slightly larger at the boundaries but otherwise smaller than the interval score. It is not surprising that the HPD CIs outperform the (otherwise best) equal-tailed CIs in terms of the generalized interval score since this score corrects the (otherwise unfair) weights of the non-coverage penalty. If the interval score is not used as a proper scoring rule but for its intuitive appeal combining width and coverage, the generalized interval score is a reasonable alternative. It is however computationally more involved and the exceedance probabilities need to be calculable.

Strictly speaking, proper scoring rules have only been defined for probabilistic predictions, where the outcome of interest is assumed to come from a probability distribution. The definition depends heavily on this assumption. The same assumption holds in the Bayesian estimation setting but the frequentist setting assumes a fixed unknown parameter. A simple option is to understand proper scoring rules for the frequentist estimation setting as a loss function that would be a proper scoring rule in a setting where the unknown parameter is assumed to have a distribution (as in the Bayesian setting). This is the definition used in this master thesis since no source with a clear definition could be found. In Gneiting and Raftery (2007), proper scoring rules for estimation are described as attractive loss functions. In Buja *et al.* (2005), they are defined as loss functions with the property that the expected loss (w.r.t. the probability

distribution of the data) is minimized under the true parameter. With this definition, also the score

$$\mathrm{W}(l, u) - \mathbb{1}(l \leq y \leq u),$$

for interval limits l and u and true parameter y, would be a proper scoring rule. The minimum of the score is -1 and is only attained for the degenerate (point) interval $l = u = y$. However, this score leads to a paradox in Casella *et al.* (1993) and Gneiting and Raftery (2007) argues that this paradox can be solved using proper scoring rules. So, apparently, propriety means more than just having a global minimum at the true parameter. It should also cause a sensible ordering of different estimators, which is not the case for the paradoxical example. Indeed, with the definition from this master thesis where an unknown distribution is assumed for π, it could be shown that the paradoxical score is not proper (by choosing $l = u < q_1$, where q_1 denotes the true lower quantile, as a counterexample). How the concept of propriety for the frequentist estimation setting is understood in this master thesis could maybe be related to the concept of *confidence distributions*, see Xie and Singh (2013), where the parameter of interest is equipped with some sort of probability distribution.

The results from this master thesis could be extended in at least two ways. For one, the Bayesian intervals could be compared using the weighted interval score that approximates the CRPS. The Bayesian setting is more similar to the prediction setting for which the interval score has been developed and proper scoring rules have a clear definition in this setting. It would be reasonable to further explore the Bayesian intervals because of their good performance. Moreover, the new approach from this master thesis could also be used to compare different confidence intervals for other parameters than the binomial proportion. For example, heterogeneity variances in meta-analysis have many different interval estimators. This approach could be valuable to better understand the resulting confidence intervals.

Appendix A

Software

All analyses were performed in the `R` programming language (R Core Team, 2021), R version 3.6.2 (2019-12-12), using the packages displayed in the session info.

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=German_Switzerland.1252  LC_CTYPE=German_Switzerland.1252
## [3] LC_MONETARY=German_Switzerland.1252 LC_NUMERIC=C
## [5] LC_TIME=German_Switzerland.1252
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] doParallel_1.0.15 iterators_1.0.12  foreach_1.4.7     ggpubr_0.3.0
## [5] scales_1.1.0      ggplot2_3.2.1     knitr_1.27
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.0.0  xfun_0.12         purrr_0.3.3       haven_2.2.0
##  [5] lattice_0.20-38   carData_3.0-3     colorspace_2.0-0  vctrs_0.2.2
##  [9] generics_0.0.2    rlang_0.4.4       pillar_1.4.3      foreign_0.8-72
## [13] glue_1.3.1        withr_2.1.2       readxl_1.3.1      lifecycle_0.1.0
## [17] stringr_1.4.0     munsell_0.5.0     ggsignif_0.6.0    gtable_0.3.0
## [21] cellranger_1.1.0  zip_2.0.4         codetools_0.2-16  evaluate_0.14
## [25] labeling_0.3      rio_0.5.16        forcats_0.5.0     curl_4.3
## [29] highr_0.8         broom_0.5.4       Rcpp_1.0.3        backports_1.1.5
## [33] abind_1.4-5       farver_2.0.3      hms_0.5.3         digest_0.6.23
## [37] stringi_1.4.5     openxlsx_4.1.4    rstatix_0.5.0     dplyr_0.8.4
## [41] grid_3.6.2        cowplot_1.0.0     tools_3.6.2       magrittr_1.5
## [45] lazyeval_0.2.2    tibble_2.1.3      crayon_1.3.4      car_3.0-7
```

```
## [49] tidyr_1.0.2       pkgconfig_2.0.3   data.table_1.12.8 assertthat_0.2.1
## [53] R6_2.4.1          nlme_3.1-147      compiler_3.6.2
```

Appendix B

R code

Code provided by Held and Sabanés Bové (2020) was used to implement confidence intervals, smoothed coverage probabilities and a plot function for the coverage probability for the binomial proportion. The implementation of confidence intervals also used code from package `biostatUZH` for the Wald, Wilson, Agresti-Coull, Jeffreys equal-tailed and Clopper-Pearson confidence intervals.

```r
# ==============================================================================
# Checks if x is a whole number
# ==============================================================================
is.wholenumber <- function(x) {abs(x - round(x)) < .Machine$double.eps^0.5}

# ==============================================================================
# Computes a Wald confidence interval
# - x is the number of successes
# - n is the sample size
# - conf.level is the confidence level
# overshoot is truncated to [0,1]
# degenerate zero width interval for x=0 and x=n
# ==============================================================================
waldCI <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
  q <- qnorm(p = (1 + conf.level)/2)
  p <- x/n
  lower <- max(p - q*sqrt(p*(1 - p)/n), 0)
  upper <- min(p + q*sqrt(p*(1 - p)/n), 1)
  return(cbind(lower, upper))
}

# ==============================================================================
# Computes a (modified) logit Wald confidence interval
# - x is the number of successes
# - n is the sample size
# - conf.level is the confidence level
# modification: add 0.5 successes and 0.5 failures
# (otherwise for x=0 and x=n no interval computable)
```

```r
# ============================================================================
logitwaldCI <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
  q <- qnorm(p = (1 + conf.level)/2)
  x <- x + 0.5
  n <- n + 1
  lower <- plogis(log(x/(n-x)) - q*sqrt(1/x + 1/(n - x)))
  upper <- plogis(log(x/(n-x)) + q*sqrt(1/x + 1/(n - x)))
  return(cbind(lower, upper))
}

# ============================================================================
# Computes a variance-stabilized Wald confidence interval
# - x is the number of successes
# - n is the sample size
# - conf.level is the confidence level
# overshoot is truncated to [0,pi/2] on var.-stab. scale
# ============================================================================
varstabwaldCI <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
  q <- qnorm(p = (1 + conf.level)/2)
  p <- x/n
  lower <- sin(max(asin(sqrt(p)) - q*sqrt(1/(4*n)), 0))^2
  upper <- sin(min(asin(sqrt(p)) + q*sqrt(1/(4*n)), pi/2))^2
  return(cbind(lower, upper))
}

# ============================================================================
# Computes a Wilson confidence interval
# - x is the number of successes
# - n is the sample size
# - conf.level is the confidence level
# ============================================================================
wilsonCI <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
  q <- qnorm(p = (1 + conf.level)/2)
  p <- x/n
  pseudo.est <- (x + q^2/2)/(n + q^2)
  pseudo.se <- sqrt(n)/(n + q^2) * sqrt(p*(1 - p) + q^2/(4*n))
  lower <- pseudo.est - q*pseudo.se
  upper <- pseudo.est + q*pseudo.se
  return(cbind(lower, upper))
}

# ============================================================================
# Computes an Agresti-Coull confidence interval
```

```r
# - x is the number of successes
# - n is the sample size
# - conf.level is the confidence level
# overshoot is truncated to [0,1]
# ============================================================================
agrestiCI <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
  q <- qnorm(p = (1 + conf.level)/2)
  x <- x + 2
  n <- n + 4
  p <- x/n
  lower <- max(p - q*sqrt(p*(1 - p)/n), 0)
  upper <- min(p + q*sqrt(p*(1 - p)/n), 1)
  return(cbind(lower, upper))
}

# ============================================================================
# Computes a Clopper-Pearson confidence interval
# - x is the number of successes
# - n is the sample size
# - conf.level is the confidence level
# ============================================================================
clopperpearsonCI <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
  alpha <- 1 - conf.level
  if(x == 0) {
    lower <- 0
    upper <- 1 - (alpha/2)^(1/n)
  }
  else if(x == n) {
    lower <- (alpha/2)^(1/n)
    upper <- 1
  }
  else {
    lower <- qbeta(p = alpha/2, shape1 = x, shape2 = n - x + 1)
    upper <- qbeta(p = 1 - alpha/2, shape1 = x + 1, shape2 = n - x)
  }
  return(cbind(lower, upper))
}

# ============================================================================
# Computes a likelihood confidence interval
# - x is the number of successes
# - n is the sample size
# - conf.level is the confidence level
# numerical solutions for 0<x<n
# ============================================================================
```

```r
likelihoodCI <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
  p <- x/n
  eps <- 1e-12
  loglik <- function(p, x, n) {x*log(p) + (n - x)*log(1 - p)}
  f <- function(theta) {
    loglik(p = theta, x = x, n = n) - loglik(p = p, x = x, n = n) +
      1/2*qchisq(p = conf.level, df = 1)
  }
  if(x == 0) {
    lower <- 0
    upper <- 1 - exp(-1/2*qchisq(p = conf.level, df = 1)/n)
  }
  else if(x == n) {
    lower <- exp(-1/2*qchisq(p = conf.level, df = 1)/n)
    upper <- 1
  }
  else {
    lower <- uniroot(f, interval = c(eps, p))$root
    upper <- uniroot(f, interval = c(p, 1 - eps))$root
  }
  return(cbind(lower, upper))
}

# ============================================================================
# Computes a Jeffreys equal-tailed credible interval
# - x is the number of successes
# - n is the sample size
# - conf.level is the credibility level
# ============================================================================
jeffreysET <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
  alpha <- 1 - conf.level
  lower <- qbeta(p = alpha/2, shape1 = x + 0.5, shape2 = n - x + 0.5)
  upper <- qbeta(p = 1 - alpha/2, shape1 = x + 0.5, shape2 = n - x + 0.5)
  return(cbind(lower, upper))
}

# ============================================================================
# Computes a uniform equal-tailed credible interval
# - x is the number of successes
# - n is the sample size
# - conf.level is the credibility level
# ============================================================================
uniformET <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
```

```
  alpha <- 1 - conf.level
  lower <- qbeta(p = alpha/2, shape1 = x + 1, shape2 = n - x + 1)
  upper <- qbeta(p = 1 - alpha/2, shape1 = x + 1, shape2 = n - x + 1)
  return(cbind(lower, upper))
}

# ================================================================================
# Helper function for HPD intervals with beta posteriors,
# returns the probability of all points for which the beta density is smaller
# than h (two tails) as well as the two boundaries of these regions
# - p1 and p2 are the parameters of the beta density
# - h is the function value of the beta density (height)
# ================================================================================
outerdens <- function(h, p1, p2){
  modus <- (p1 - 1)/(p1 + p2 - 2)
  schnitt.l <- uniroot(function(x){dbeta(x, p1, p2) - h},
                       interval = c(0, modus))$root
  schnitt.u <- uniroot(function(x){dbeta(x, p1, p2) - h},
                       interval = c(modus, 1))$root
  tails <- pbeta(schnitt.l, p1, p2) + pbeta(schnitt.u, p1, p2,
                                            lower.tail = FALSE)
  return(c(tails, schnitt.l, schnitt.u))
}

# ================================================================================
# Computes a Jeffreys HPD interval
# - x is the number of successes
# - n is the sample size
# - conf.level is the credibility level
# extreme cases x=0, x=n handled differently (no mode in these cases)
# ================================================================================
jeffreysHPD <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
  alpha <- 1 - conf.level
  p1 <- x + 0.5
  p2 <- n - x + 0.5
  modus <- (p1 - 1)/(p1 + p2 - 2)
  eps <- 1e-15
  if(x == 0) {
    lower <- 0
    upper <- qbeta(p = conf.level, shape1 = 0 + 0.5, shape2 = n - 0 + 0.5)
  }
  else if(x == n) {
    lower <- qbeta(p = 1 - conf.level, shape1 = n + 0.5, shape2 = n - n + 0.5)
    upper <- 1
  }
  else {
    height <- uniroot(function(h) {
```

```
      outerdens(h = h, p1 = p1, p2 = p2)[1] - alpha},
      interval = c(eps, dbeta(modus, p1, p2) - 10*eps))[["root"]]
    lower <- outerdens(h = height, p1 = p1, p2 = p2)[2]
    upper <- outerdens(h = height, p1 = p1, p2 = p2)[3]
  }
  return(cbind(lower, upper))
}

# ============================================================================
# Computes a uniform HPD interval
# - x is the number of successes
# - n is the sample size
# - conf.level is the credibility level
# extreme cases x=0, x=n handled differently (no mode in these cases)
# ============================================================================
uniformHPD <- function(x, n, conf.level) {
  stopifnot(is.wholenumber(x), is.wholenumber(n), x <= n, n >= 1, x >= 0,
            conf.level > 0, conf.level < 1)
  alpha <- 1 - conf.level
  p1 <- x + 1
  p2 <- n - x + 1
  modus <- (p1 - 1)/(p1 + p2 - 2)
  eps <- 1e-15
  if(x == 0) {
    lower <- 0
    upper <- qbeta(p = conf.level, shape1 = 0 + 1, shape2 = n - 0 + 1)
  }
  else if(x == n) {
    lower <- qbeta(p = 1 - conf.level, shape1 = n + 1, shape2 = n - n + 1)
    upper <- 1
  }
  else {
    height <- uniroot(function(h) {
      outerdens(h = h, p1 = p1, p2 = p2)[1] - alpha},
      interval = c(eps, dbeta(modus, p1, p2) - 10*eps))[["root"]]
    lower <- outerdens(h = height, p1 = p1, p2 = p2)[2]
    upper <- outerdens(h = height, p1 = p1, p2 = p2)[3]
  }
  return(cbind(lower, upper))
}

# ============================================================================
# Returns a list with lower and upper limits for all possible x (nb of success)
# for the CIs: Wald, Logit Wald, Var stab Wald, Wilson, Agresti, Likelihood,
#              Clopper-Pearson, , Jeffreys HPD, Jeffreys equal-tailed,
#              Uniform HPD, Uniform equal-tailed
# - n is the sample size
# - conf.level is the confidence level
# ============================================================================
```

```r
CIprop <- function(n, conf.level) {
  x <- 0:n
  nwald <- t(sapply(x, waldCI, n = n, conf.level = conf.level))
  lwald <- t(sapply(x, logitwaldCI, n = n, conf.level = conf.level))
  vwald <- t(sapply(x, varstabwaldCI, n = n, conf.level = conf.level))
  wil <- t(sapply(x, wilsonCI, n = n, conf.level = conf.level))
  agr <- t(sapply(x, agrestiCI, n = n, conf.level = conf.level))
  lik <- t(sapply(x, likelihoodCI, n = n, conf.level = conf.level))
  CP <- t(sapply(x, clopperpearsonCI, n = n, conf.level = conf.level))
  jeffET <- t(sapply(x, jeffreysET, n = n, conf.level = conf.level))
  jeffHPD <- t(sapply(x, jeffreysHPD, n = n, conf.level = conf.level))
  unifET <- t(sapply(x, uniformET, n = n, conf.level = conf.level))
  unifHPD <- t(sapply(x, uniformHPD, n = n, conf.level = conf.level))
  return(list(nwald, lwald, vwald, wil, agr, lik, CP,
              jeffHPD, jeffET, unifHPD, unifET))
}

# ============================================================================
# Computes the width of a confidence interval
# - ci is a confidence interval (vector with lower and upper limit)
# ============================================================================
width <- function(ci) {
  lower <- ci[1]
  upper <- ci[2]
  return(upper - lower)
}

# ============================================================================
# Computes the mean width
# - true.pi is the true proportion
# - n is the sample size
# - ci contains all possible confidence intervals (for all possible x = 0:n),
#   a matrix where each row contains lower and upper limit for one x
# conf.level is not needed here but in CIpropmeasures() because of the score
# ============================================================================
meanwidth <- function(true.pi, n, ci, conf.level) {
  sum(dbinom(x = 0:n, size = n, prob = true.pi)*apply(ci, 1, width))
}

# ============================================================================
# Computes the coverage of a confidence interval
# - ci is a confidence interval (vector with lower and upper limit)
# - true.pi is the true proportion
# ============================================================================
coverage <- function(ci, true.pi) {
  lower <- ci[1]
  upper <- ci[2]
  return((true.pi >= lower) & (true.pi <= upper))
}
```

```r
# ============================================================================
# Computes the coverage probability
# - true.pi is the true proportion
# - n is the sample size
# - ci contains all possible confidence intervals (for all possible x = 0:n),
#   a matrix where each row contains lower and upper limit for one x
# conf.level is not needed here but in CIpropmeasures() because of the score
# ============================================================================
meancoverage <- function(true.pi, n, ci, conf.level) {
  sum(dbinom(x = 0:n, size = n, prob = true.pi)*
        apply(ci, 1, coverage, true.pi = true.pi))
}

# ============================================================================
# Computes the interval score of a confidence interval
# - ci is a confidence interval (vector with lower and upper limit)
# - true.pi is the true proportion
# - conf.level is the corresponding (!) confidence level of ci
# ============================================================================
score <- function(ci, true.pi, conf.level) {
  alpha <- 1 - conf.level
  return(width(ci) +
           2/alpha*min(abs(true.pi - ci))*(1 - coverage(ci, true.pi)))
}

# ============================================================================
# Computes the mean interval score
# - true.pi is the true proportion
# - n is the sample size
# - ci contains all possible confidence intervals (for all possible x = 0:n),
#   a matrix where each row contains lower and upper limit for one x
# - conf.level is the corresponding (!) confidence level of ci
# - varstab.rescale decides if the var.-stab. transformation should be used
# ============================================================================
meanscore <- function(true.pi, n, ci, conf.level, varstab.rescale = FALSE) {
  if(varstab.rescale == TRUE) {true.pi <- sin(true.pi)^2}
  sum(dbinom(x = 0:n, size = n, prob = true.pi)*
        apply(ci, 1, score, conf.level = conf.level, true.pi = true.pi))
}
# needs to be vectorized for the numerical integration
meanscore <- Vectorize(meanscore, vectorize.args = c("true.pi"))

# ============================================================================
# Computes the weighted interval score of confidence intervals for different
# confidence levels
# - ci is a vector with confidence intervals (lower and upper limit),
#   concatenated for the different levels
# - true.pi is the true proportion
```

```r
# - conf.level is a vector with the corresponding (!) confidence levels of ci
# ============================================================================
wscore <- function(ci, true.pi, conf.level) {
  ci <- matrix(ci, ncol = 2, byrow = TRUE)
  res <- apply(ci, 1, score, true.pi = true.pi, conf.level = conf.level)
  return(sum(diag(res)))
}

# ============================================================================
# Computes the mean weighted interval score
# - true.pi is the true proportion
# - n is the sample size
# - ci contains all possible confidence intervals (for all possible x = 0:n),
#   a matrix where each row contains lower and upper limits for one x,
#   concatenated for all considered levels (per row)
# - conf.level is a vector with the corresponding (!) confidence levels of ci
# - varstab.rescale decides if the var.-stab. transformation should be used
# ============================================================================
meanwscore <- function(true.pi, n, ci, conf.level, varstab.rescale = FALSE) {
  if(varstab.rescale == TRUE) {true.pi <- sin(true.pi)^2}
  sum(dbinom(x = 0:n, size = n, prob = true.pi)*
        apply(ci, 1, wscore, conf.level = conf.level, true.pi = true.pi))
}
# needs to be vectorized for the numerical integration
meanwscore <- Vectorize(meanwscore, vectorize.args = c("true.pi"))

# ============================================================================
# Returns mean width, coverage probability and mean interval score for a finite
# grid of true proportions and the following confidence intervals:
# Wald, Logit Wald, Var stab Wald, Wilson, Agresti, Likelihood, Clopper-Pearson,
# Jeffreys HPD, Jeffreys equal-tailed, Uniform HPD, Uniform equal-tailed
# (as a concatenated vector)
# - n is the sample size
# - conf.level is the confidence level
# - nbpoints is the number of points in the equidist. grid of true proportions
# measures symmetric around 0.5 --> grid of (0,0.5]
# ============================================================================
CIpropmeasures <- function(n, conf.level, nbpoints) {
  CIs <- CIprop(n, conf.level)
  pvector <- seq(0, 0.5, length = nbpoints + 1)[-1]
  methodwisemeanmeasure <- function(ci, measure.fct) {
    res <- sapply(pvector, measure.fct, n = n, ci = ci, conf.level = conf.level)
    return(res)
  }
  w <- lapply(CIs, methodwisemeanmeasure, measure.fct = meanwidth)
  w <- unlist(w)
  c <- lapply(CIs, methodwisemeanmeasure, measure.fct = meancoverage)
  c <- unlist(c)
  s <- lapply(CIs, methodwisemeanmeasure, measure.fct = meanscore)
```

```
  s <- unlist(s)
  return(data.frame("width" = w, "coverage" = c, "score" = s))
}

# ============================================================================
# Returns the integral of the mean interval score over the unit interval of
# underlying true proportions for the following confidence intervals:
# Wald, Logit Wald, Var stab Wald, Wilson, Agresti, Likelihood, Clopper-Pearson,
# Jeffreys HPD, Jeffreys equal-tailed, Uniform HPD, Uniform equal-tailed
# (in a vector with the same order)
# - n is the sample size
# - conf.level is the confidence level
# - nbpoints is the number of subdivisions used for the integration
# - varstab.rescale decides if the var.-stab. transformation should be used
# ============================================================================
CIpropscoreintegral <- function(n, conf.level, nbpoints,
                                varstab.rescale = FALSE) {
  CIs <- CIprop(n, conf.level)
  if(varstab.rescale == FALSE) {u <- 1}
  if(varstab.rescale == TRUE) {u <- pi/2}
  res <- lapply(CIs, function(ci) integrate(meanscore, n = n, ci = ci,
                                            conf.level = conf.level,
                                            varstab.rescale = varstab.rescale,
                                            lower = 0, upper = u,
                                            subdivisions = nbpoints)$value)
  return(unlist(res))
}

# ============================================================================
# Returns the mean weighted interval scores for a finite grid of true
# proportions and confidence intervals: Wald, Logit Wald, Var stab Wald, Wilson,
#                                       Agresti, Likelihood, Clopper-Pearson,
#                                       Jeffreys HPD, Jeffreys equal-tailed,
#                                       Uniform HPD, Uniform equal-tailed
# (as a concatenated vector)
# - n is the sample size
# - conf.level is a vector of the used confidence levels
# - nbpoints is the number of points in the equidist. grid of true proportions
# uses the linearity of the expectation
# ============================================================================
CIpropwscore <- function(n, conf.level, nbpoints) {
  score.matrix <- sapply(conf.level,
                         function(x) CIpropmeasures(conf.level = x,
                                                    nbpoints = nbpoints,
                                                    n = n)[, "score"])
  return(apply(score.matrix, 1, sum))
}

# ============================================================================
```

```r
# Returns the integral of the mean weighted interval score over the unit
# interval of underlying true proportions for the confidence intervals:
# Wald, Logit Wald, Var stab Wald, Wilson, Agresti, Likelihood, Clopper-Pearson,
# Jeffreys HPD, Jeffreys equal-tailed, Uniform HPD, Uniform equal-tailed
# (in a vector with the same order)
# - n is the sample size
# - conf.level is a vector of the used confidence levels
# - nbpoints is the number of subdivisions used for the integration
# - tol is the relative tolerance (of errors) in the integration
# - varstab.rescale decides if the var.-stab. transformation should be used
# ============================================================================
CIpropwscoreintegral <- function(n, conf.level, nbpoints, tol,
                                 varstab.rescale = FALSE) {
  res <- lapply(conf.level, CIprop, n = n)
  CIs <- res[[1]]
  for(i in 2:length(res)) {
    CIs <- mapply(cbind, CIs, res[[i]], SIMPLIFY = FALSE)
  }
  if(varstab.rescale == FALSE) {u <- 1}
  if(varstab.rescale == TRUE) {u <- pi/2}
  res <- lapply(CIs, function(ci) integrate(meanwscore, n = n, ci = ci,
                                            conf.level = conf.level,
                                            varstab.rescale = varstab.rescale,
                                            lower = 0, upper = u,
                                            subdivisions = nbpoints,
                                            rel.tol = tol)$value)
  return(unlist(res))
}

# ============================================================================
# Asymptotical reference for expected (weighted) interval score/ expected width
# and integrals thereof
# - true.pi is the true proportion
# - n is the sample size
# - conf.level is the confidence level
# - varstab.rescale decides if the var.-stab. transformation should be used
# - nbpoints is the number of subdivisions used for the integration
# ============================================================================
refmeanscore <- function(true.pi, n, conf.level, varstab.rescale = FALSE) {
  if(varstab.rescale == TRUE) {true.pi <- sin(true.pi)^2}
  q <- qnorm(p = (1 + conf.level)/2)
  s <- sqrt(true.pi*(1-true.pi)/n)
  2*s*dnorm(q)/(1-pnorm(q))
}

refmeanwscore <- function(true.pi, n, conf.level, varstab.rescale = FALSE) {
  if(varstab.rescale == TRUE) {true.pi <- sin(true.pi)^2}
  res <- sapply(conf.level, refmeanscore, true.pi = true.pi, n = n,
                varstab.rescale = varstab.rescale)
```

```r
  return(sum(res))
}
# needs to be vectorized for the numerical integration
refmeanwscore <- Vectorize(refmeanwscore, vectorize.args = c("true.pi"))

refmeanwidth <- function(true.pi, n, conf.level, varstab.rescale = FALSE) {
  if(varstab.rescale == TRUE) {true.pi <- sin(true.pi)^2}
  q <- qnorm(p = (1 + conf.level)/2)
  s <- sqrt(true.pi*(1-true.pi)/n)
  2*q*s
}

refmeanscoreintegral <- function(n, conf.level, nbpoints,
                                 varstab.rescale = FALSE) {
  if(varstab.rescale == FALSE) {u <- 1}
  if(varstab.rescale == TRUE) {u <- pi/2}
  res <- integrate(refmeanscore, n = n, conf.level = conf.level,
                   varstab.rescale = varstab.rescale,
                   lower = 0, upper = u,
                   subdivisions = nbpoints)$value
  return(res)
}

refmeanwscoreintegral <- function(n, conf.level, nbpoints,
                                  varstab.rescale = FALSE) {
  if(varstab.rescale == FALSE) {u <- 1}
  if(varstab.rescale == TRUE) {u <- pi/2}
  res <- integrate(refmeanwscore, n = n, conf.level = conf.level,
                   varstab.rescale = varstab.rescale,
                   lower = 0, upper = u,
                   subdivisions = nbpoints)$value
  return(res)
}

# ============================================================================
# Returns local average of coverage probabilities (as in BayarriBerger2004)
# for a finite grid of true proportions and the following confidence intervals:
# Wald, Logit Wald, Var stab Wald, Wilson, Agresti, Likelihood, Clopper-Pearson,
# Jeffreys HPD, Jeffreys equal-tailed, Uniform HPD, Uniform equal-tailed
# (as a concatenated vector)
# - n is the sample size
# - conf.level is the confidence level
# - nbpoints is the number of points in the equidist. grid of true proportions
# measures symmetric around 0.5 --> grid of (0,0.5]
# ============================================================================
CIproplocalcoverage <- function(n, conf.level, nbpoints) {
  CIs <- CIprop(n, conf.level)
  x <- 0:n
  pvector <- seq(0, 0.5, length = nbpoints + 1)[-1]
```

```
  a <- function(p) {
    if(p <= 0.025) {
      NA # (p*(1-p)*p^(-2) - 1)*p #1 - 2*0.025
    }
    else if(p >= (1 - 0.025)) {
      NA # (p*(1-p)*(1-p)^(-2) - 1)*p #1/0.025 - 3 + 2*0.025
    }
    else {
      (p*(1-p)*0.025^(-2) - 1)*p
    }
  }
  local.meancoverage <- function(p, ci) {
    ap <- a(p)
    a1mp <- a(1-p)
    alpha <- ap + x
    beta <- a1mp + n - x
    values.gamma <- (lchoose(n, x)
                     + lgamma(ap + a1mp) - lgamma(ap) - lgamma(a1mp)
                     + lgamma(ap + x) + lgamma(a1mp + n - x)
                     - lgamma(ap + a1mp + n))
    values.integral <- log(pbeta(ci[, 2], alpha, beta)
                           - pbeta(ci[, 1], alpha, beta))
    return(sum(exp(values.gamma + values.integral)))
  }
  methodwisemeanmeasure <- function(ci) {
    res <- sapply(pvector, local.meancoverage, ci = ci)
    return(res)
  }
  c <- lapply(CIs, methodwisemeanmeasure)
  return(unlist(c))
}

# ============================================================================
# GIS for HPD intervals (computation of tail probabilities, generalized interval
# score and expected generalized interval score)
# - n is the sample size
# - true.pi is the true proportion
# - prior is the used prior (uniform or Jeffreys)
# - tailprobs: ci is a vector with lower limit, upper limit, nb of success x
# - gscore: ci is a vector with lower limit, upper limit, tail prob. 1 and 2
# - meangscore: ci is a matrix with rows for all x=0:n, vector like in gscore
#   for each row
# - varstab.rescale decides if the var.-stab. transformation should be used
# ============================================================================
tailprobs <- function(ci, n, prior) {
  lower <- ci[1]
  upper <- ci[2]
  x <- ci[3]
  if (prior == "uniform") s <- 1
```

```r
  if (prior == "jeffreys") s <- 0.5
  if (x == 0) {
    alpha1 <- 0
    alpha2 <- pbeta(upper, x + s, n - x + s, lower.tail = FALSE)
  }
  else if (x == n) {
    alpha1 <- pbeta(lower, x + s, n - x + s)
    alpha2 <- 0
  }
  else {
    alpha1 <- pbeta(lower, x + s, n - x + s)
    alpha2 <- pbeta(upper, x + s, n - x + s, lower.tail = FALSE)
  }
  return(c(alpha1, alpha2))
}

gscore <- function(ci, true.pi) {
  lower <- ci[1]
  upper <- ci[2]
  alpha1 <- ci[3]
  alpha2 <- ci[4]
  if (alpha1 == 0) {
    k <- 1/alpha2
  }
  else if (alpha2 == 0) {
    k <- 1/alpha1
  }
  else {
    k <- ifelse(abs(true.pi - lower) < abs(true.pi - upper), 1/alpha1, 1/alpha2)
  }
  k <- k*min(abs(c(true.pi - lower, true.pi - upper)))
  return(width(ci) + k*(1 - coverage(ci, true.pi)))
}

meangscore <- function(true.pi, n, ci, varstab.rescale = FALSE) {
  if(varstab.rescale == TRUE) {true.pi <- sin(true.pi)^2}
  sum(dbinom(x = 0:n, size = n, prob = true.pi)*
        apply(ci, 1, gscore, true.pi = true.pi))
}
# needs to be vectorized for the numerical integration
meangscore <- Vectorize(meangscore, vectorize.args = c("true.pi"))

# ============================================================================
# Plot function for binomial proportion
# - df is a dataframe with columns x (x values), y (y values), CI (method of
#   confidence interval), type (sample size or confidence level setting)
# - xlab and ylab are the labels of the x- and y-axis
# - layers decides if we have several plots (types)
# - ylimfree decides if a common y range should be used
```

```r
# ============================================================================
library(ggplot2)
library(scales)
ggprop <- function(df, xlab, ylab, layers = FALSE, ylimfree = FALSE) {
  colours <- hue_pal()(6)
  colours <- c(colours[c(1,4,3)], "red", colours[c(6,2)], "orange", "grey",
               "black", colours[5], "blue")
  if(layers == FALSE) {
    relevelnb <- order(df$y[1:11])
    ranking10 <- levels(df$CI)[relevelnb]
    df$CI <- factor(df$CI, levels = ranking10)
    colours <- colours[relevelnb]
  }
  names(colours) <- levels(df$CI)
  p <- ggplot(df, aes(x = x, y = y, group = CI, colour = CI)) +
    geom_line(alpha = 0.5) +
    scale_colour_manual(name = "CI", values = colours) +
    labs(x = xlab, y = ylab) +
    theme_bw() +
    theme(legend.justification = "top", aspect.ratio = 1)
  if(layers == TRUE) p <- p + facet_wrap(~ type)
  if(ylimfree == TRUE) p <- p + facet_wrap(~ type, scales = "free")
  return(p)
}
```

Bibliography

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52**, 119–126. 2, 7

Altman, D. G., Machin, D., Bryant, T. N., and Gardner, M. J. (2000). *Statistics with confidence*. BMJ Books, second edition. 1

Bayarri, M. J. and Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, **19**, 58–80. 10, 15, 19

Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2021). Evaluating epidemic forecasts in an interval format. *PLoS Computational Biology*, **17**, 1–15. 14, 28

Brehmer, J. and Gneiting, T. (2021). Scoring interval forecasts: equal-tailed, shortest, and modal interval. *Bernoulli*, **27**, 1993–2010. 13

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, **16**, 101–117. 1, 2, 9, 28

Buja, A., Stuetzle, W., and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: structure and applications. Manuscript available from: www-stat.wharton.upenn.edu/~buja/. 29

Casella, G., Hwang, J. T. G., and Robert, C. (1993). A paradox in decision-theoretic interval estimation. *Statistica Sinica*, **3**, 141–155. 29, 30

Cervera, J. L. and Muñoz, J. (1996). Proper scoring rules for fractiles. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 5*, 513–519. Oxford University Press, Oxford, U.K. 12

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413. 5

Gillibert, A., Bénichou, J., and Fallisard, B. (2021). Two-sided confidence interval of a binomial proportion: how to choose? Preprint available from: https://arxiv.org/abs/2103.10463. 1, 5, 10, 28

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 243–268. 1

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378. iii, 1, 9, 10, 13, 19, 29, 30

Held, L. and Sabanés Bové, D. (2020). *Likelihood and Bayesian inference - with applications in biology and medicine*. Springer, second edition. 1, 2, 3, 4, 5, 6, 9, 10, 19, 28, 33

Jaynes, E. T. (1976). Confidence intervals vs Bayesian intervals. In Harper, W. L. and Hooker, C. A., editors, *Foundations of probability theory, statistical inference, and statistical theories of science*. Reidel, Dordrecht, Netherlands. 29

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous univariate distributions, volume 1*. Wiley. 17

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, **17**, 857–872. 1, 2, 6, 9, 13, 29

Newcombe, R. G. (2011). Measures of location for confidence intervals for proportions. *Communications in Statistics – Theory and Methods*, **40**, 1743–1767. 6

Newcombe, R. G. (2013). *Confidence intervals for proportions and related measures of effect size*. Chapman & Hall/CRC Biostatistics Series, first edition. 5, 6, 8, 29

Pires, A. M. and Amado, C. (2008). Interval estimators for a binomial proportion: comparison of twenty methods. *REVSTAT – Statistical Journal*, **6**, 165–197. 1, 2, 5, 6, 8, 9, 10, 19, 29

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 31

Rindskopf, D. (2000). Commentary: Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **54**, 88. 6, 7

Robert, C. P. (2007). *The Bayesian choice*. Springer. 15

Rothman, K. J. (1986). *Modern epidemiology*. Little Brown. 1

Rubin, D. B. and Schenker, N. (1987). Logit-based interval estimation for binomial data using the Jeffreys prior. *Sociological Methodology*, **17**, 131–144. 6

Tuyl, F. A. W. M. (2007). *Estimation of the binomial parameter: in defence of Bayes*. PhD thesis, University of Newcastle, New South Wales, Australia. 29

Vollset, S. E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, **12**, 809–824. 2, 6, 9

Wald, A. and Wolfowitz, J. (1939). Confidence limits for continuous distribution functions. *The Annals of Mathematical Statistics*, **10**, 105–118. 6

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, **9**, 60–62. 4

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209–212. 5

Xie, M.-g. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, **81**, 3–39. 30