Imperial College London

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Variance Stabilizing Transformations for Meta-Analysis of Binary Data

Vuistiner Philippe

Prof. Elena Kulinskaya (Imperial College London)
Prof. Stephan Morgenthaler (EPFL Lausanne)

Master project
Autumn 2009

Acknowledgements

I thank the professor Stephan Morgenthaler to have offered me the opportunity of doing my Master thesis at Imperial College in London. I am also very grateful to professor Elena Kulinskaya for all the time spent on my project and her advices that always allowed me to go on with the work even when problems occured. It has been a great pleasure for me to work under her supervision.

Summary

Variance stabilization constitutes a new approach for computing confidence intervals to compare binomial probabilities. Kulinskaya et al. (2009) developp the method for the risk difference and the results are better than those of the highly applied Newcombe (1998) interval. This Master thesis develops the same approach for both relative risk and odds ratio.

The transformation from risk difference to either relative risk or odds ratio causes some bias and the coverage of the obtained intervals is not satisfying. Thus we compute conditional confidence intervals with the hypergeometric distribution and the obtained results are as good as those of other well-known methods.

This approach is then applied to combine evidences in meta-analysis and simulations are run to compare the results with other methods. We found that our approach performs much better than the widely used inverse variance method and is competitive with the best known Mantel and Haenszel (1959) method.

Contents

List of Figures

Introduction

The binary data are often used in the medical field or in social sciences to study an effect between two groups. It can be for example the effect of a treatment between a treated and a control group. The outcome will be the number of successes and the total number of trials in each group. Let us call X_1, X_2 the successes and n_1, n_2 the number of trials, then X_1 and X_2 are binomial random variables with probabilities p_1 and p_2 respectively. The difference between these two probabilities will attest the difference of the effect between the two groups. Several statistics may be used to compare two probabilities. In this paper we deal with the risk difference $\Delta = p_1 - p_2$, the relative risk $\rho = \frac{p_1}{p_2}$ and the odds ratio $\gamma = \{p_1(1-p_2)\}/\{(1-p_1)p_2\}$.

Since the success probabilities are estimated, so is the parameter of interest, thus a confidence interval should be given. The mainly used method proposes an estimate of the variance and computes the Wald confidence interval. The problem is that, as the variance is estimated and depends on the true value of the parameter, the coverage of these intervals might be much worst than the expected level. Many papers discuss about ways of computing a confidence interval using different approaches: Storer and Kim (1990) developp exact tests, Newcombe (1998) proposes a method based on the Wilson (1927) score method for the single proportion. He compares eleven different methods and concludes that his new way of computing a confidence interval is the most satisfying. Martín Andrés and Tapia García (2004) study unconditional asymptotic tests, Agresti and Min (2005) compare the effect of different Bayesian priors and methods inverting a score test. Brown and Li (2005) compare six methods and conclude that Newcombe's approach is the best one.

Kulinskaya et al. (2009) propose a variance stabilizing transformation (vst) so the obtained statistic has unit variance. Their theory is developed for the risk difference Δ and simulations show that the coverage of their confidence intervals is really satisfying. The new approach works better than the previously known methods and even better than the Newcombe's interval (1998). Since the method of Kulinskaya et al. (2009) works pretty well for the risk difference we will apply the same approach on the two other mostly used statistics, namely the relative risk and the odds ratio to verify if the results are as good as those for the risk difference.

Kulinskaya et al. (2009) combine studies in a meta-analysis with these transformed statistics. Adding the evidences with weights equal to the square root of the sample sizes leads to a combined statistic with unit variance. A confidence interval can then be computed whose coverage is much more reliable than the widely applied inverse variance approach which is still used in most of softwares. The authors besides suggest never to apply this last method.

In this paper we use this variance stabilization for the risk difference and adapt it for both relative risk and odds ratio (Chapter 1). In Chapter 2 simulations of the coverage of the obtained confidence intervals show that the results are not as satisfying as for risk difference and different modifications are tried to attempt to get a good coverage. Finally conditional confidence intervals are used for odds ratio, given the total number of successes in both groups (Chapter 3). Simulations show that this method gives satisfying results, much better than the unconditional previous approach. In Chapter 4 we apply this method to combine evidences in a meta-analysis and compare the results with other well-known methods. Finally a discussion is drawn to summarize the obtained results and to suggest further research interests.

Chapter 1

Variance Stabilizing Transformations

1.1 Variance Stabilisation of the Risk Difference

For two binomial random variables $X_1 \sim \mathcal{B}(n_1, p_1)$ and $X_2 \sim \mathcal{B}(n_2, p_2)$, the risk difference is defined as $\Delta = p_1 - p_2$. Because both probabilities p_1 and p_2 are unknown we need to define a nuisance parameter $\psi = Ap_1 + (1-A)p_2$ for any $0 \leq A \leq 1$. There seems to be no simple solution for the choice of A so in the following we will mostly fix $A = \frac{1}{2}$ as suggested by Kulinskaya et al. (2009). We can express the two probabilities as functions of Δ and ψ as follows:

$$p_1 = \psi + (1 - A)\Delta \qquad \text{and} \qquad p_2 = \psi - A\Delta. \tag{1.1}$$

To satisfy $0 < p_i < 1$ $(i = 1, 2)$ the following condition should be verified:

$$\max\left\{\frac{\psi - 1}{A}, \frac{\psi}{A - 1}\right\} < \Delta < \min\left\{\frac{\psi}{A}, \frac{\psi - 1}{A - 1}\right\} \tag{1.2}$$

for $A \notin \{0, 1\}$; there is no constraint otherwise.

The usual maximum likelihood estimator (MLE) of Δ is $\hat{\Delta} = \frac{x_1}{n_1} - \frac{x_2}{n_2}$ (x_i is a realisation of X_i, $i = 1, 2$) which is unbiased and has a variance equal to

$$\text{var}[\hat{\Delta}] = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}.$$

Actually the values of p_1 and p_2 are unknown so in practice we use the estimated variance $\widehat{\text{var}}[\hat{\Delta}] = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$. This variance is used to compute the well-known Wald confidence interval but the coverage is improved by replacing the MLE $\hat{p}_i$ by $\tilde{p}_i = \frac{x_i + c}{n_i + 2c}$ for different values of c (the most widely used is $c = 0.5$ (Anscombe 1956)). The main problem here is that the variance of $\hat{\Delta}$ depends on the values of p_1 and p_2 and so on its expectation. The aim of a *vst* is to make the variance of a statistic constant.

Let Y be a random variable that depends on a parameter θ with expectation $\mathbb{E}_\theta[Y] = \mu(\theta) = \mu$. If the variance of Y is $\text{var}_\theta[Y] = \sigma^2(\mu)$ then a *vst* is a function $h(y)$ that satisfies $h'(y) = \frac{1}{\sigma(y)}$. The variance of the transformed variable is then approximately one.

Kulinskaya et al. (2009) (Eq. 2.3) show that the *vst* of the risk difference Δ is

$$T_A^\Delta(\hat{\Delta}; \psi, \Delta_0) = \sqrt{\frac{2Nq(1-q)}{u}} \left\{ \arcsin\left(\frac{u\hat{\Delta} + v}{w}\right) - \arcsin\left(\frac{u\Delta_0 + v}{w}\right) \right\}, \tag{1.3}$$

which they refer as the evidence function, where

$$\begin{aligned} N &= n_1 + n_2, \\ q &= \frac{n_2}{N}, \\ u &= 2\left\{(1-A)^2 q + A^2(1-q)\right\}, \\ v &= (1-2\psi)(A-q) \text{ and} \\ w &= \sqrt{2u\psi(1-\psi) + v^2}. \end{aligned}$$

Notice that the arcsine function is only defined on the interval $[-1, 1]$, imposing a condition on Δ to be between $-\frac{w+v}{u}$ and $\frac{w-v}{u}$.

Under the null hypothesis $\Delta = \Delta_0$ the statistic $T_A^\Delta(\hat{\Delta}; \psi, \Delta_0)$ follows a normal $\mathcal{N}(0, 1)$ distribution. We can easily test this null hypothesis or find a confidence interval as the set of all Δ_0 such that $\left|T_A^\Delta(\hat{\Delta}; \psi, \Delta_0)\right| \leq z_{1-\frac{\alpha}{2}}$, the $(1-\frac{\alpha}{2})$-quantile of the normal distribution, leading to the following confidence interval (Kulinskaya et al. 2009, Eq. 2.4):

$$\frac{\hat{w}}{u} \sin\left\{ \arcsin\left(\frac{u\hat{\Delta} + \hat{v}}{\hat{w}}\right) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{u}{2Nq(1-q)}} \right\} - \frac{\hat{v}}{u}, \tag{1.4}$$

where $\hat{v} = (1-2\hat{\psi})(A-q)$ and $\hat{w} = \sqrt{2u\hat{\psi}(1-\hat{\psi}) + \hat{v}^2}$.

1.1.1 Special case when $A = \frac{1}{2}$

The notations simplify if we fix the parameter $A = \frac{1}{2}$. In this case the nuisance parameter becomes $p = \frac{p_1+p_2}{2}$, the mean risk. The *vst* given in Eq.(1.3) reduces to

$$T^\Delta(\hat{\Delta}; p, \Delta_0) = 2\frac{\sqrt{RN}}{R+1} \left\{ \arcsin\left(\frac{\hat{\Delta} + 2v}{2w}\right) - \arcsin\left(\frac{\Delta_0 + 2v}{2w}\right) \right\}, \tag{1.5}$$

(Kulinskaya 2009, Eq. 2) where $R = \frac{n_1}{n_2}$, $v = 2\left(\frac{1}{2} - p\right)\left(\frac{1}{2} - \frac{1}{R+1}\right)$ and $w = \sqrt{p(1-p) + v^2}$. In this case Δ has to be between $-2(w+v)$ and

$2(w-v)$ to satisfy the constraints of the arcsine function. Note that if $p = \frac{1}{2}$ or if $R = 1$ then $v = 0$ and $w = \sqrt{p(1-p)}$ which is the standard error of a Bernoulli random variable with probability p.

Thus the limits of the confidence interval are

$$2\hat{w}\sin\left\{\arcsin\left(\frac{\hat{\Delta}+2\hat{v}}{2\hat{w}}\right) \pm z_{1-\frac{\alpha}{2}}\frac{R+1}{2\sqrt{RN}}\right\} - 2\hat{v}. \tag{1.6}$$

1.2 Variance Stabilisation of the Risk Ratio

In Section 1.1 the risk difference Δ was the parameter of interest but other statistics may also be used, for example the risk ratio $\rho = \frac{p_1}{p_2}$ which is always a positive number. Using Eq.(1.1) the risk ratio can be expressed as a function of Δ and the nuisance parameter ψ as follows:

$$\rho = g(\Delta;\psi,A) = \frac{\psi + (1-A)\Delta}{\psi - A\Delta}. \tag{1.7}$$

The risk difference is between -1 and 1 but the risk ratio ρ is always positive so Δ must lie in the interval

$$\left(\max\left\{-1, \frac{\psi}{A-1}\right\}, \min\left\{1, \frac{\psi}{A}\right\}\right)$$

for $A \notin \{0,1\}$. When $A = 0$ or $A = 1$ then ρ is always positive for all $\Delta \in (-1,1)$. On this domain the function g is monotone increasing so an inverse can be computed as

$$\Delta = g^{-1}(\rho;\psi,A) = \frac{\psi(\rho-1)}{A(\rho-1)+1}. \tag{1.8}$$

The denominator would be equal to zero if $\rho = \frac{A-1}{A}$ (for $A \neq 0$) but this quantity is less than or equal to zero and we know that ρ is positive so the transformation is always defined. Nevertheless Δ should lie between -1 and 1 which imposes a constraint between ψ and ρ that is $\psi < \{A(\rho-1)+1\}\min\left\{1,\rho^{-1}\right\}$. This condition is illustrated in Figure 1.1 for different values of A. So ρ must be larger than $\max\left\{0, 1+\frac{\psi-1}{A}\right\}$ for any value of ψ and moreover if $\psi > A$ then ρ must be smaller than $\frac{1-A}{\psi-A}$.

The *vst* for the risk ratio ρ is found using the composition of the *vst* for Δ and the inverse transformation g^{-1}, giving the following evidence function:

$$T_A^{\rho}(\hat{\rho};\psi,\rho_0) = T_A^{\Delta}\left(g^{-1}(\hat{\rho};\psi,A);\psi,\Delta_0\right). \tag{1.9}$$

To find a confidence interval for ρ we can simply use Eq.(1.4), replace $\hat{\Delta}$ by $g^{-1}(\hat{\rho};\psi,A)$ and then apply the transformation g to the obtained limits.

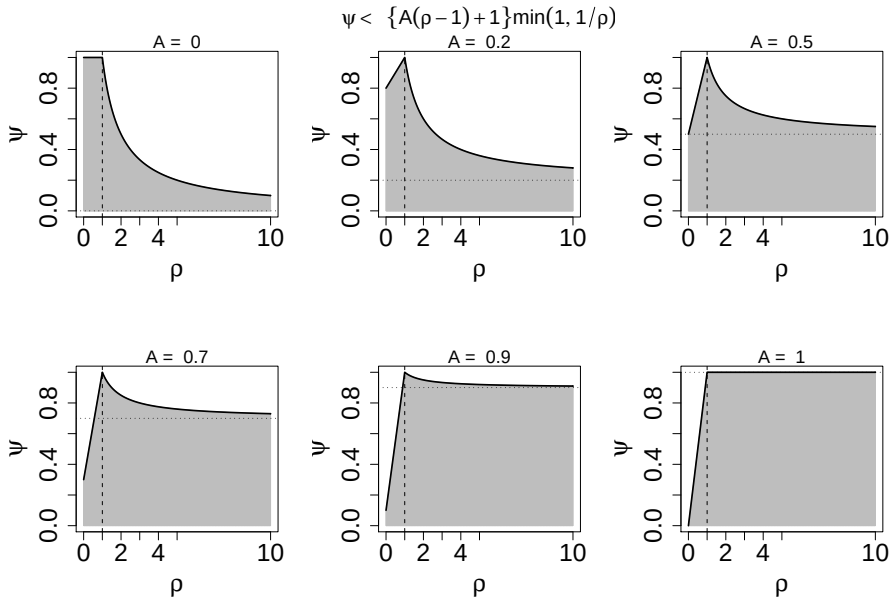

Figure 1.1: Restrictions between ρ and the nuisance parameter ψ for different values of A, $A \in \{0, 0.2, 0.5, 0.7, 0.9, 1\}$.

Nevertheless we have to check whether the limits of the confidence interval for Δ lie between $\frac{\psi}{A-1}$ and $\frac{\psi}{A}$ before applying the transformation g, otherwise it will return negative values. If the lower limit of the confidence interval for Δ is smaller than $\frac{\psi}{A-1}$ then the lower limit for ρ will be zero. If the upper limit for Δ is larger than $\frac{\psi}{A}$ we take the interval for ρ up to infinity.

1.2.1 Special case when $A = \frac{1}{2}$

As in the previous section we compute the vst for a fixed value of $A = \frac{1}{2}$. In this case the transformation simplifies to

$$\rho = g(\Delta; p) = \frac{2p + \Delta}{2p - \Delta}$$

and the inverse transformation is

$$\Delta = g^{-1}(\rho; p) = \frac{2p(\rho - 1)}{\rho + 1}.$$

The function g is not defined for $\Delta = 2p = p_1 + p_2$ but as $p_2 > 0$ then it is impossible that Δ takes this value. The inverse transformation is defined for all $\rho > 0$. To satisfy the condition for Δ to be between -1 and 1, p must be smaller than $\frac{\rho+1}{2} \min\left\{1, \rho^{-1}\right\}$. This condition is illustrated in Figure 1.1 on the top right panel. If $p \leq 0.5$ all values of ρ are allowed but when p is

larger ρ must lie between $2p-1$ and $\frac{1}{2p-1}$. Nevertheless if p is very close to zero and ρ is quite large then p_2 must be almost zero. In this case when the binomial random variable X_2 is generated it will be zero most of the time and so the probability is estimated by $\hat{p}_2 = \frac{1}{2(n_2+1)}$. The problem is when n_2 is small because then $\hat{p}_2 \gg p_2$ and so $\hat{\rho} \ll \rho$. So the confidence interval will never cover the true value of the risk ratio.

Thus for the risk ratio the only condition to check is the constraint between ρ and the nuisance parameter ψ otherwise all the given transformations are well-defined whatever the values of the different parameters are.

1.3 Variance Stabilisation of the Odds Ratio

Another statistic of interest is the odds-ratio $\gamma = \frac{p_1(1-p_2)}{(1-p_1)p_2}$ that we can also express as a function of Δ and ψ. Using Eq.(1.1) and doing few computation we get

$$\gamma = f(\Delta; \psi, A) = \frac{a\Delta^2 + b\Delta + c}{a\Delta^2 + (b-1)\Delta + c} \tag{1.10}$$

with $a = A(1-A)$, $b = \psi A + (1-\psi)(1-A)$ and $c = \psi(1-\psi)$. We first consider the case where A is different from 0 or 1. To get a positive odds-ratio the condition on Δ is the following:

$$\Delta \in \left(\max\left\{-1, \frac{-b+\sqrt{b^2-4ac}}{2a}\right\}, \min\left\{1, \frac{1-b-\sqrt{b^2-4ac+1-2b}}{2a}\right\}\right).$$

In the cases when $A = 0$ or $A = 1$ the domain of definition for Δ is

$$\Delta \in \left(\max\left\{-1, -\frac{c}{b}\right\}, \min\left\{1, \frac{c}{1-b}\right\}\right).$$

In both domains the function f is monotone increasing so an inverse can be defined as

$$\Delta = f^{-1}(\gamma; \psi, A) = \begin{cases} \dfrac{\gamma - b(\gamma-1) - \sqrt{\{b(\gamma-1)-\gamma\}^2 - 4ac(\gamma-1)^2}}{2a(\gamma-1)} & \\ & \text{, if } A \neq 0 \text{ and } A \neq 1 \\ \dfrac{c(\gamma-1)}{(1-b)\gamma + b} & \text{, if } A = 0 \text{ or } A = 1. \end{cases} \tag{1.11}$$

Of course f^{-1} is not defined for $\gamma = 1$ ($A \notin \{0, 1\}$) but in this case we know that $\Delta = 0$.

The argument under the square root in Eq.(1.11) can be written as

$$(b^2 - 4ac + 1 - 2b)\gamma^2 - 2(b^2 - 4ac - b)\gamma + b^2 - 4ac$$

and the factor of the quadratic term is thought to be always positive (it is easy to prove it for $A = 0.5$). To satisfy the non-negativity of the square root, γ needs to lie outside the interval

$$\left[\frac{b^2 - 4ac - b - 2\sqrt{ac}}{b^2 - 4ac - 2b + 1}, \frac{b^2 - 4ac - b + 2\sqrt{ac}}{b^2 - 4ac - 2b + 1}\right],$$

the bounds being defined by the two roots of the polynomial.

To compute a confidence interval for γ we use Eq.(1.4), check if the limits of this interval fall in the domain of definition of f and then apply the transformation f on the two limits. If the lower limit of the interval for Δ is outside the domain then the lower bound for γ is zero and if the upper limit of Δ is outside then the upper bound for γ goes up to infinity.

1.3.1 Special case when $A = \frac{1}{2}$

With $A = \frac{1}{2}$, the transformation f simplifies to

$$\gamma = f(\Delta; p) = \frac{\Delta^2 + 2\Delta + 4p(1-p)}{\Delta^2 - 2\Delta + 4p(1-p)}$$

and since this is a monotone function in Δ, the inverse can be computed as follows:

$$\Delta = f^{-1}(\gamma; p) = \begin{cases} \dfrac{\gamma + 1 - \sqrt{(\gamma+1)^2 - 4p(1-p)(\gamma-1)^2}}{\gamma - 1} & \text{, if } \gamma \neq 1 \\ 0 & \text{, otherwise.} \end{cases}$$

The transformation f is not defined for $\Delta = 1 - \sqrt{1 - 4p(1-p)}$ in which case the denominator is zero. The quadratic polynomials give two solutions but only this one is valid because Δ should lie between -1 and 1. Moreover the argument under the square root must be non-negative but this is always the case because $4p(1-p) < 0$ for all p as $p = \frac{p_1+p_2}{2}$ and so

$$(\gamma+1)^2 - 4p(1-p)(\gamma-1)^2 \geq (\gamma+1)^2 - (\gamma-1)^2 = 4\gamma > 0.$$

Thus the function f^{-1} is always well-defined.

Chapter 2

Simulations for One Study

The transformations presented in Chapter 1 allow to compute confidence intervals for both relative risk and odds ratio starting from the results of Kulinskaya et al. (2009). With monotone transformations the behaviour of the confidence intervals would be the same as for the risk difference Δ but these functions have to be restricted in a particular domain to be monotone. In this chapter simulations test the coverage of these intervals to attest the efficiency of the method.

2.1 Confidence Intervals for the Risk Ratio

For each simulation the input is the total sample size N, the proportion between the two samples $R = \frac{n_1}{n_2}$ (this allows to find $n_2 = \left\lceil \frac{N}{R+1} \right\rceil$ and $n_1 = N - n_2$) and a value for the nuisance parameter $\psi = Ap_1 + (1-A)p_2$ for a given A. Then for any value of ρ the probabilities $p_1 = \frac{\psi\rho}{A(\rho-1)+1}$ and $p_2 = \frac{\psi}{A(\rho-1)+1}$ are determined and binomial samples are generated with respective parameters (n_1, p_1) and (n_2, p_2). We can choose whether or not to remove studies with 0 or n_i events $(i = 1, 2)$ in both arms and there is also an option to decide how to estimate the probability of success $\hat{p}_i = \frac{x_i+c}{n_i+2c}$ $(i = 1, 2)$. We can give the value of c and decide to add this c always or only if $x_i = 0$ or $x_i = n_i$ $(i = 1, 2)$ and use the MLE otherwise. For each pair of samples the 95% confidence interval defined in Section 1.2 is computed. We also test whether the true value of ρ lies in the interval and thus we can compute the mean coverage level. The average length of the confidence intervals is also computed, in the case where the length is finite.

To present the results of the simulations we can either fix the value of ρ and give the coverage of the intervals for a range of ψ-values, or fix ψ and plot the results as a function of ρ. Another way is to represent the results on the log scale to better see what is happening for ρ between 0 and 1. When plotting as a function of ψ for a fixed value of ρ we represent the x-axis as

the average probability $\bar{p}$, so if A is changed the results are still presented with the same scale. To do this we simply express $\bar{p}$ as a function of ρ, ψ and A using Eq.(1.1) and (1.8). We get the following result:

$$\bar{p} = \frac{\rho + 1}{2A(\rho - 1) + 2}\psi,$$

so, given a range of values for ψ, we can plot the results as a function of $\bar{p}$ using this simple transformation which is basically only a rescaling of the x-axis.

To attest the quality of the variance stabilized confidence intervals, they need to be compared with the ones obtained with other well-known methods. We use three other confidence intervals. Two very similar methods are those of Woolf (1955) and Gart (1966). The Woolf method is described by Agresti and Min (2002) and studied by Brown (1981) who concludes that this method is reasonable for large sample sizes. Agresti (1999) reports both Woolf and Gart methods which define confidence intervals for the log relative risk with the estimated variance

$$\text{var}[\log(\hat{\rho})] = \frac{1}{x_1} + \frac{1}{x_2} - \frac{1}{n_1} - \frac{1}{n_2},$$

so the confidence limits are given with $\log(\hat{\rho}) \pm z_{1-\frac{\alpha}{2}}\text{var}[\log(\hat{\rho})]$ and the exponential of these limits gives the confidence interval for ρ. The difference between these two methods lies in the way of estimating the risk ratio $\hat{\rho}$. Woolf (1955) defines $\hat{\rho} = \frac{x_1/n_1}{x_2/n_2}$ and modifies it only if one of the x_i's is equal to 0 or n_i replacing x_i by $x_i + 0.5$ and n_i by $n_i + 1$ $(i = 1, 2)$. This avoids a null variance if both $x_1 = n_1$ and $x_2 = n_2$. Gart (1966) always estimates $\hat{\rho}$ using $x_i + 0.5$ and $n_i + 1$ whatever the values of x_i $(i = 1, 2)$ are. These two methods are known to perform pretty well for large sample sizes.

The third method is based on inverting a score test and is suggested by Koopman (1984) and Miettinen and Nurminen (1985). Agresti and Min (2005) compare this method to Bayesian confidence intervals and recommend the use of score intervals that tend to be better in terms of coverage probability. The code used for the simulations is available[1] and has been written by Y. Min where confidence intervals for all risk difference, relative risk and odds ratio are computed.

Figures 2.1 to 2.4 present the results of the simulations. A thousand replications are performed and the average coverage probability and length of the confidence intervals are computed. The x-axis of these plots displays a hundred equally-spaced values of the average probability $\bar{p} = \frac{p_1+p_2}{2}$. Different values of the relative risk ρ and allocation ratio $R = \frac{n_1}{n_2}$ are tested with a total sample size of $N = 100$ and with $A = \frac{1}{2}$. The plots present the coverage of Δ given by Eq.(1.4) and the coverage of the convidence interval

[1] `http://www.stat.ufl.edu/~aa/cda/R/two_sample/R2` visited in January 2010.

for ρ found by transforming the interval for Δ so we can see if the transformation from Δ to ρ affects the coverage. Three other methods are used to compare the results: the methods of Gart (1966), Woolf (1955) and Agresti and Min (2005). The black line represents the coverage of the confidence intervals found with the modified transformation based on the expected value of ρ described later in Section 2.3.

Figures 2.1 and 2.2 show that when $\rho = 1$ the *vst* method (red line) performs quite well, being a little bit conservative for very small or large values of $\bar{p}$ but for all values between 0.1 and 0.9, the coverage remains close to the expected 95% in both balanced and unbalanced cases. Notice that the coverage with the *vst* is exactly the same as the coverage of Δ which shows that, in this case, applying the transformation does not affect the coverage. The other methods perform well too, except the Woolf method for $\bar{p}$ between 0.7 and 0.9 when $R = 3$. In the unbalanced case the *vst* and score methods behave very similarly for all $\bar{p}$. Nevertheless, both Gart and Woolf methods are too conservative for small $\bar{p}$ so the score method would be prefered here. Moreover the average length of the score intervals is much smaller than all other methods especially for small $\bar{p}$.

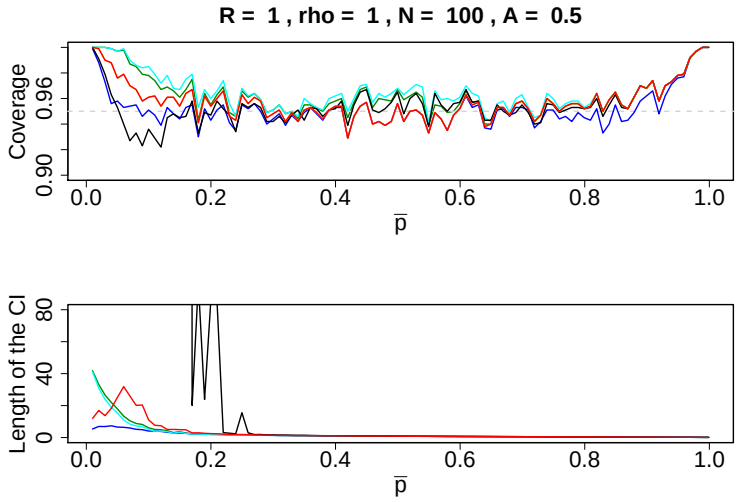

Figure 2.1: Plots of the coverage probabilities (top) and the lengths (bottom) of the nominal 95% confidence intervals given with the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively), the score method (blue line) and the *vst* with the modified transformation (black line). The orange line represents the coverage for Δ with the *vst* method from Kulinskaya et al. (2009). The horizontal line is at 0.95. These plots are based on balanced sampling with $n_1 = n_2 = 50$. The relative risk is taking the value $\rho = 1$ and the x-axis displays the whole possible range of values for the nuisance parameter.

The troubles appear when $\rho = 3$. In the balanced case (Figure 2.3), when $\bar{p}$ is larger than 0.4, the coverage of the *vst* (red line) starts going

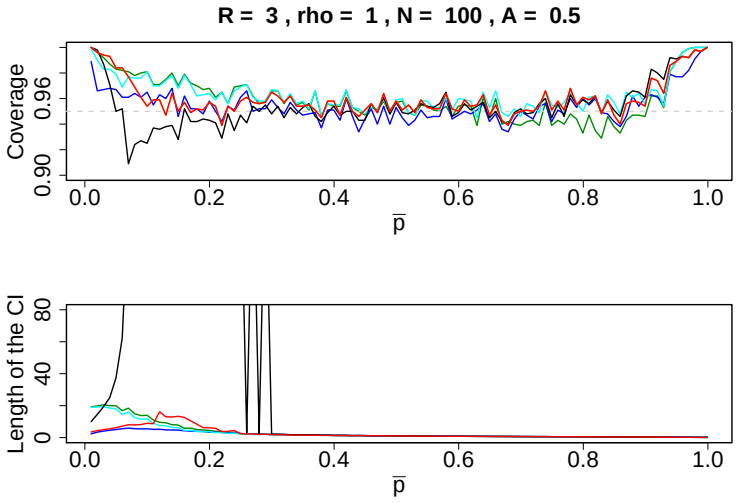

Figure 2.2: Plots of the coverage probabilities (top) and the lengths (bottom) of the nominal 95% confidence intervals given with the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively), the score method (blue line) and the *vst* with the modified transformation (black line). The orange line represents the coverage for Δ with the *vst* method from Kulinskaya et al. (2009). The horizontal line is at 0.95. These plots are based on unbalanced sampling with $n_1 = 75$, $n_2 = 25$. The relative risk is taking the value $\rho = 1$ and the x-axis displays the whole possible range of values for the nuisance parameter.

down, falling under 90% for $\bar{p} > 0.55$ whereas it was too conservative for small $\bar{p}$. We see here that the problem comes from the transformation from Δ to ρ because the coverage of Δ (orange line) is very satisfying all over the range of $\bar{p}$. It becomes even much worst in the unbalanced case (Figure 2.4). The coverage of Δ is too conservative for small values of $\bar{p}$ but when $\bar{p} > 0.15$ it perfoms very well. For ρ the coverage is disastrous for all $\bar{p} > 0.15$ which clearly shows the bias caused by the transformation. As expected the score method is again the one that behaves the best, giving a good coverage with much shorter intervals than the Gart and Woolf methods, even if their coverages are quite similar.

We also tried to simulate with the same parameters but with $N = 1000$ to see if the *vst* method behaves better for large sample sizes. All the results are not shown here but for $\rho = 1$ all methods give similar coverage and the lengths of the intervals are much shorter than when $N = 100$. The methods are not conservative any more in the unbalanced case but no big difference is noticeable. When $\rho = 3$ the Gart, Woolf and score methods are all very good but the *vst* still has a wrong behaviour either in the balanced or unbalanced case. Figure 2.5 shows the unbalanced case when $\rho = 3$. The level of coverage is even worst than with $N = 100$ (Figure 2.4) with a coverage falling down to 80% for large $\bar{p}$. The initial coverage of Δ (orange line) is

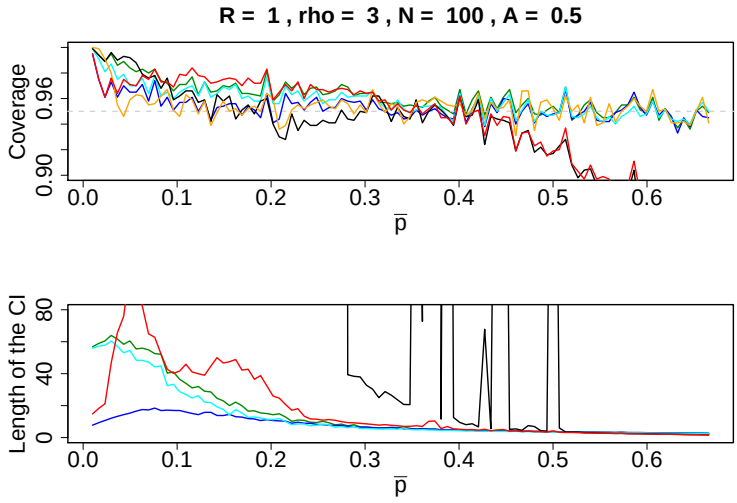

Figure 2.3: Plots of the coverage probabilities (top) and the lengths (bottom) of the nominal 95% confidence intervals given with the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively), the score method (blue line) and the *vst* with the modified transformation (black line). The orange line represents the coverage for Δ with the *vst* method from Kulinskaya et al. (2009). The horizontal line is at 0.95. These plots are based on balanced sampling with $n_1 = n_2 = 50$. The relative risk is taking the value $\rho = 3$ and the x-axis displays the whole possible range of values for the nuisance parameter..

improved by increasing the sample sizes which shows that the problem really comes from the transformation from Δ to ρ.

2.2 Modification of the Confidence Interval

The transformation of the confidence interval for Δ into a confidence interval for ρ causes a lost of the coverage due to the fact that the limits have to be restricted to lie into $(\frac{\psi}{A-1}, \frac{\psi}{A})$, the definition domain of the function g given in Eq.(1.7). To avoid loosing coverage the idea is that if one limit of the confidence interval for Δ is restricted to lie into the definition domain of g then we move the other limit such that the coverage of the interval remains 95%. If the lower bound is restricted to $\frac{\psi}{A-1}$ then the upper bound is modified as

$$U = T_A^{-1}\left(\Phi^{-1}\left[0.95 + \Phi\left\{T_A\left(\frac{\psi}{A-1}\right)\right\}\right]\right),$$

(see Appendix A.1 for details of computation) and if the upper bound is restricted to $\frac{\psi}{A}$ then the lower bound is

$$L = T_A^{-1}\left(\Phi^{-1}\left[0.95 - \Phi\left\{T_A\left(\frac{\psi}{A}\right)\right\}\right]\right).$$

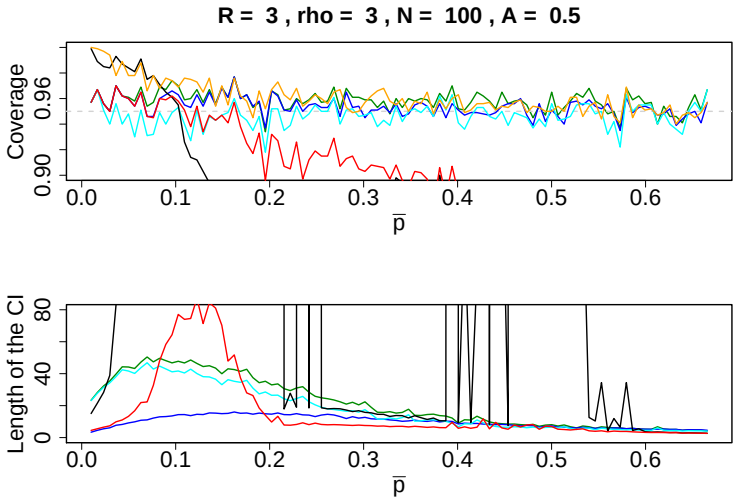

Figure 2.4: Plots of the coverage probabilities (top) and the lengths (bottom) of the nominal 95% confidence intervals given with the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively), the score method (blue line) and the *vst* with the modified transformation (black line). The orange line represents the coverage for Δ with the *vst* method from Kulinskaya et al. (2009). The horizontal line is at 0.95. These plots are based on balanced sampling with $n_1 = 75$, $n_2 = 25$. The relative risk is taking the value $\rho = 3$ and the x-axis displays the whole possible range of values for the nuisance parameter.

The explicit inverse function of T_A is easily computed. Unfortunately if both limits lie outside the domain we can do nothing to avoid loosing coverage but simulations show that this hardly ever happens. Once we get this new confidence interval for Δ we can apply the transformation g to these limits and obtain a confidence interval for ρ.

Figure 2.6 illustrates the coverage of the confidence intervals for both Δ and ρ before and after moving the limits. The results are presented for fixed nuisance parameter ψ and allocation ratio R, and for the whole range of possible values of Δ. We see that when $\psi = 0.5$ (top plots) the restriction of the limits does not change the coverage of the intervals (the red and the blue lines, corresponding to the coverage of Δ before and after the restrictions respectively, are superposed) but once they are transformed into intervals for ρ their coverage does not remain as good as they were for Δ because of the bias induced by the transformation. In the bottom plots, when $\psi = 0.1$, the restriction of the interval astonishingly leads to quite a large loss of coverage. This is due to large frequently needed restrictions. In this case the definition domain of the function g narrows to $(-0.2, 0.2)$ so if we restrict one limit of the interval we can not move the other one far enough such that the coverage remains 95%. Restrictions are needed for almost all simulated confidence intervals when $\psi = 0.1$: for negative values of Δ the lower bound

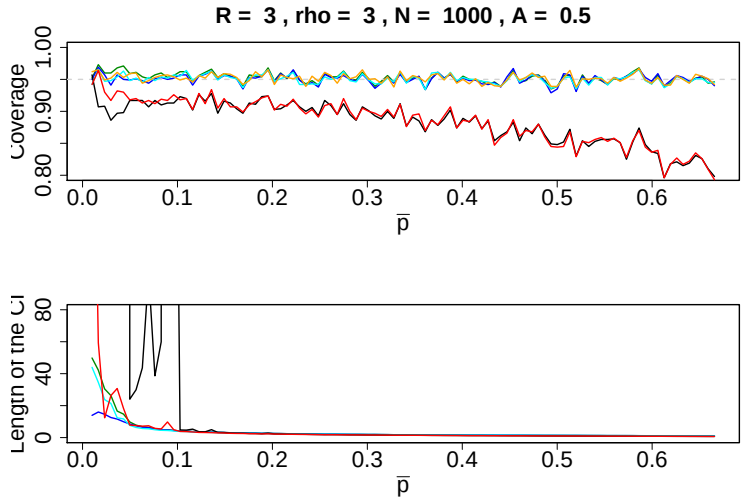

Figure 2.5: Plots of the coverage probabilities (top) and the lengths (bottom) of the nominal 95% confidence intervals given with the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively), the score method (blue line) and the *vst* with the modified transformation (black line). The orange line represents the coverage for Δ with the *vst* method from Kulinskaya et al. (2009). The horizontal line is at 0.95. These plots are based on balanced sampling with $n_1 = 750$, $n_2 = 250$. The relative risk is taking the value $\rho = 3$ and the x-axis displays the whole possible range of values for the nuisance parameter.

must be moved and for positive values of Δ the upper bound is moved up to more than 90% of the time whereas the modification was needed in about only 20% of the cases with $\psi = 0.5$. The more extreme the value of Δ is the more often a modification is needed that is why the transformation is less biased in the middle of the plots.

Exactly the same procedure can be applied for the odds ratio where the restriction domain is given in Section 1.3. Since these modifications of the confidence intervals do not give the expected results we do not use this approach anymore in the remaining of this study. We then try another modification based on the expected value.

2.3 Modification of the Transformation Based on the Expected Value of the Risk Ratio

In Section 1.2 the risk ratio ρ was computed as $g(\Delta)$ but if we use the evidence function given in Eq.(1.3) we get

$$\kappa = h(\Delta) = \frac{1}{\sqrt{N}} T_A^{\Delta}(\Delta).$$

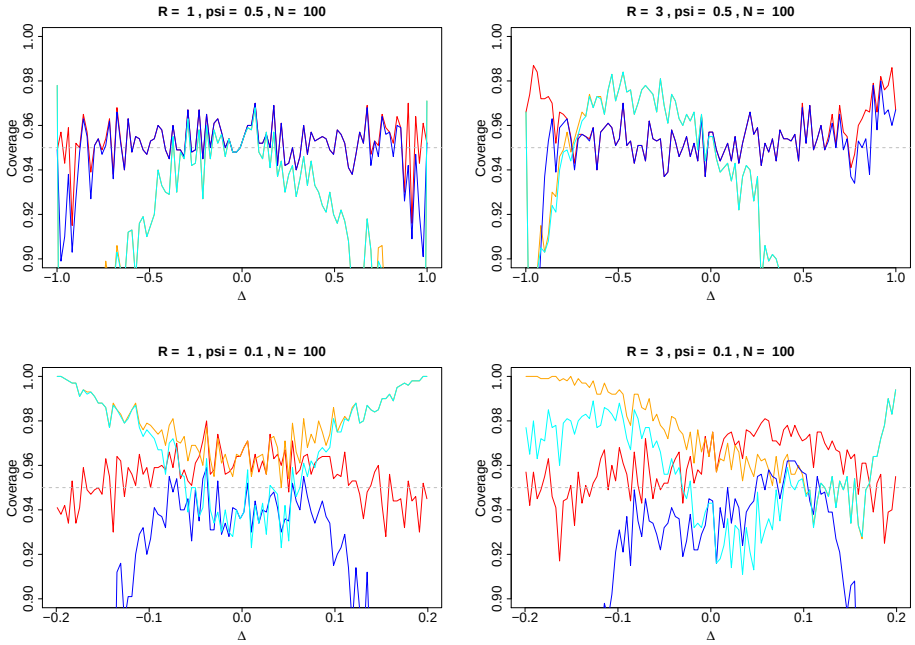

Figure 2.6: Coverage of the nominal 95% confidence intervals as a function of Δ for different values of R, $R \in \{1, 3\}$, and ψ, $\psi \in \{0.1, 0.5\}$. The total sample size is $N = 100$. The red line shows the coverage of the confidence interval for Δ of Eq.(1.4) and the blue line presents the coverage of this interval once the limits have been modified according to the above explanation. These two intervals are transformed into confidence intervals for ρ (orange and cyan lines respectively).

Then we can write ρ as a function of κ as follows:

$$\rho = (g \circ h^{-1})(\kappa)$$

and compute the expected value of ρ. For any function f the expected value of $f(X)$ is

$$\mathbb{E}[f(X)] \approx f(\mathbb{E}[X]) + \frac{f''(\mathbb{E}[X])}{2}\text{var}[X].$$

Here $\mathbb{E}[\kappa] = 0$ and $\text{var}[\kappa] = \frac{1}{N}$ as, under $H_0 : \Delta = \Delta_0$, $T_A^{\Delta} \sim \mathcal{N}(0, 1)$, thus

$$\mathbb{E}[\rho] = \frac{\psi + (1 - A)\Delta_0}{\psi - A\Delta_0} - \frac{\psi(u\Delta_0 + v)}{4Nq(1 - q)(\psi - A\Delta_0)^2} + \frac{A\psi\{w^2 - (u\Delta_0 + v)^2\}}{2Nuq(1 - q)(\psi - A\Delta_0)^3} \tag{2.1}$$

(see Appendix A.2 for the details of computation).

The idea is to use Eq.(2.1) to transform the confidence interval for Δ into an interval for ρ as we did before with the function g (Eq.(1.7)). The above transformation is basically only a correction of the function g. Once we get the interval for Δ we apply this transformation on the bounds to

get an interval for ρ. The coverage of the confidence intervals found with this transformation is shown in Figures 2.1 to 2.4 (black line). It appears that the coverage is even worst than the one with the transformation g (red line). Moreover the lengths of these intervals are often huge, giving no precise information on the value of the parameter. It would be interesting to test whether the length is most of the time large or only rarely. However we are not able to see it applying the mean length, a more robust method should be applied like a trimmed mean or the median. In the simulation with a total sample size of $N = 1000$ (Figure 2.5) the coverage is very close to the one with the standard vst (red line) because as N increases $\mathbb{E}[\rho]$ tends to $g(\Delta)$ so the two methods become similar. In this case the lengths are much shorter except for very small $\bar{p}$.

A similar computation can be performed for the odds ratio leading to

$$\mathbb{E}[\gamma] = \frac{a\Delta_0^2 + b\Delta_0 + c}{a\Delta_0^2 + (b-1)\Delta_0 + c} - \frac{(-a\Delta_0 + c)(u\Delta_0 + v)}{4Nq(1-q)\{a\Delta_0^2 + (b-1)\Delta_0 + c\}^2} + \frac{\{2a^2\Delta_0^3 - 6ac\Delta_0 - 2c(b-1)\}\{w^2 - (u\Delta_0 + v)^2\}}{4Nuq(1-q)\{a\Delta_0^2 + (b-1)\Delta_0 + c\}^3}.$$

(see Appendix A.3 for the details of computation) but since the results are not improved with this transformation we will not apply it in the later simulations.

2.4 Relation between the True and the Simulated Values of the Risk Ratio

To check whether the vst gives the expected results we compute the bias between the true value of the variance stabilized relative risk and the value obtained with the simulated values. We also compute the sample variance of the simulated values to see if it lies close to one.

Two binomial samples of $10,000$ replications are generated, for each couple the evidence function $T_A^{\rho}(\hat{\rho}; \psi, \rho_0)$ given in Eq.(1.9) is computed with $\hat{\rho}$ as well as with the modified transformation from Eq.(2.1). Then the mean and sample variance of these functions are computed. Figures 2.7 to 2.9 show the results of the simulations for different values of ψ, R and N. The black curves are obtained using $\hat{\rho}$ whereas the red ones are those based on the expected value. On all these figures we see that a large N (right panels) reduces the bias. When $\psi = 0.5$ (Figure 2.7) the sample variance is always close to one even for small N but for $\psi = 0.1$ or 0.9 (Figure 2.8 and Figure 2.9 respectively) it can be quite different from one especially for small values of N (left panels). Notice that when ψ is large there is no big difference between the value with $\hat{\rho}$ and with the modified transformation but otherwise the difference can be quite important. The bias is always positive and larger with the transformation based on the expected value.

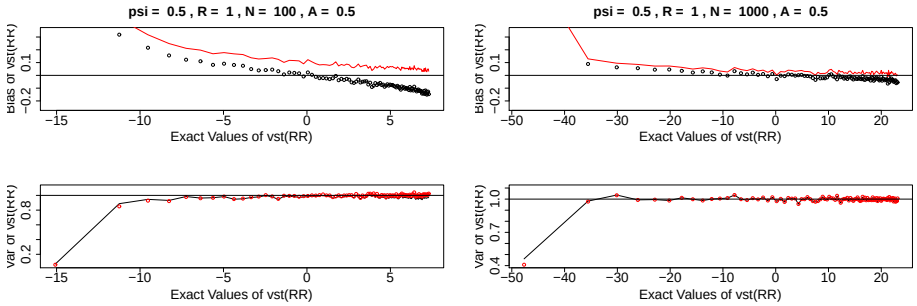

Figure 2.7: Bias and sample variance (top and bottom panels respectively) of the variance stabilized value of the relative risk for balanced sample sizes of $n_1 = n_2 = 50$ (left panels) and $n_1 = n_2 = 500$ (right panels). The value of the nuisance parameter is $\psi = 0.5$. The black line represents the results with $\hat{\rho}$ whereas a red line is used for the transformation based on $\mathbb{E}[\rho]$. In the labels RR denotes the relative risk.

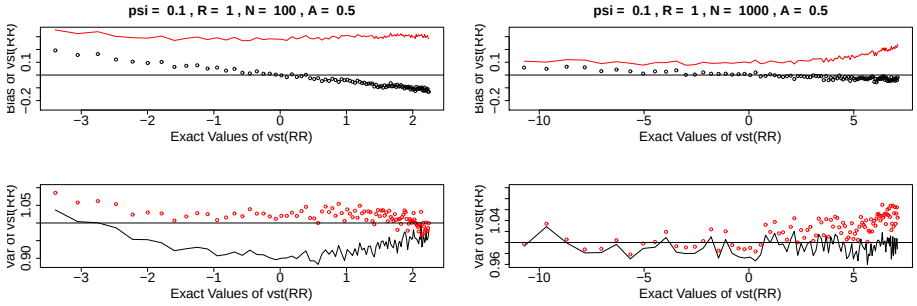

Figure 2.8: Bias and sample variance (top and bottom panels respectively) of the variance stabilized value of the relative risk for balanced sample sizes of $n_1 = n_2 = 50$ (left panels) and $n_1 = n_2 = 500$ (right panels). The value of the nuisance parameter is $\psi = 0.1$. The black line represents the results with $\hat{\rho}$ whereas a red line is used for the transformation based on $\mathbb{E}[\rho]$. In the labels RR denotes the relative risk.

2.5 Confidence Intervals for the Odds Ratio

The procedure to compute the coverage of confidence intervals for odds ratio is the same as for relative risk except the value of γ instead of ρ in the input. After having generated the binomial samples the confidence interval is determined using the results of Section 1.3. The presentation of the results is the same as for the risk ratio, namely as a function of ψ or as a function of γ or its logarithm. If the results are expressed as a function of ψ the x-axis of the plot will be $\bar{p}$ so a different choice of A does not affect the scale. To express $\bar{p}$ as a function of ψ and γ we first compute Δ using Eq.(1.11) and

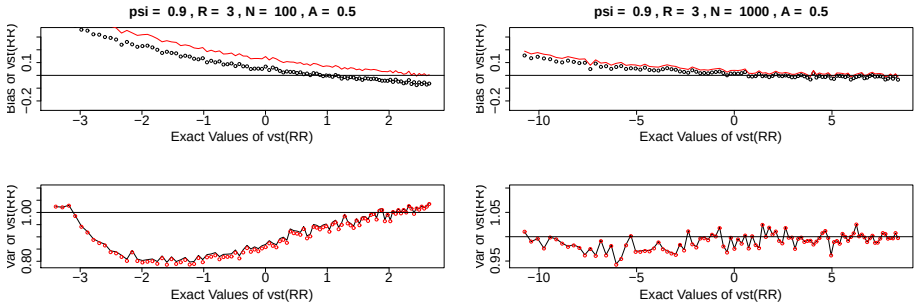

Figure 2.9: Bias and sample variance (top and bottom panels respectively) of the variance stabilized value of the relative risk for unbalanced sample sizes of $n_1 = 75$, $n_2 = 25$ (left panels) and $n_1 = 750$, $n_2 = 250$ (right panels). The value of the nuisance parameter is $\psi = 0.9$. The black line represents the results with $\hat{\rho}$ whereas a red line is used for the transformation based on $\mathbb{E}[\rho]$. In the labels RR denotes the relative risk.

then $\bar{p} = \psi + \frac{1-2A}{2}\Delta$.

The same methods are also used to compare the results. The Woolf (1955) and Gart (1966) confidence intervals for the log odds ratio are computed as follows:

$$\log(\hat{\gamma}) \pm z_{1-\frac{\alpha}{2}} \text{var}[\log(\hat{\gamma})],$$

with the sampling variance

$$\text{var}[\log(\hat{\gamma})] = \frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_2} + \frac{1}{n_2 - x_2}. \tag{2.2}$$

The approach recommended by Agresti and Min (2005) that consists of inverting a score test is also applied. The code has been written by Y. Min based on the papers of Cornfield (1956) and Miettinen and Nurminen (1985).

Figures 2.10 to 2.13 show the results of the simulations for different values of γ and R with $N = 100$ and for the whole range of values of ψ. The four different methods are represented and the coverage of Δ with the vst is also plotted so we can compare it with the coverage for odds ratio to see whether the transformation from Δ to γ causes some bias.

In Figures 2.10 and 2.11, where the odds ratio is one, we see that the coverage is pretty good for all tested methods for $\bar{p}$ between 0.1 and 0.9 and they are conservative for the other values of $\bar{p}$. We do not see the coverage of Δ on these plots because it is exactly the same as the one with the vst (red line). This shows that the applied transformation works perfectly in this case when $\gamma = 1$. As with the relative risk, the score method should be prefered for very small or large values of $\bar{p}$ because it is less conservative than the others and the lengths of the intervals are much smaller. But in

the other cases, our *vst* method performs a bit better, especially in the unbalanced case where the score method is often a bit liberal, even if all methods give very similar coverage and length for $\bar{p}$ between 0.2 and 0.8. The main drawback of the score method is the huge needed computation time so when the results are similar we would recommend another method.

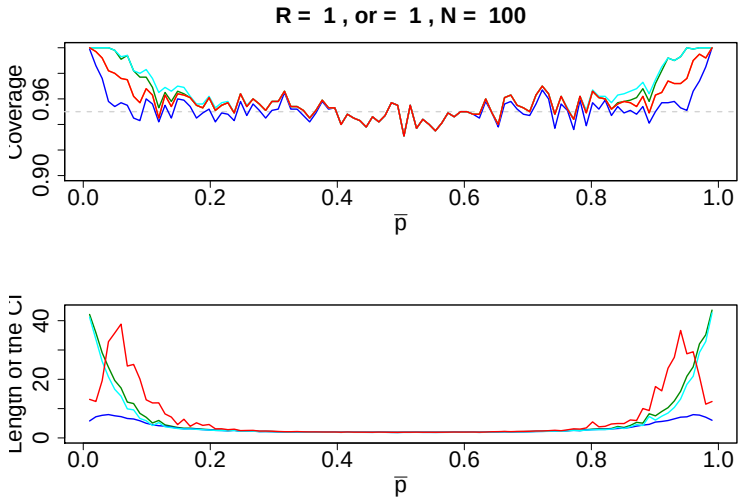

Figure 2.10: Plots of the coverage probabilities (top) and the lengths (bottom) of the nominal 95% confidence intervals given with the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively) and the score method (blue line). The orange line represents the coverage for Δ with the *vst* method from Kulinskaya et al. (2009). The horizontal line is at 0.95. These plots are based on balanced sampling with $n_1 = n_2 = 50$. The odds ratio is taking the value $\gamma = 1$ and the x-axis displays the whole possible range of values for the nuisance parameter.

When the odds ratio equals three (Figures 2.12 and 2.13) a bias appears with the transformation from Δ to γ. In the balanced case the coverage of Δ (orange line) is very satisfying but it becomes too conservative once transformed into intervals for γ (red line). The difference is much more important in the unbalanced case where the *vst* method can clearly not be used. Even if the results are not as dramatic as with the relative risk, the coverage is clearly biased by the transformation, it is liberal for $\bar{p} < 0.5$ and conservative for larger $\bar{p}$. The methods of Gart and Woolf are too conservative and once again the score method is the best one, not being too conservative and with much smaller intervals.

Either with relative risk or odds ratio the coverage of the confidence intervals computed with the *vst* is satisfying under the null hypothesis when the parameter equals one. For a larger value of the parameter of interest the coverage becomes far unsatisfying especially with unbalanbed studies because the function applied to transform Δ causes some bias. We tried to improve the results either by moving the limits before the transformation

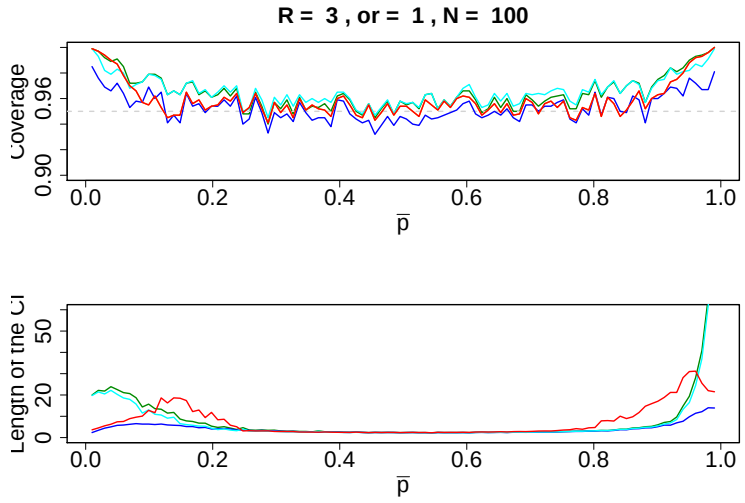

Figure 2.11: Plots of the coverage probabilities (top) and the lengths (bottom) of the nominal 95% confidence intervals given with the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively) and the score method (blue line). The orange line represents the coverage for Δ with the *vst* method from Kulinskaya et al. (2009). The horizontal line is at 0.95. These plots are based on unbalanced sampling with $n_1 = 75$, $n_2 = 25$. The odds ratio is taking the value $\gamma = 1$ and the x-axis displays the whole possible range of values for the nuisance parameter.

or by modifying the transformation but none of the modifications gave acceptable results. Thus the *vst* method is satisfying when the parameter of interest is taking the value $\rho = 1$ ($\gamma = 1$) giving results similar to the other methods and with a short computational time. Nevertheless the bias caused by the transformation is too important to be able to use this method for values of the relative risk or odds ratio different from one and especially for unbalanced sample sizes.

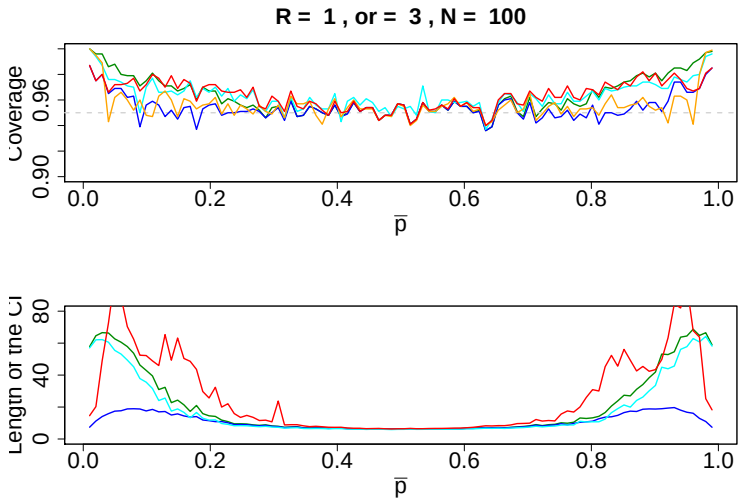

Figure 2.12: Plots of the coverage probabilities (top) and the lengths (bottom) of the nominal 95% confidence intervals given with the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively) and the score method (blue line). The orange line represents the coverage for Δ with the *vst* method from Kulinskaya et al. (2009). The horizontal line is at 0.95. These plots are based on balanced sampling with $n_1 = n_2 = 50$. The odds ratio is taking the value $\gamma = 3$ and the x-axis displays the whole possible range of values for the nuisance parameter.

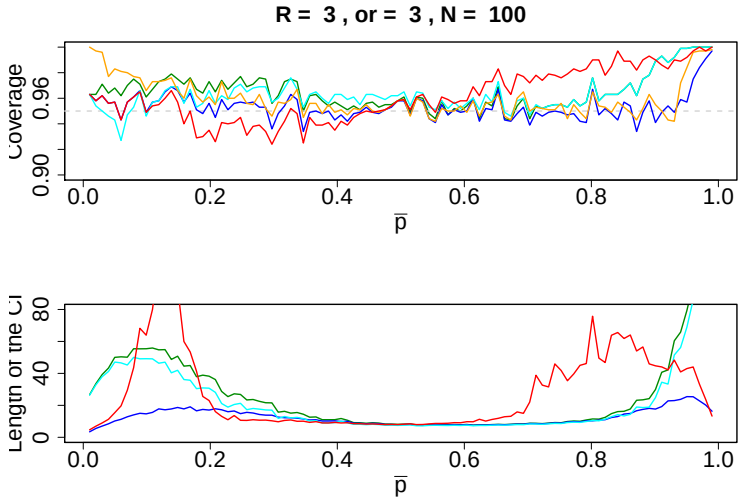

Figure 2.13: Plots of the coverage probabilities (top) and the lengths (bottom) of the nominal 95% confidence intervals given with the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively) and the score method (blue line). The orange line represents the coverage for Δ with the *vst* method from Kulinskaya et al. (2009). The horizontal line is at 0.95. These plots are based on unbalanced sampling with $n_1 = 75$, $n_2 = 25$. The odds ratio is taking the value $\gamma = 3$ and the x-axis displays the whole possible range of values for the nuisance parameter.

Chapter 3

Conditional Confidence Intervals

Since the results of the previous chapter are not satisfactory in each case we developp here a different approach conditionally on the total number of successes.

3.1 Odds Ratio

Two binary random variables $X_i \sim B(n_i, p_i)$, $i = 1, 2$, return x_i successes and $n_i - x_i$ failures. Suppose that $m = x_1 + x_2$, n_1 and $N = n_1 + n_2$ are known then, given x_1, all the values of the 2×2 table can be found. Given the odds ratio γ, estimated by $\hat{\gamma} = \frac{x_1(n_2 - x_2)}{x_2(n_1 - x_1)}$, x_1 follows the hypergeometric distribution

$$f(x_1; m, n_1, N, \gamma) = \frac{\binom{n_1}{x_1}\binom{N-n_1}{m-x_1}\gamma^{x_1}}{\sum_u \binom{n_1}{u}\binom{N-n_1}{m-u}\gamma^u},$$

where the sum goes from $u = \max\{0, m - n_2\}$ up to $\min\{m, n_1\}$.

We can use this distribution in the simulations instead of the binomial. The given parameters m, N, $R = \frac{n_1}{n_2}$ and γ allow to find both n_1, n_2 and then x_1 is generated with the hypergeometric distribution and $x_2 = m - x_1$. Once we get these values we use the same procedures as in Section 2.5 to compute confidence intervals and determine the coverage of them. In this method $A = \frac{R}{R+1}$ is fixed thus the nuisance parameter is now completely determined as m is given by the relation $m = N\psi$. The major advantage to remove the nuisance parameter is that the confidence interval is not estimated anymore, using Eq.(1.4) the parameters v and w are known.

Figures 3.1 to 3.4 show the results of the simulations for 1000 replications. The mean coverage and the average length of the confidence intervals are computed and the results are presented for all values of $m \in \{1, \ldots, N - 1\}$. In this case the x-axis of the plot displays the values of the average

probability $\bar{p}$ which is found as in Section 2.5 with $\psi = \frac{m}{N}$.

The coverage of these confidence intervals has more variation than with the unconditional method but the level of coverage is satisfying for all the four applied methods and it would be difficult to say which of them gives the best results even if the score method would probably be chosen because of the shorter length both for small and large $\bar{p}$. Nevertheless the *vst* method performs quite well in all cases now and is much faster to compute than the score method (about 100 times faster). Whereas the coverage was unacceptable with the unconditional intervals it gives now a good coverage even for $\gamma = 3$ (Figures 3.3 and 3.4) for both balanced and unbalanced cases. It is a little bit conservative for small and large values of $\bar{p}$ but the other methods do not perform better in this case. Moreover the length of the intervals remains quite small except for rare cases but this is perhaps only due to the use of the mean which is not robust at all.

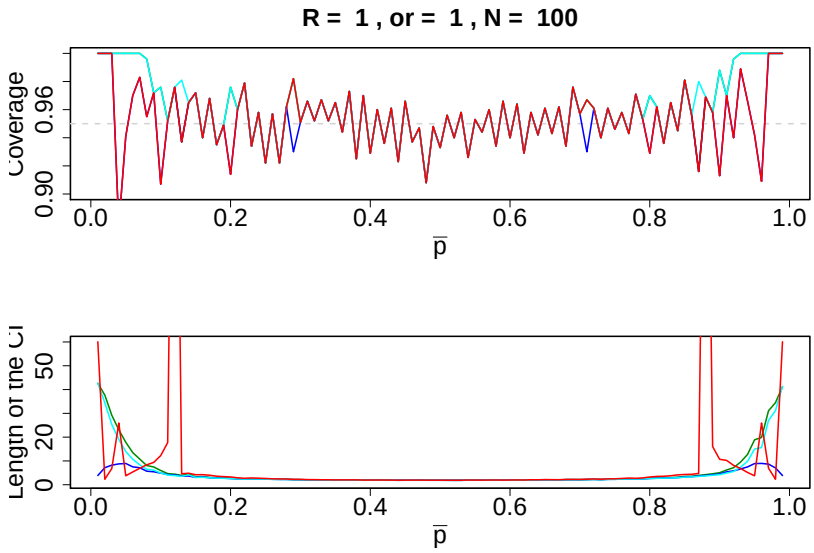

Figure 3.1: Coverage probabilities (top) and average lengths (bottom) of the nominal 95% confidence intervals based on the conditional distribution with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the score method (blue line). The sample sizes are balanced with $n_1 = n_2 = 50$ and the odds ratio is $\gamma = 1$.

3.2 Risk Ratio

The approach in the previous section uses the odds ratio to generate the data with the hypergeometric distribution. If a confidence interval for the relative risk is required we need to transform it into the odds ratio. This can be achieved writing p_1 and p_2 as functions of ρ and ψ as in Section 2.1

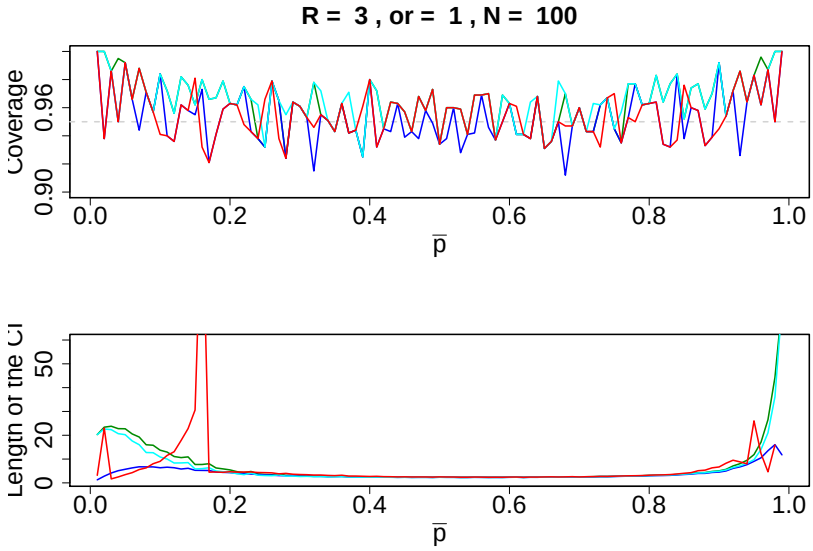

Figure 3.2: Coverage probabilities (top) and average lengths (bottom) of the nominal 95% confidence intervals based on the conditional distribution with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the score method (blue line). The sample sizes are unbalanced with $n_1 = 75$, $n_2 = 25$ and the odds ratio is $\gamma = 1$.

leading to

$$\gamma = \rho \frac{A(\rho - 1) + 1 - \psi}{A(\rho - 1) + 1 - \psi\rho}.$$

This transformation is not defined for $\rho = \frac{1-A}{\psi-A}$ but the constraint between ψ and ρ illustrated in Figure 1.1 avoids this case happening.

It seems that the transformation from ρ to γ might cause bias and thus we can not use it straightforward. Due to lack of time we did not investigate this point further. Nevertheless it would be of interest to be able to use the conditional method for the relative risk too.

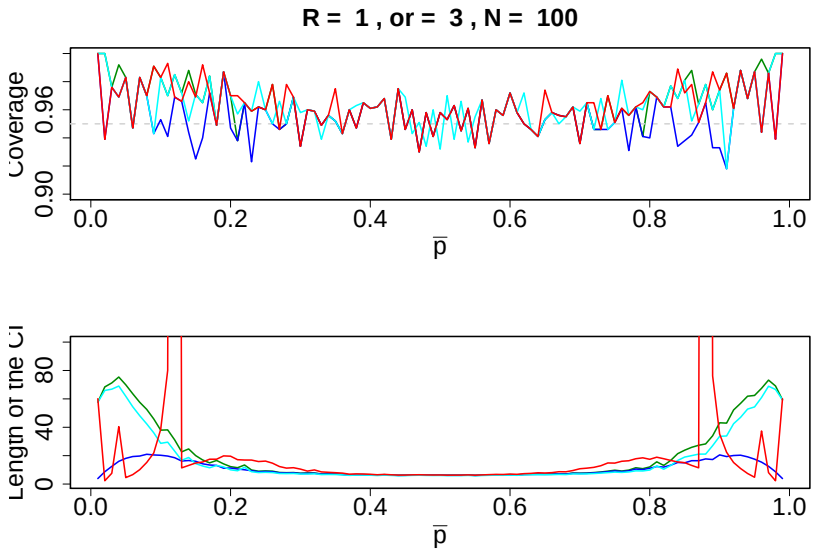

Figure 3.3: Coverage probabilities (top) and average lengths (bottom) of the nominal 95% confidence intervals based on the conditional distribution with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the score method (blue line). The sample sizes are balanced with $n_1 = n_2 = 50$ and the odds ratio is $\gamma = 3$.

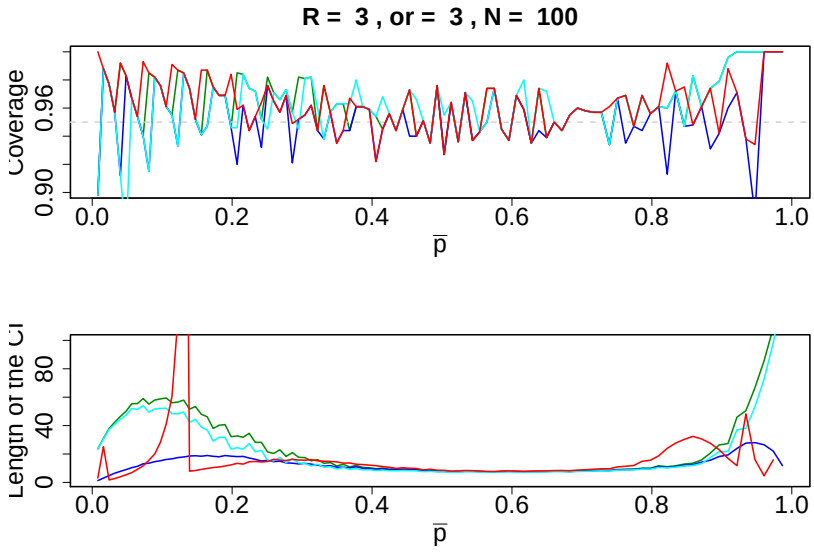

Figure 3.4: Coverage probabilities (top) and average lengths (bottom) of the nominal 95% confidence intervals based on the conditional distribution with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the score method (blue line). The sample sizes are unbalanced with $n_1 = 75$, $n_2 = 25$ and the odds ratio is $\gamma = 3$.

Chapter 4

Meta-analysis

4.1 Combining the Risk Difference

To find a combined effect using the *vst*, a weighted average is performed. Since the variance stabilized parameters $T_A^{\Delta}(\hat{\Delta}_k; \psi_k, \Delta_0), k = 1, \ldots, K$, follow a normal distribution with unit variance, they can be averaged with known weights $\sqrt{N_k}$ giving the following combined effect:

$$T_{\text{comb}}(\Delta_0) = \frac{\sum_{k=1}^{K} \sqrt{N_k} T_A^{\Delta}(\hat{\Delta}_k; \psi_k, \Delta_0)}{\sqrt{\sum_{k=1}^{K} N_k}} \tag{4.1}$$

(Kulinskaya et al. 2009, Eq. 2.4). The combined effect in Eq.(4.1) still has unit variance. A confidence interval for T_{comb} is the set of all Δ_0 such that $|T_{\text{comb}}(\Delta_0)| \leq z_{1-\frac{\alpha}{2}}$. As T_{comb} is monotone decreasing in Δ_0 any root-finding algorithm might be used to find the limits of the confidence interval.

If we want a confidence interval for the relative risk or the odds ratio exactly the same method can be applied. We write the quantity of interest as a function of Δ and ψ and replace it in Eq.(4.1).

4.2 Simulations Design

The coverage of the confidence intervals is computed by simulation. We compare our method with three other well-known approaches. These methods are used in `RevMan`[1], software of the Cochrane Collaboration, and described by Deeks and Higgins (2007). Two of them are based on the inverse variance (IV) weights, namely the methods of Woolf (1955) and Gart (1966). Here θ

[1] Review Manager (RevMan). [Computer Program]. Version 5.0. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2008.
See `http://www.cc-ims.net/revman` (visited in January 2010).

denotes the log odds ratio. The combined estimator is

$$\hat{\theta}_{\mathrm{IV}} = \frac{\sum_{k=1}^{K} w_k \hat{\theta}_k}{\sum_{k=1}^{K} w_k}$$

with the weights equal to $w_k = \mathrm{var}[\hat{\theta}_k]^{-1}$ and where the estimated variance is given in Eq.(2.2). The sampling variance of the pooled estimate is

$$\mathrm{var}[\hat{\theta}_{\mathrm{IV}}] = \left(\sum_{k=1}^{K} w_k \right)^{-1}.$$

A confidence interval is thus easily computed. The difference between the two methods has been explained in Section 2.1.

The Mantel and Haenszel (1959) method (MH) which is described by Agresti and Hartzel (2000) computes the weights in another way, that is $w_k = \frac{(n_{1k} - x_{1k})x_{2k}}{N_k}$. The pooled odds ratio is in this case

$$\gamma_{\mathrm{MH}} = \frac{\sum_{k=1}^{K} w_k \hat{\gamma}_k}{\sum_{k=1}^{K} w_k}$$

and the sampling variance of its logarithm is

$$\mathrm{var}[\log(\gamma_{\mathrm{MH}})] = \frac{1}{2} \left(\frac{E}{R^2} + \frac{F+G}{RS} + \frac{H}{S^2} \right),$$

where

$$\begin{aligned}
R &= \sum_{k=1}^{K} \frac{x_{1k}(n_{2k} - x_{2k})}{N_k}, \\
S &= \sum_{k=1}^{K} \frac{x_{2k}(n_{1k} - x_{1k})}{N_k}, \\
E &= \sum_{k=1}^{K} \frac{(x_{1k} + n_{2k} - x_{2k})x_{1k}(n_{2k} - x_{2k})}{N_k^2}, \\
F &= \sum_{k=1}^{K} \frac{(x_{1k} + n_{2k} - x_{2k})x_{2k}(n_{1k} - x_{1k})}{N_k^2}, \\
G &= \sum_{k=1}^{K} \frac{(x_{2k} + n_{1k} - x_{1k})x_{1k}(n_{2k} - x_{2k})}{N_k^2}, \\
H &= \sum_{k=1}^{K} \frac{(x_{2k} + n_{1k} - x_{1k})x_{2k}(n_{1k} - x_{1k})}{N_k^2}.
\end{aligned}$$

We can then compute a confidence interval for the log odds ratio and transform it into an interval for the odds ratio.

We follow the simulation's design described by Sánchez-Meca and Marín-Martínez (2000) which is based on 30 real meta-analyses in the field of health and behavioural sciences. This same setup has also been used by Kulinskaya (2009). The total number of studies takes the values $K = 10$, 20 and 40. The total sample sizes are selected as follows (Sánchez-Meca and Marín-Martínez 2000):

- the set $\{24, 24, 32, 32, 36, 36, 40, 40, 168, 168\}$ with an average size $\bar{N} = 60$,
- the set $\{64, 64, 72, 72, 76, 76, 80, 80, 208, 208\}$ with $\bar{N} = 100$
- and the set $\{124, 124, 132, 132, 136, 136, 140, 140, 268, 268\}$ with $\bar{N} = 160$.

For $K = 20$ these values are repeated twice and for $K = 40$ four times. The allocation ratio is $R \in \{1, 2, 3\}$ when simulating under the null hypothesis $\gamma = 1$ and $R \in \{1, 1/2, 2, 1/3, 3\}$ under the alternatives $\gamma = 1.3$ and $\gamma = 1.7$. For each of the 45 configurations of the above parameters 1000 simulations are run and the average coverage, level and length of the confidence intervals are computed for eight different values of ψ, $\psi \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.5, 0.85, 0.95\}$. In simulations with $\gamma = 1$ the proportion of rejection of the null hypothesis is the estimated Type I error rate whereas with $\gamma \in \{1.3, 1.7\}$ this proportion represents the estimated power.

4.3 Results under the Null Hypothesis

This section presents the results of simulations under the null hypothesis $\gamma = 1$. Figures 4.1 to 4.3 show the mean coverage of the confidence intervals with the four compared approaches for different values of the average sample size and allocation ratio and for a different number of studies.

Figure 4.1 shows the results for $K = 10$ studies. For the balanced case (left column) all the four methods perform quite well but we would prefer MH (black line) because the others are a bit conservative, especially for small and large values of ψ. When $R = 2$ all methods perform really similar but when the allocation ratio is $R = 3$ the Gart method (cyan line) becomes quite liberal and the increase of the sample size does not improve its coverage. Overall the *vst* method (red line) performs very well even if it is a bit conservative in the edges.

When the number of studies is $K = 20$ (Figure 4.2) the methods are still conservative in the balanced case except MH whose coverage remains always pretty close to the 95% nominal level. In the unbalanced cases the

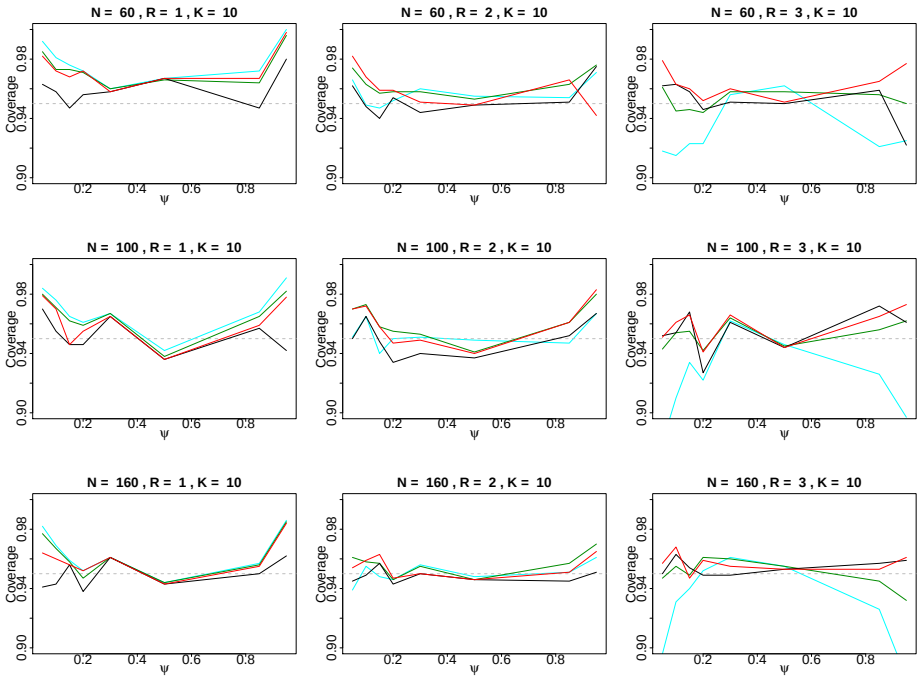

Figure 4.1: Mean coverage of the nominal 95% confidence intervals determined with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the Mantel–Haenszel method (black line). The horizontal dashed line represents the 95% level. The number of studies is $K = 10$, the allocation ratio takes values $R = 1$ (left column), $R = 2$ (central column) and $R = 3$ (right column) and the average sample size is $\bar{N} = 60$ (top row), $\bar{N} = 100$ (central row) and $\bar{N} = 160$ (bottom row).

two inverse variance approaches (cyan and green lines) become much too liberal especially for $R = 3$ (right column). The results are similar even for large sample sizes whereas both *vst* and Mantel–Haenszel methods perform well. The only case where MH drops is when $\psi = 0.95$ (top right panel) and this is also observed when $K = 10$ and $K = 40$. Otherwise there is no big difference between these two methods.

With $K = 40$ studies (Figure 4.3) we see that the inverse variance methods should really be avoided for unbalanced sample sizes. Their coverages fall down under 90% for the majority of values of ψ which is worst than with 20 studies. In this case the results are unacceptable even when $R = 2$. The *vst* method seems to perform at least as well as the Mantel–Haenszel's except in the balanced case where the last method is still prefered, the *vst* being more conservative. In the unbalanced cases these two approaches behave quite similar but the *vst* would be prefered when $R = 3$ and $\bar{N} = 60$ (top right panel) because MH drops under 90% for $\psi = 0.05$ and 0.95 whereas the *vst* always has a coverage above 93%.

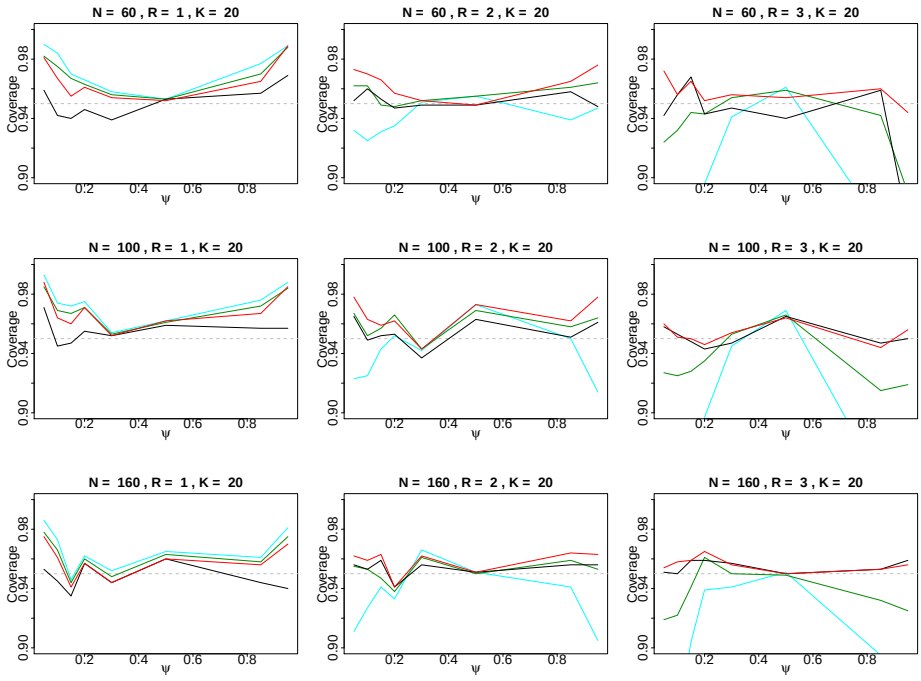

Figure 4.2: Mean coverage of the nominal 95% confidence intervals determined with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the Mantel–Haenszel method (black line). The horizontal dashed line represents the 95% level. The number of studies is $K = 20$, the allocation ratio takes values $R = 1$ (left column), $R = 2$ (central column) and $R = 3$ (right column) and the average sample size is $\bar{N} = 60$ (top row), $\bar{N} = 100$ (central row) and $\bar{N} = 160$ (bottom row).

All these observed results tally with the ones of Kulinskaya (2009) for the risk difference. Moreover the simulations return the average length of the confidence intervals. The results are not shown but there is no significative difference between the four methods, the intervals being a bit longer for small and large values of the risk ψ and smaller for ψ around 0.5 but the lengths cannot be used to decide which method is preferable.

4.4 Results under the Alternatives

In this section the simulations are first performed with a true value of the odds ratio $\gamma = 1.3$ with the same sample sizes and number of studies as in the previous section. This alternative value is chosen to have a reasonable power of the test. The power is proportional to both sample size $\bar{N}$ and number of studies K. For fixed $\bar{N}$ and K the power is usually maximum for ψ around 0.5. The allocation ratio R has little effect on the power. For

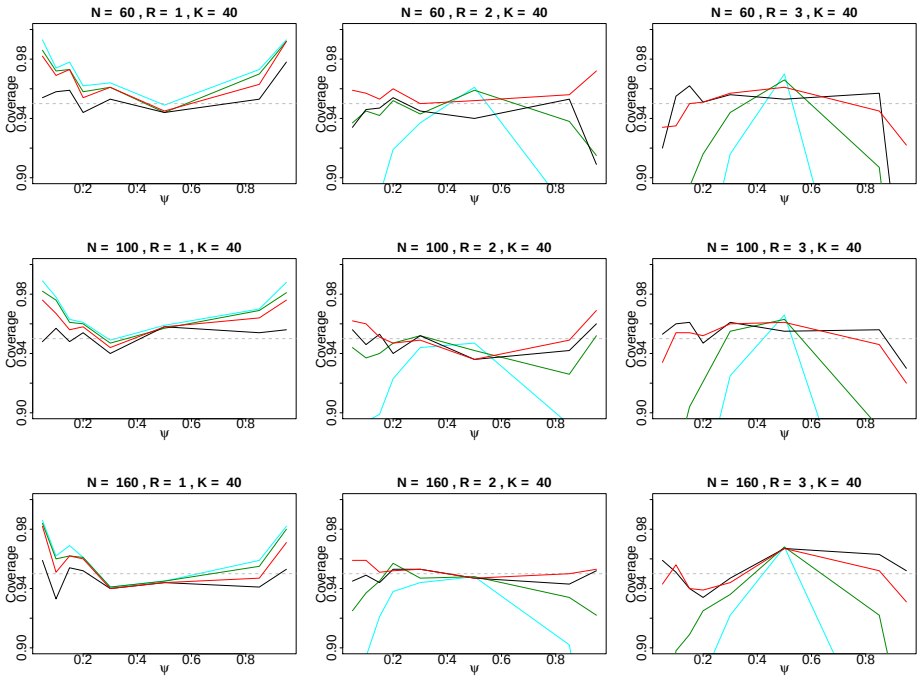

Figure 4.3: Mean coverage of the nominal 95% confidence intervals determined with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the Mantel–Haenszel method (black line). The horizontal dashed line represents the 95% level. The number of studies is $K = 40$, the allocation ratio takes values $R = 1$ (left column), $R = 2$ (central column) and $R = 3$ (right column) and the average sample size is $\bar{N} = 60$ (top row), $\bar{N} = 100$ (central row) and $\bar{N} = 160$ (bottom row).

example with $\bar{N} = 60$ and $K = 10$ the power remains under 40% but grows up to 80% with $\bar{N} = 160$. The increase is about 20% with $K = 20$ and another 20% when $K = 40$ such that the power is always close to 100% in this last situation.

Figures 4.4 to 4.6 show the coverage of the confidence intervals for $K = 10$, 20 and 40 studies respectively. As in the previous section with $\gamma = 1$ the results become worst when increasing the number of studies and the inverse variance methods (cyan and green lines) are far unsatisfying in unbalanced cases and should be avoided. Moreover in the balanced case the Mantel–Haenszel method (black line) is the best one, the others being too conservative.

When $K = 10$ (Figure 4.4) the *vst* (red line) and MH perform quite similarly even if the *vst* is a bit more conservative. For large sample sizes (bottom panels) these two approaches have a very satisfying coverage for all values of ψ.

With 20 studies (Figure 4.5) both Woolf and Gart methods become worst

in every unbalanced case but the *vst* and Mantel–Haenszel still perform quite well in most of the cases. The only bad point for MH is with $\bar{N} = 60$, $R = \frac{1}{3}$ and $\psi = 0.95$ (fourth panel on the top row) where the coverage drops under 90% and when $R = \frac{1}{2}$ the intervals are a bit liberal for large values of ψ. As before the *vst* and Mantel–Haenszel behave quite similarly especially for $\bar{N} = 160$ and it would be difficult to decide for the best one.

The results deteriorate when the number of studies is $K = 40$ (Figure 4.6) in the unbalanced cases. For $R = 1$ the coverage is as described previously, satisfying but a bit conservative except with Mantel–Haenszel. All methods deteriorate in the unbalanced cases especially with $R = 3$ and $R = \frac{1}{3}$. As usual the inverse variance methods are the worst ones. For large sample sizes (central and bottom panels) the coverage with the *vst* and MH are good except for the *vst* with extreme values of ψ. Basically the two approaches do not behave that differently for all ψ between 0.2 and 0.8. When $\bar{N} = 60$ (top panels) the coverage is not very satisfying for all methods especially for $R = \frac{1}{3}$ where the coverage drops even for values of ψ close to 0.5. Since no method is satisfactory enough in these situations we have tried to run simulations with $\bar{N} = 320$ (bottom row) to check if a larger sample size improves the coverage. This sample is defined as the sum of the three other samples with $\bar{N} = 60$, 100 and 160. The two inverse variance methods still do not behave correctly in the unbalanced cases but both *vst* and MH are now better than previously; the intervals everywhere have a coverage above 92% and the two methods are close from each other except for very few values where MH would be prefered (see for example the leftmost point in the bottom right panel).

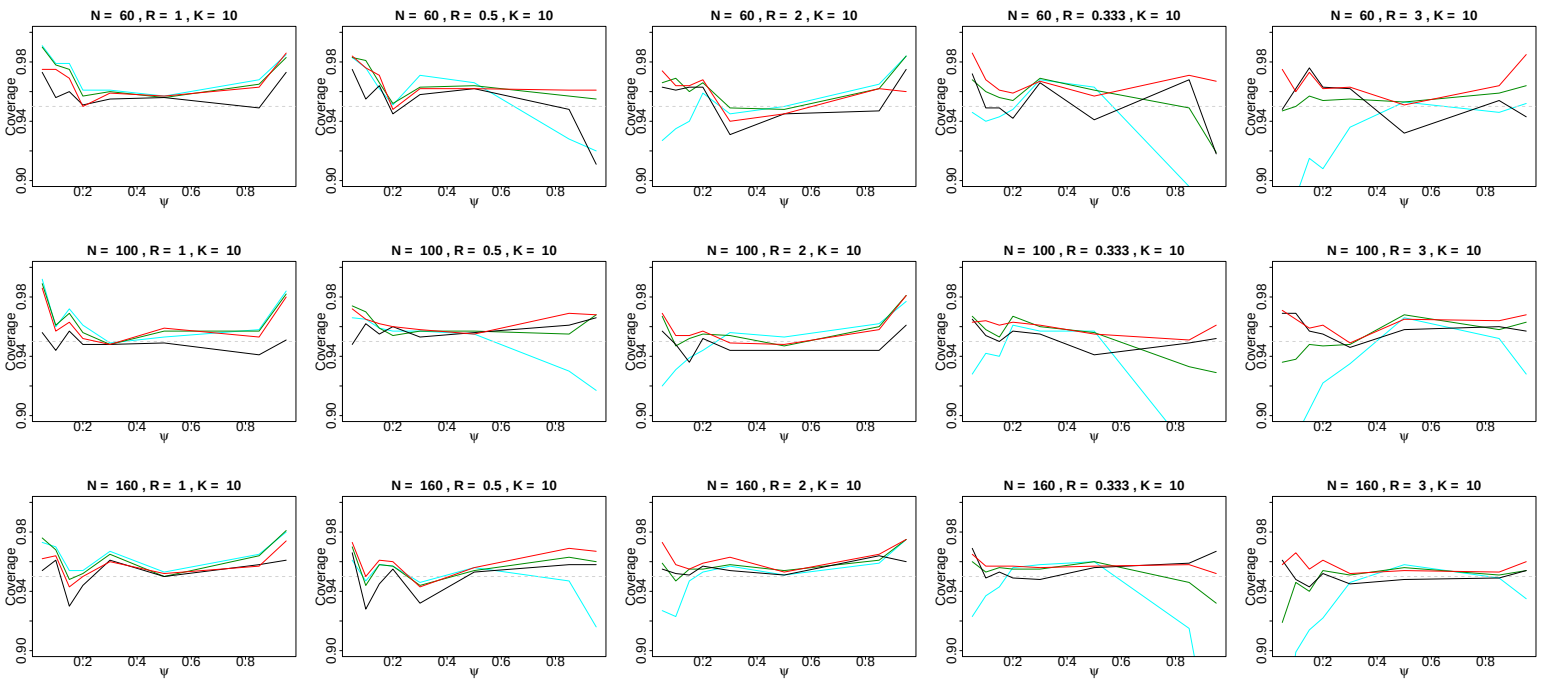

Figure 4.4: Mean coverage of the nominal 95% confidence intervals determined with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the Mantel–Haenszel method (black line). The horizontal dashed line represents the 95% level. The number of studies is $K = 10$, the allocation ratio takes values $R = 1$, $\frac{1}{2}$, 2, $\frac{1}{3}$ and 3 (from left to right columns) and the average sample size is $\bar{N} = 60$ (top row), $\bar{N} = 100$ (central row) and $\bar{N} = 160$ (bottom row). The true odds ratio is $\gamma = 1.3$.

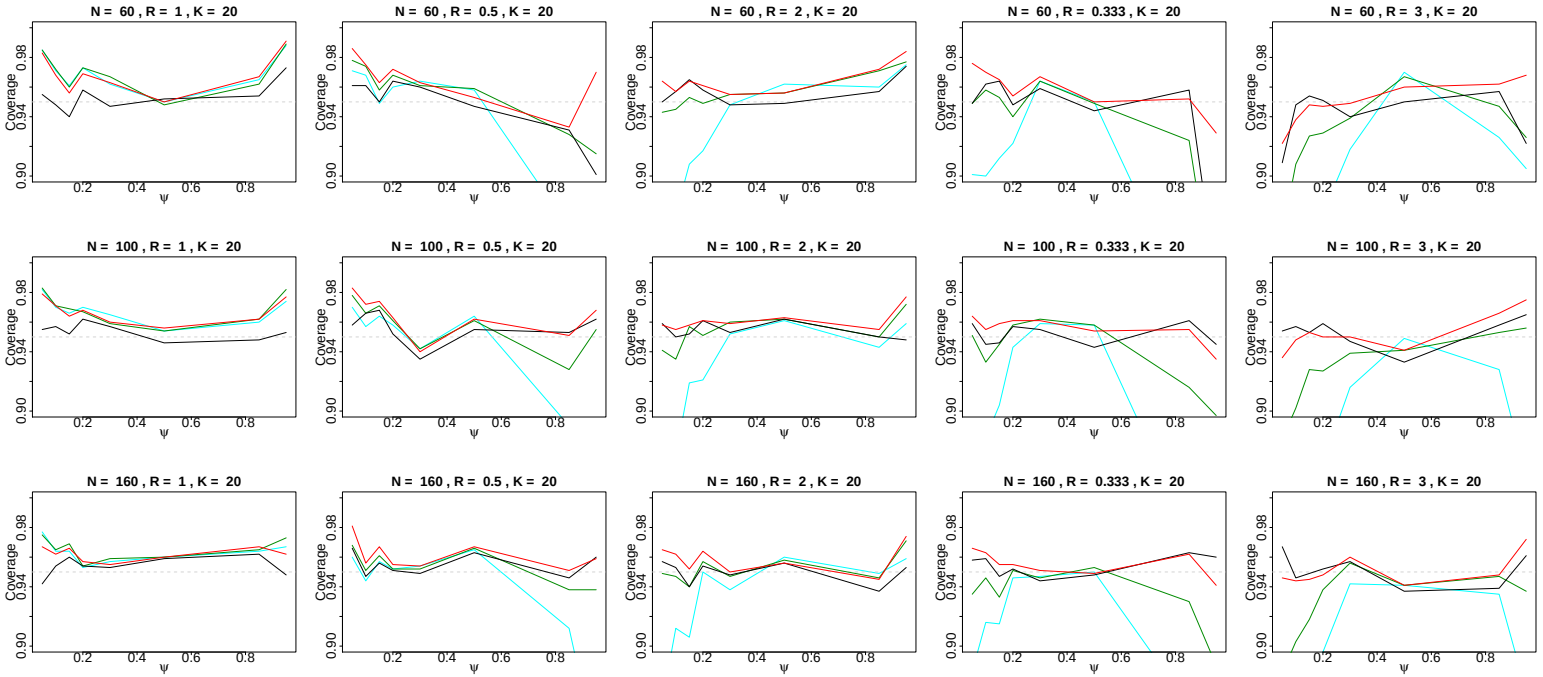

Figure 4.5: Mean coverage of the nominal 95% confidence intervals determined with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the Mantel–Haenszel method (black line). The horizontal dashed line represents the 95% level. The number of studies is $K = 20$, the allocation ratio takes values $R = 1$, $\frac{1}{2}$, 2, $\frac{1}{3}$ and 3 (from left to right columns) and the average sample size is $\bar{N} = 60$ (top row), $\bar{N} = 100$ (central row) and $\bar{N} = 160$ (bottom row). The true odds ratio is $\gamma = 1.3$.

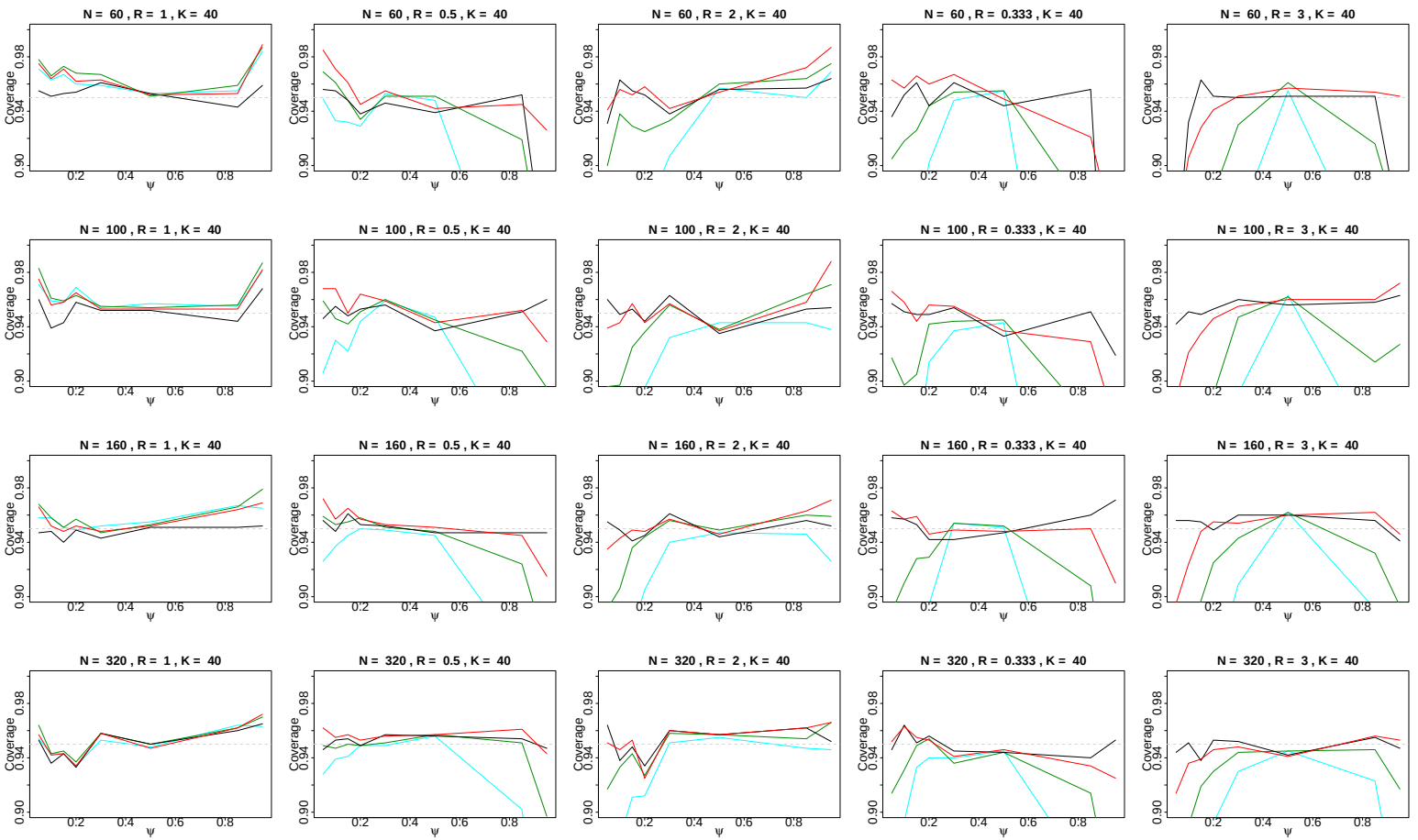

Figure 4.6: Mean coverage of the nominal 95% confidence intervals determined with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the Mantel–Haenszel method (black line). The horizontal dashed line represents the 95% level. The number of studies is $K = 40$, the allocation ratio takes values $R = 1$, $\frac{1}{2}$, 2, $\frac{1}{3}$ and 3 (from left to right columns) and the average sample size is $\bar{N} = 60$ (top row), $\bar{N} = 100$ (second row), $\bar{N} = 160$ (third row) and $\bar{N} = 320$ (bottom row). The true odds ratio is $\gamma = 1.3$.

We also run simulations with a much larger value of the odds ratio $\gamma = 1.7$ to verify if the coverage remains correct for more extreme cases. Figures 4.7 to 4.9 show the results of simulations for the same configurations of parameters as above with $\gamma = 1.3$. As before we notice that the results for unbalanced cases are symmetric. For example the coverage for a small ψ when $R = 2$ is similar to the coverage for a large ψ and $R = \frac{1}{2}$.

For $K = 10$ studies (Figure 4.7) all four approaches are too conservative in the balanced cases (left panels) whatever the value of ψ is. In unbalanced cases the inverse variance methods are far too liberal. The Gart intervals (cyan lines) fail to reach the nominal level even for $R = 2$ or $R = \frac{1}{2}$ and both Gart and Woolf (green lines) have a poor behaviour when $R = 3$ or $R = \frac{1}{3}$. The two other methods, namely the *vst* and MH (red and black lines respectively) perform relatively well in all cases but we notice that their coverage is not as satisfying as when $\gamma = 1.3$ (Figure 4.4).

When $K = 20$ (Figure 4.8) all four methods behave well when $R = 1$ but the coverage of the two inverse variance methods drops in unbalanced cases. When $R = 2$ or $\frac{1}{2}$ the *vst* and MH still behave very well for large sample sizes. They are sometimes a little bit conservative but for $\bar{N} = 60$ (top panels) it happens that the coverage fails to reach the nominal 95% level. In the most unbalanced cases ($R = \frac{1}{3}$ and 3) the coverage of these two methods is satisfying for $\bar{N} = 160$ but for small sample sizes and extreme values of ψ the coverage sometimes drops under 90%. This happens for both *vst* and MH and there is basically no great difference between these two approaches.

Figure 4.9 shows that the results are globally unsatisfying for $K = 40$ studies. As usual the inverse variance methods are the worst ones but here Gart intervals (cyan lines) behave badly even in balanced cases for small or large values of ψ. In unbalanced cases both Gart and Woolf are much too liberal, Gart intervals never reach the 95% level when $R = 3$ (rightmost panels). The *vst* and MH perform pretty well when $R = 1$ but all four approaches behave badly in unbalanced cases with small sample sizes with a coverage under 90% for some values of ψ. For large $\bar{N}$ Mantel–Haenszel intervals have a very satisfying coverage even in very unbalanced cases but the *vst* is not as reliable. Even if it performs much better than the inverse variance methods its coverage sometimes fails to reach an acceptable level. However we see that the increase of the sample sizes really has a postive effect on the behaviour of the confidence intervals of any methods.

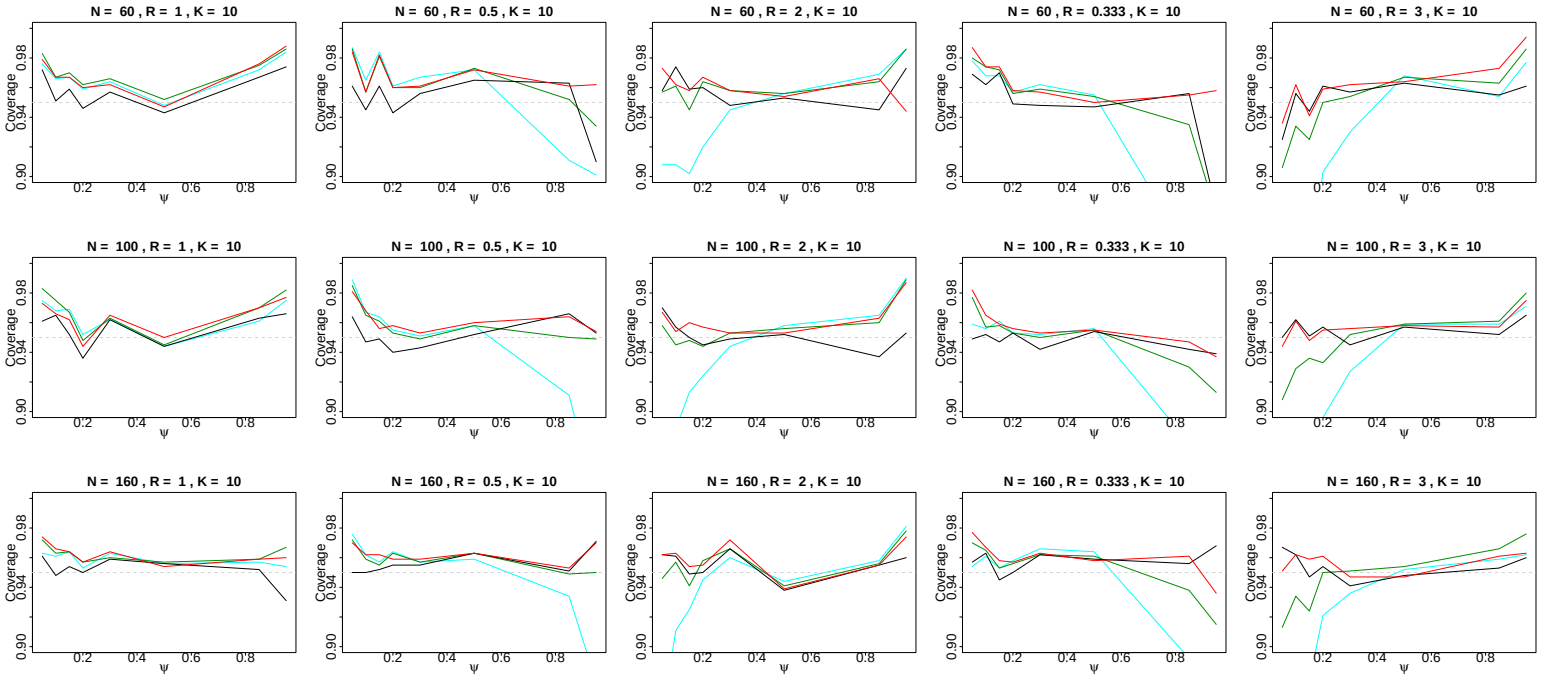

Figure 4.7: Mean coverage of the nominal 95% confidence intervals determined with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the Mantel–Haenszel method (black line). The horizontal dashed line represents the 95% level. The number of studies is $K = 10$, the allocation ratio takes values $R = 1$, $\frac{1}{2}$, 2, $\frac{1}{3}$ and 3 (from left to right columns) and the average sample size is $\bar{N} = 60$ (top row), $\bar{N} = 100$ (central row) and $\bar{N} = 160$ (bottom row). The true odds ratio is $\gamma = 1.7$.

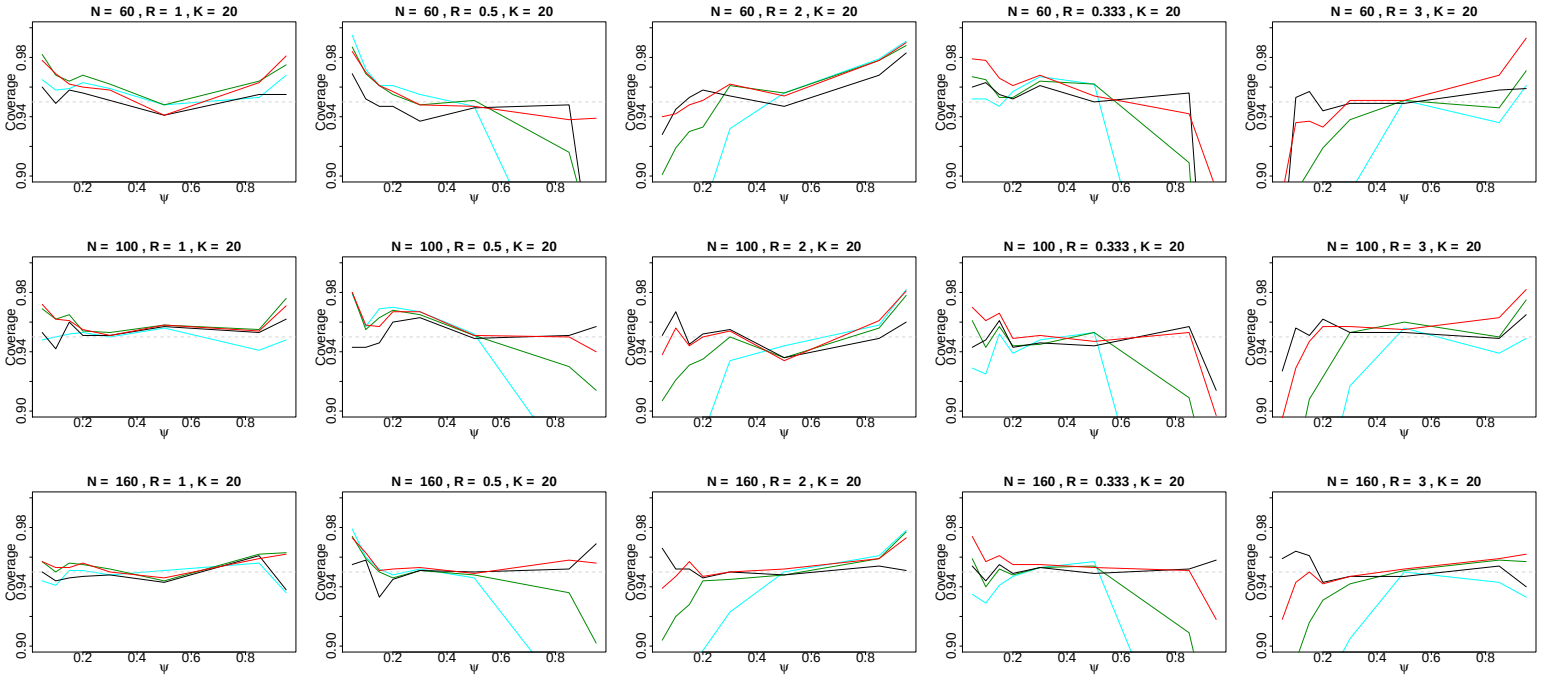

Figure 4.8: Mean coverage of the nominal 95% confidence intervals determined with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the Mantel–Haenszel method (black line). The horizontal dashed line represents the 95% level. The number of studies is $K = 20$, the allocation ratio takes values $R = 1$, $\frac{1}{2}$, 2, $\frac{1}{3}$ and 3 (from left to right columns) and the average sample size is $\bar{N} = 60$ (top row), $\bar{N} = 100$ (central row) and $\bar{N} = 160$ (bottom row). The true odds ratio is $\gamma = 1.7$.

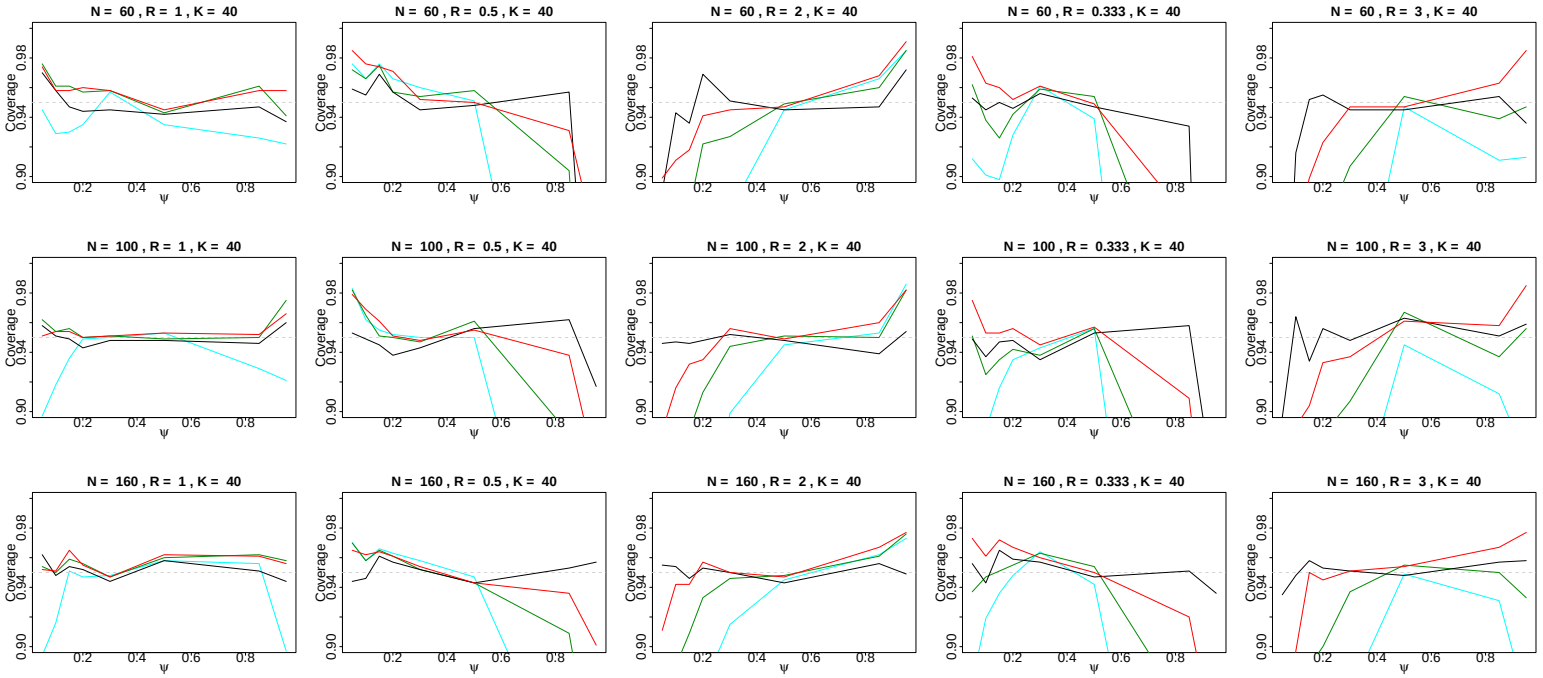

Figure 4.9: Mean coverage of the nominal 95% confidence intervals determined with the *vst* (red line), the Woolf and Gart methods (green and cyan lines respectively) and the Mantel–Haenszel method (black line). The horizontal dashed line represents the 95% level. The number of studies is $K = 40$, the allocation ratio takes values $R = 1$, $\frac{1}{2}$, 2, $\frac{1}{3}$ and 3 (from left to right columns) and the average sample size is $\bar{N} = 60$ (top row), $\bar{N} = 100$ (central row) and $\bar{N} = 160$ (bottom row). The true odds ratio is $\gamma = 1.7$.

4.5 Examples with Fixed Numbers of Successes

In the previous sections the value of the odds ratio was fixed and the simulations were run for different values of ψ. Here we fix the total number of successes from real datasets and simulate the confidence intervals for different values of the odds ratio. In what follows the data give the number of deaths among the total number of patients. Thus the probability of success is actually the probability to die.

4.5.1 Angiotensin-Converting Enzyme Data

The data used for this first example come from Garg and Yusuf (1995) and consist of a meta-analysis of 32 studies testing the effect of angiotensin-converting enzyme (ACE) inhibitors on mortality in patients with heart failure. Different kinds of agents are tested and there is a control group to see the efficiency of the treatment. For more details about the studies the reader is refered to Garg and Yusuf (1995). Figure 4.10 gives the estimates of the odds ratio and 95% confidence intervals for each study computed with the Woolf method in addition to the number of events and total number of subjects in both treated and control groups. The four studies with zero event in both arms are not displayed. The Mantel–Haenszel method returns a pooled odds ratio of 0.77 and 95% confidence limits of $[0.67, 0.88]$ which indicate a significant effect of the treatment. Table 4.1 gives the values of the pooled odds ratio and 95% confidence intervals for each of the four methods, namely the *vst*, Woolf, Gart and MH. We see that the results are very similar even if the *vst* returns slightly smaller values. We compute the vectors m and N of total successes and sample sizes respectively and run the simulations with this setup for different values of the odds ratio.

Most of the studies have a really small number of deaths in both treatment and control groups. Except four or five studies with more successes most of the estimated probabilities are much smaller than 10% in both treatment and control.

Results are presented in Figure 4.11 that shows the mean coverage, power and length of simulated confidence intervals for different values of the odds ratio $\gamma \in \{0.666, 0.8, 0.9, 0.95, 0.975, 1, 1.025, 1.05, 1.1, 1.2, 1.5\}$. All methods except the *vst* perform very similarly and quite well which is not surprising: MH is always satisfying and so do the inverse variance methods with such balanced studies. Nevertheless the *vst* is worst than the other methods. A too conservative coverage, small power and longer length of the intervals indicate that this method should not be used with these studies. Indeed with very small success probabilities the Mantel–Haenszel method would usually be recommended. We notice that the two inverse variance methods perform well too.

This is an example where the *vst* should not be used. In what follows

we run the same simulations for other datasets to assess the performance of the developed method.

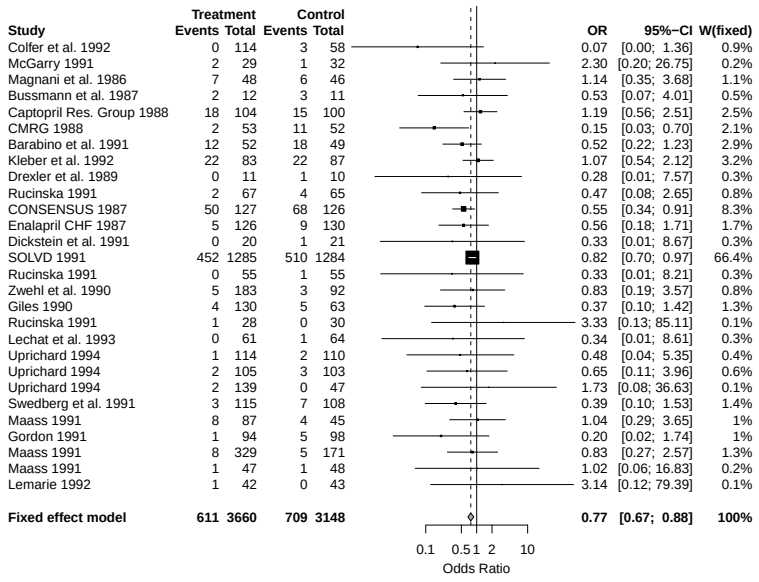

Figure 4.10: Studies testing the effect of ACE inhibitors with a forest plot representing the estimate of the odds ratio and the limits of a 95% confidence interval computed with the Woolf method. The rightmost column gives the weights used to compute the pooled odds ratio represented in the last row and determined with MH.

	γ_{pooled}	L	U
vst	0.745	0.641	0.864
Woolf	0.775	0.678	0.887
Gart	0.775	0.679	0.885
MH	0.767	0.671	0.875

Table 4.1: Pooled odds ratio (first column) and 95% confidence limits (lower bound and upper bound) computed with the *vst*, Woolf, Gart and MH methods.

4.5.2 Pre-eclampsia Data

The data presented in Figure 4.12 have been collected by Collins et al. (1985) and have also been later studied by Hardy and Thompson (1996) and Viechtbauer (2006). Nine studies are reviewed concerning the effectiveness of taking diuretics during pregnancy for preventing pre-eclampsia. These studies involve a total of about 7000 patients. The pooled odds ratio of 0.67

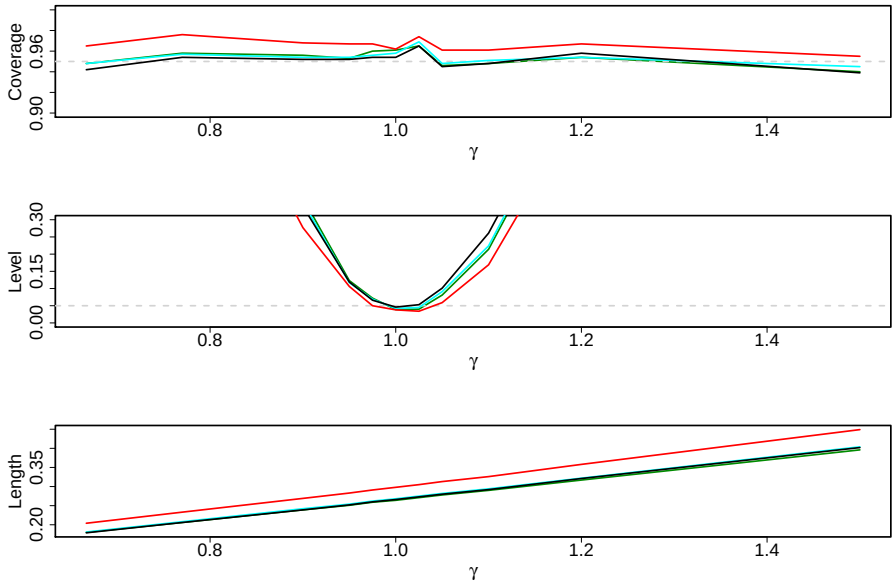

Figure 4.11: Coverage (top), power (central) and average length (bottom) of the confidence intervals based on the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively) and the Mantel–Haenszel method (black line).

with confidence limits of $[0.56, 0.79]$ computed with MH indicate a positive effect of the treatment. Table 4.2 gives the pooled odds ratio estimated with each of the four methods and the corresponding confidence limits. As in the previous example the *vst* estimate is a bit smaller than the others.

Results of the simulations are presented in Figure 4.13 showing the mean coverage, power and length of the confidence intervals. The coverage (top panel) is very similar for all four compared methods, staying very close to the nominal 95% for the whole range of values of the odds ratio. The power (middle panel) is also relatively the same for every method even if the *vst* is slightly lower and MH is a bit higher for larger odds ratios. As in the previous example the *vst* gives longer confidence intervals (bottom panel) than the three other methods.

In this example it would be difficult to say which method performs the best because they all are very similar but it seems that the *vst* is a bit worst and MH has a slightly better power.

4.5.3 Angiotensin Receptor Blockers Data

This data (explained in Figure 4.14) come from Jong et al. (2002) who identify 17 relevant studies to determine the effect of angiotensin receptor blockers (ARBs) on mortality in patients with heart failure. A total of 12469 patients participated to these studies: 7060 with ARBs and 5409 used

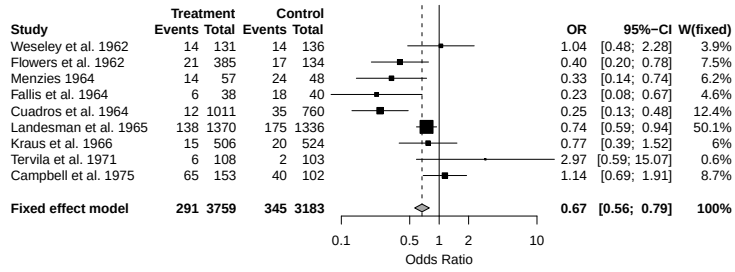

Figure 4.12: Studies testing the effect of diuretics for preventing pre-eclampsia with a forest plot representing the estimate of the odds ratio and the limits of a 95% confidence interval computed with the Woolf method. The rightmost column gives the weights used to compute the pooled odds ratio represented in the last row and determined with MH.

	γ_{pooled}	L	U
vst	0.631	0.522	0.759
Woolf	0.672	0.564	0.800
Gart	0.673	0.566	0.801
MH	0.668	0.562	0.793

Table 4.2: Pooled odds ratio (first column) and 95% confidence limits (lower bound and upper bound) computed with the *vst*, Woolf, Gart and MH methods.

placebo or ACE inhibitors as controls. Table 4.3 represents the pooled odds ratios and corresponding 95% confidence intervals obtained with the four compared methods. There is almost no difference between the methods in this example. The value of the odds ratio close to 1.03 with confidence limits $(0.93, 1.15)$ indicate no statistical difference in the mortality rate between the treated and control groups.

The probabilities of death are small in most of the studies (less than 10% in all studies but two) but the particularity of this meta-analysis is that most of the studies have unbalanced sample sizes. Some of them are balanced (ADEPT 2001, ELITE II 2000, Hamroff 1999) and others are strongly

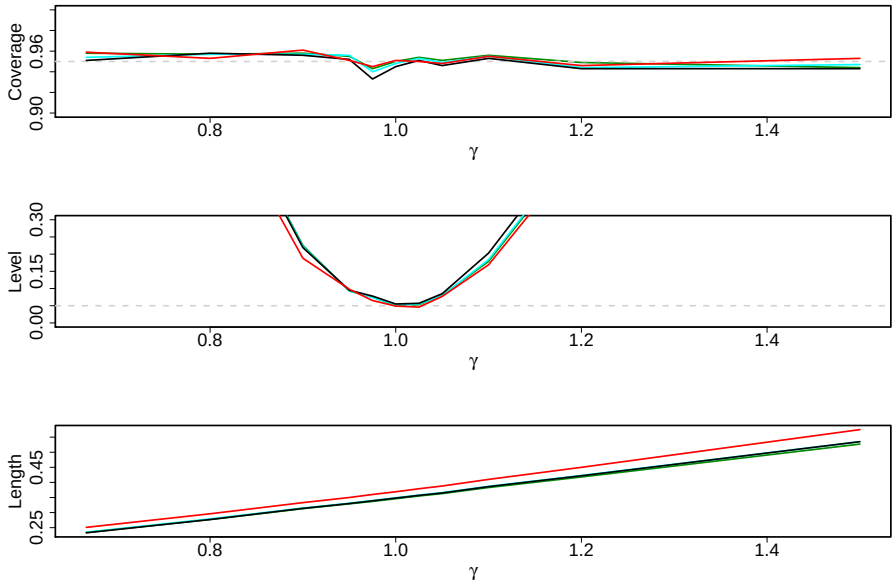

Figure 4.13: Coverage (top), power (central) and average length (bottom) of the confidence intervals based on the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively) and the Mantel–Haenszel method (black line).

unbalanced (Crozier 1995, RESOLVD 1999). However if we compute the allocation ratio of the total number of patients we get $R = \frac{7060}{5409} \approx 1.3$ which is still much less than $R = 2$ or $R = 3$ used for the simulations in the previous sections.

Figure 4.15 shows the results of the simulations. We notice that the coverage of the *vst* (red line) is a bit conservative for small values of the odds ratio (top panel) but is very satisfying for all values larger than 0.9 whereas the two inverse variance methods become too liberal when the odds ratio increases, especially the Woolf method (green line). The power of all four methods is really close but MH is still a bit higher for large values of the odds ratio. As in the previous examples the length of the *vst* intervals is larger than the others but the difference tends to be smaller.

In this particular example the Woolf method should be avoided and the *vst* performs better than previously. Nevertheless MH remains the best method as in all other considered examples. Even if the studies are unbalanced the inverse variance methods do not perform so bad because the sample sizes are quite large (larger than in the two previous examples) and the studies are not as unbalanced as in the simulations from the previous sections. Moreover the pooled odds ratio is 0.96 corresponding to the simulations under the null hypothesis of Section 4.3. The average sample size is here $\bar{N} = 733$, much more than the values previously tested in the simu-

lations. This explains why there is much less variation is these results than in Figure 4.2.

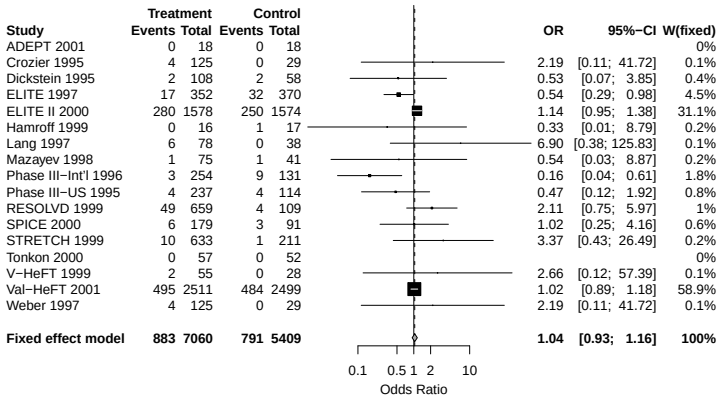

Figure 4.14: Studies testing the effect of ARBs with a forest plot representing the estimate of the odds ratio and the limits of a 95% confidence interval computed with the Woolf method. The rightmost column gives the weights used to compute the pooled odds ratio represented in the last row and determined with MH.

	γ_{pooled}	L	U
vst	1.034	0.924	1.159
Woolf	1.035	0.930	1.153
Gart	1.032	0.927	1.149
MH	1.039	0.934	1.156

Table 4.3: Pooled odds ratio (first column) and 95% confidence limits (lower bound and upper bound) computed with the *vst*, Woolf, Gart and MH methods.

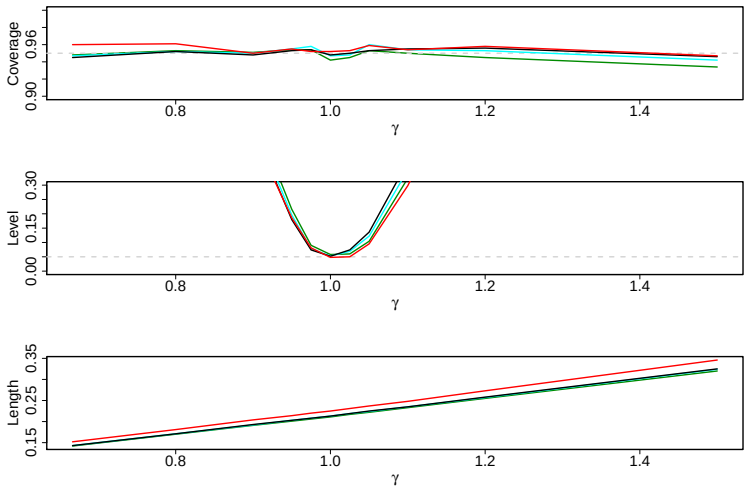

Figure 4.15: Coverage (top), power (central) and average length (bottom) of the confidence intervals based on the *vst* (red line), the Gart and Woolf methods (cyan and green lines respectively) and the Mantel–Haenszel method (black line).

Conclusion

Kulinskaya et al. (2009) developed a new approach for computing confidence intervals for the difference of two binomial probabilities using a variance stabilization. Results of their simulations show very good performances: their method is even better than the highly regarded Newcombe procedure. It was then of interest to test if the same kind of approach could be adapted for the relative risk and the odds ratio. We defined transformations to compute confidence intervals for both risk ratio and odds ratio. However these transformations induce some bias so that the coverage of the obtained intervals fails to reach an acceptable level. This bias is due to the nuisance parameter that needs to be estimated. It performs well only under the null hypothesis of no difference between the two groups. Under alternatives the results for the odds ratio are better than those for the relative risk but they remain unacceptable especially in unbalanced cases. Several corrections have been implanted but none of them have succeeded in improving the performance of the confidence intervals.

To obtain better results we have developed a new approach based on a conditional distribution with fixed number of successes to remove the nuisance parameter. Simulations show that these confidence intervals now perform very well for all configurations of the parameters. The coverage of the obtained confidence intervals is as good as the applied score method recommended by Agresti and Min (2005). Based on this satisfying method we have then combined the results of many studies to get a confidence interval for the global result of a meta-analysis. The simulations under the null hypothesis of an odds ratio equal to one show that even if the *vst* method is sometimes a bit conservative it performs quite well and never has a too low level of coverage. This method is much better than the widely applied inverse variance intervals and performs as well as Mantel–Haenszel's which is one of the best known approach. Under alternatives the *vst* still performs quite well especially for a small number of studies. For more studies results are not as satisfying but Mantel–Haenszel does not perform much better and both are far more reliable than the inverse variance approaches. For a larger odds ratio the performances decrease for all methods and larger sample sizes are needed to achieve a satisfying level of coverage.

In further researches it would be of interest to investigate several types

of correction in the transformation from the risk difference to either the relative risk or the odds ratio. Appropriate corrections would allow to use an unconditional distribution in the simulations of the confidence intervals. Another point that needs to be developed is to compute conditional intervals for the relative risk too. It seems that the transformation from relative risk to odds ratio introduces bias so corrections would perhaps be needed.

In summary the use of a variance stabilization is a recent but very promising way of computing confidence intervals either to compare two binomial proportions or to combine evidences in a meta-analysis. Its behaviour performs much better than widely used inverse variance methods and is often as reliable as the actual best known methods. Moreover the *vst* intervals can be easily computed and need a much shorter computation time than for example the score method. Further work is needed to really highlight the quality of this approach.

References

Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples. *Biometrics* ***55***, 597–602.

Agresti, A. and J. Hartzel (2000). Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine* ***19***, 1115–1139.

Agresti, A. and Y. Min (2002). Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics* ***3***, 379–386.

Agresti, A. and Y. Min (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2×2 contingency tables. *Biometrics* ***61***, 515–523.

Anscombe, F. J. (1956). On estimating binomial response relations. *Biometrika* ***43***, 461–464.

Brown, C. C. (1981). The validity of approximation methods for interval estimation of the odds ratio. *American Journal of Epidemiology* ***113***, 474–480.

Brown, L. and X. Li (2005). Confidence intervals for two sample binomial distribution. *Journal of Statistical Planning and Inference* ***130***, 359–375.

Collins, R., S. Yusuf, and R. Peto (1985). Overview of randomised trials of diuretics in pregnancy. *British Medical Journal* ***290***, 17–23.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* ***4*** (J. Neyman (ed)), 135–148.

Deeks, J. J. and J. P. T. Higgins (2007). *Statistical algorithms in Review Manager 5.*

Garg, R. and S. Yusuf (1995). Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure. *Journal of the American Medical Association* ***273***, 1450–1456.

Gart, J. J. (1966). Alternative analyses of contingency tables. *Journal of the Royal Statistical Society* ***28*** (Series B), 164–179.

Hardy, R. J. and S. G. Thompson (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* ***15***, 619–629.

Jong, P., C. Demers, R. S. McKelvie, and P. P. Liu (2002). Angiotensin receptor blockers in heart failure: meta-analysis of randomized controlled trials. *Journal of the American College of Cardiology* ***39***, 463–470.

Koopman, P. A. R. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics* ***40***, 513–517.

Kulinskaya, E. (2009). Variance stabilisation in meta-analysis: combining the evidence. *57th Session of the International Statistical Institute, Durban, 14–22 August 2009* (paper 733).

Kulinskaya, E., S. Morgenthaler, and R. G. Staudte (2009). Variance stabilizing the risk difference. *submitted* .

Mantel, N. and W. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* ***22***, 719–748.

Martín Andrés, A. and J. M. Tapia García (2004). Optimal unconditional asymptotic test in 2×2 multinomial trials. *Communication in Statistics – Simulation and Computation* ***33***, 83–97.

Miettinen, O. and M. Nurminen (1985). Comparative analysis of two rates. *Statistics in Medicine* ***4***, 213–226.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* ***17***, 873–890.

Sánchez-Meca, J. and F. Marín-Martínez (2000). Testing the significance of a common risk difference in meta-analysis. *Computational Statistics & Data Analysis* ***33***, 299–313.

Storer, B. E. and C. Kim (1990). Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association* ***85***, 146–155.

Viechtbauer, W. (2006). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine* ***26***, 37–52.

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

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics* ***19***, 251–253.

Appendix A

Details of Computations

A.1 Modification of the Confidence Limits

To find how to modify the limits of the confidence interval such that the coverage remains 95% we proceed as follows: we only explain the case when the lower bound is restricted to $\frac{\psi}{A-1}$ because the procedure is the same for the upper bound. We want to find U such that $\mathbb{P}\left\{\frac{\psi}{A-1} < \Delta < U\right\} = 0.95$ so we apply the evidence function defined in Eq.(1.3) because we know that $T_A(\Delta) \sim \mathcal{N}(0,1)$. So we get $\Phi\left\{T_A(U)\right\} - \Phi\left\{T_A\left(\frac{\psi}{A-1}\right)\right\} = 0.95$, where Φ is the standard normal distribution function. Taking the inverse of these functions we find $U = T_A^{-1}\left(\Phi^{-1}\left[0.95 + \Phi\left\{T_A\left(\frac{\psi}{A-1}\right)\right\}\right]\right)$ as given in Section 2.2.

A.2 Expected Value of the Risk Ratio

We need the second derivative of $g \circ h^{-1}$ which is given as

$$(g \circ h^{-1})'' = (g'' \circ h^{-1}) \cdot (h^{-1'})^2 + (g' \circ h^{-1}) \cdot h^{-1''}.$$

The first and second derivatives of g and h^{-1} are first computed:

$$\begin{aligned} g(\Delta) &= \frac{\psi + (1-A)\Delta}{\psi - A\Delta} \\ g'(\Delta) &= \frac{\psi}{(\psi - A\Delta)^2} \\ g''(\Delta) &= \frac{2A\psi}{(\psi - A\Delta)^3} \end{aligned}$$

$$h^{-1}(\kappa) = \frac{w}{u}\sin\left\{\arcsin\left(\frac{u\Delta_0+v}{w}\right)+\sqrt{\frac{u}{2q(1-q)}}\kappa\right\}-\frac{v}{u}$$
$$h^{-1'}(\kappa) = \frac{w}{u}\sqrt{\frac{u}{2q(1-q)}}\cos\left\{\arcsin\left(\frac{u\Delta_0+v}{w}\right)+\sqrt{\frac{u}{2q(1-q)}}\kappa\right\}$$
$$h^{-1''}(\kappa) = -\frac{w}{2q(1-q)}\sin\left\{\arcsin\left(\frac{u\Delta_0+v}{w}\right)+\sqrt{\frac{u}{2q(1-q)}}\kappa\right\}.$$

Since $\kappa = \frac{1}{\sqrt{N}}T_A^{\Delta}(\Delta)$ and under $H_0 : \Delta = \Delta_0$, T_A^{Δ} follows a standard normal distribution, the expected value of κ is 0 and its variance is N^{-1} so the expected value of ρ is $\mathbb{E}[\rho] = (g \circ h^{-1})(0) + \frac{(g\circ h^{-1})''(0)}{2N}$. Evaluating the above functions at 0 gives the following:

$$h^{-1}(0) = \Delta_0$$
$$h^{-1'}(0) = \frac{w}{u}\sqrt{\frac{u}{2q(1-q)}}\cos\left\{\arcsin\left(\frac{u\Delta_0+v}{w}\right)\right\}$$
$$h^{-1''}(0) = -\frac{u\Delta_0+v}{2q(1-q)},$$

and thus, putting everything together leads to the expected value of ρ

$$\mathbb{E}[\rho] = \frac{\psi+(1-A)\Delta_0}{\psi-A\Delta_0} - \frac{\psi(u\Delta_0+v)}{4Nq(1-q)(\psi-A\Delta_0)^2} + \frac{A\psi\{w^2-(u\Delta_0+v)^2\}}{2Nuq(1-q)(\psi-A\Delta_0)^3}$$

given in Eq.(2.1).

A.3 Expected Value of the Odds Ratio

The method is the same as for relative risk but here we need the derivatives of the function f given in Eq.(1.10) which are

$$f(\Delta) = \frac{a\Delta^2+b\Delta+c}{a\Delta^2+(b-1)\Delta+c}$$
$$f'(\Delta) = \frac{-a\Delta^2+c}{\{a\Delta^2+(b-1)\Delta+c\}^2}$$
$$f''(\Delta) = \frac{2a^2\Delta^3-6ac\Delta-2c(b-1)}{\{a\Delta^2+(b-1)\Delta+c\}^3}.$$

The key function h is the same as for the risk ratio so, using the above results, the expected value of γ is

$$\mathbb{E}[\gamma] = \frac{a\Delta_0^2+b\Delta_0+c}{a\Delta_0^2+(b-1)\Delta_0+c} - \frac{(-a\Delta_0+c)(u\Delta_0+v)}{4Nq(1-q)\{a\Delta_0^2+(b-1)\Delta_0+c\}^2} + \frac{\{2a^2\Delta_0^3-6ac\Delta_0-2c(b-1)\}\{w^2-(u\Delta_0+v)^2\}}{4Nuq(1-q)\{a\Delta_0^2+(b-1)\Delta_0+c\}^3}.$$