First line of title
second line of title

Master Thesis in Biostatistics (STA495)

by

Name of student

Matriculation number

supervised by

Name of responsible supervisor (with title)

Name of supervisor (with title and affiliation if external)

Zurich, month year

p-values: their use, abuse and proper use illustrated with seven facets

Mäxli Musterli

Version June 11, 2019

Contents

Preface

Howdy!

Max Muster
June 2018

Chapter 1

Introduction

Typically, we expect empirical scientific results to be distorted by random noise, such that the results are not equal to the real process that they describe. Additionally, one can reasonably assume that the results are also biased to some extent, as the statistical paradigm of precision versus bias would predict: The smaller the noise part of distortion, the more the result should be distorted by bias. However, the trade-off is not inevitable and can be minimized by scientific rigour and foresightful experimental design. Large efforts have been made both in theory and practice to improve the quality of experiments and their analysis, such as defining experimental settings that prohibit influence of expectations of the researcher (e.g. blinding) or increase sample size in experiments. The introduction of randomized clinical trials is for example seen nowadays as a benchmark in clinical science, heavily improving the reliability of its findings.

It is more and more considered a new scientific field by itself to think of and argue about circumstances that improve the quality of scientific findings. Statistics is in some sense predestined to contribute to it, because probability and chance concepts provide are already at the core of understanding of each empirical science.

There are however issues in which statistics is not thought to be able to contribute much understanding and to offer little help. For example, the selective attention that agents in modern science are paying to new, strong and clear findings in each science is considered merely a psychological, socioeconomic issue. It is often explained by the affinity of humans for stories rather than mere facts and for novelties. This issue, that scientific findings get more attention, and today, are more likely to be published, read and cited, is known for many years now, but has only gained new traction in meta science when the reproducibility crisis emerged in 2003 to 2005. It was the studies of ... that showed that even rigid prosecution of study protocolls did often not lead to reproduction of previous results when experiments were repeated, and struck empirical science in its core. Most reasonably, ... concluded that it was a large missunderstanding of statistical tools such as the p-value in the scientific community that lead to this situation. Although that some measures have been taken since then, and some progress is made, the non-reproducibility of scientific fiendings remain an accute threat to the relevance and reliability of science. Some even argue that certain scientific fields sooner or later will repeat the experience that for example psychology or medicine have made, and are yet to encounter their own reproducibility crisis.

Clearly, reproduction of experimental results is the gold standard of testing empirical science, because it reassures the universality and reliability by which certain procedures or effects appear if measured in the correct experimental context. However, there are also problems there. One can almost never fully exclude chance events to have played a major role in negating or reafirming the experimental findings (which is of course always the case where noise is present in the data). Furthermore, reproduction studies are generally costly and the precise experimental conditions might be hard to reproduce because of lack of information or other reasons. There is another way to assess the strength of scientific results if the experiment has been conducted more than once (however not with the same study protocoll as in reproduction studies): Meta-analyses. In a meta-analyses, results of multiple (fairly identical) experiments (studies) are summarized to a single result. Very often, meta-analyses are not only a synthesis of results, but of evidence, thus reflecting the overall and summarized evidence regarding to a scientific question. It is on purpose that selective attention to positive, strong results in the scientific literature have been mentioned beforehand, because it is now clear how they will affect this second way of reassurance of scientific findigs: The meta-analysis will again lead to irreproducible findings if it is based on a set of results that are all positive since they are easier available and more prominent in the scientific literature. Thus, meta-analyses will in this case not reach their purpose of assessing reliability of science, but worsen the problem by reinforcing the confidence in the overestimated effects in the experiment. However, and this is the main topic in this masters thesis, there are some ways to detect irregularities in the body of results that can provide indirect hints that a selective rather than a random sample out of a hypothetical population of experiments is present. The abundance of the methods to link some features of the sample of experiments to publication bias, as the tendency of scientific literature to include over-proportionally large effects is generally termed, also speaks for the relevance of the topic. A review in 2017 identified alone 147 (!) conceptually differing methods.

This speaks as well for the inevitable difficulties that such methods encounter. First of all, it is most often impossible to estimate the real selection process, that is the rate by which smaller effects go unnoticed because of selective publication and reporting, because the number of missing results in the studies is not known at the first place and impossible to retrieve. Secondly, there is almost no real world data of complete and unbiased meta-analyses, such that evaluations of methods is most often dependent on simulations. So we have, after applying the methods, only indirect information of publication bias, and the extent to what it might influences scientific findings. There are, almost all the time, alternative explanations for the results of the methods that relieve the operators and publishers from the reproval of publication bias. However, given the large body of evidence for publication bias to play a role in science that has been collected in other ways, those can be expelled most of the time, such that those methods can give a quantitative measure of publication bias. Before beginning with the main part of this masters thesis, I want to recall some of the most pointed evidence for publication bias that has been collected for publication bias so far.

Meta-analysis is at the core of evidence based medicine because it allows to summarise evidence over multiple studies and provide a more broad view on success and effectivity of clinical treatments. The necessity of meta-analyses is also increased by the abundance of data and publications. Especially when the findings differ or even contradict between studies, meta-analysis is the only way to go if one wants to make decisions based on quantitative and scientific criteria.

For this, meta-analyses do not only benefit research, but also clinical practice, and may lead to better health care and prevention. However, the usefulness of meta-analysis does not restrict to clinical science, but to any empirical and quantiative science.

Usually, a meta-analysis is part of a systematic review where researchers decided to summarise all research in a given field or more specifically, that concerns a given question. Meta-analysis can be applied to all studies that are approximately identical in their experimental setup and the way the outcome of the experiments is measured. In systematic reviews where meta-analyses are used, the conclusions are most often strongly based on the results and the interpretation of the meta-analysis.

However, there are problems that potentially limit the validity of meta-analysis; the number of studies available can be incomplete or the results of the studies can be biased. Some of those problems can be solved or asserted by special statistical methods.

1.0.1 Small Study Effects or Publication Bias

When study sample size decreases, the probability of extreme and missleading results in a study increases. This becomes a problem if results are selectively published, and therefore available, based on their results. When this is the case, one speaks of a small study effect or of “Publication bias”.

The issue has been discussed extensively in the last years, most often in the context of what has came to be known as the replication crisis. The reasons for small study effects are manifold, but originate most often in the myopical acting of agents in science and the lack of statistical education. Studies are reported by scientists, published by journals and noticed by readers more often if their findigs are positive and find e.g. a substantive difference or effect. When doing a meta-analysis, one again obtains biased results.

The reason why that is less of an issue for larger studies is that extreme results are in general less likely and that due to larger effort, a result is published although there has been no clear and positive findings.

While there is generally no way to assert poor study quality, small study effect can in principle be asserted and corrected for statistically. This masters thesis will mainly be about statistical methods to detect and adjust for small study effects. It can furthermore be divided in two parts:

- Methodological part: Collection and discussion of statistical tests and correction methods for small study effects.
- Applied part: Application of the methods to studies of the Cochrane Library of systematic Reviews. Subsequent discussion of the implications of the results for clinical science.

In contrast to simulation studies, it is not possible to assess critical properties of the methods such as the power of a test, since the truth is not known. But based on the amount of data, one can of course try to make extrapolation to tendencies in clinical science in general. Moreover, it is still interesting to see how the methods behave in general, especially with respect to each other. It may, as an example, be possible to answer the question which statistical test is most conservative and which pooling method is most optimistic on average. Comparison with results from simulations may allow to speculate about the reasons when simulation and real world results diverge.

1.0.2 Cochrane and the Cochrane Database of Systematic Reviews

The Cochrane Organization has specialized on systematic reviews in clinical science. It publishes and maintains a library with a large number of systematic reviews that are available in some countries to the public.

The data analyzed in this thesis stems completely from the Cochrane Library of systematic Reviews (cite).

The reviews are arguably of good quality, since the authors are following elaborated guidelines, and there are control-mechanisms within the organisation that should prohibit conflicts of interests. This might further improve the validity and precision of findings and conclusions that have been made based on this data.

Chapter 2

Methods

2.1 Introduction and Notation

The analysis has already been said to be restricted on clinical or health care interventions. The interventions are restricted to two participant arms of which some measure of sanity or worsening of health is measured and compared. The difference between these two is referred to as the effect of the treatment. Where it is not particularly mentioned, the term treatment effect refers to any effect measure such as log risk ratio, log hazard ratio, Cohen's d or Hedges g, fisher's z score or pearson correlation.

Let us consider a meta-analysis with n study treatment effects ($n > 1$, but typically small). A study is indexed by i, and it's treatment effect by θ_i. The observed treatment effect is $\hat{\theta}_i$. The pooled treatment effect of a meta-analysis will be denoted as θ_M, and consequently, the observed pooled treatment effect as $\hat{\theta}_M$. Furthermore, each treatment effect is typically measured with some standard error s_i and an estimate of s_i is given by $\hat{s}_i$.

For continuous outcomes, let m_t be the mean of the treatment group, m_c the mean of the control group, and equivalentyl sd_t and sd_c the corresponding standard deviations. n_t and n_c are the total number of participants in the groups. In the case of binary outcomes, let e_t be the count of events in the treatment arm e_c the count of events in the control group. The observed counts in a study i are referred to as $e_{t,i}$ and analogously $e_{c,i}$.

2.2 Effect Measures and p-values

2.2.1 Continuous Outcomes

For given $(m_t, m_c), (sd_t, sd_c)$ and (n_t, n_c), one can compute mean difference as well as a standardized mean difference (here: Hedges g) and a standard error thereof. Note that the definition of Hedges g and its standard error s varies amongst the literature, the following applies for this report:

$$s = \sqrt{\frac{(n_t - 1)sd_t^2 + (n_c - 1)sd_c^2}{n_t + n_c - 2}} \qquad g = \frac{m_t - m_c}{s} \tag{2.1}$$

The mean difference θ and its standard error s can similarly be obtained by

$$\theta = m_t - m_c \qquad s = \sqrt{sd_t^2/n_t + sd_c^2/n_t} \tag{2.2}$$

Both estimators take into account that the two groups might have unequal variances. A p-value to test the null hypothesis that the mean between group is equal is commonly obtained with the students t test. The t statistic is obtained, using s and g from 2.1, by

$$t = g/(s\sqrt{(1/n_t) + (1/n_c)})$$

and the p-value can be obtained with the cumulative student's t-distibution F with $n_t + n_c - 2$ degrees of freedom:

$$p = 2(1 - F(|t|))$$

The t-test is known to be less reliable if combined sample size is small ($n < 30$), see for example Kasuya (2001).

2.2.2 Binary Outcomes

Two commonly used effect measures for binary outcome data are risk ratios and odds ratios between treatment and control group. The methods presented here can also be found, for example in Borenstein *et al.* (2011), pg. 34. Let θ be the logarithm to base 10 of the odds ratio. $\hat{\theta}$ and its variance s^2 can be obtained by

$$\hat{\theta} = \log(\frac{e_t * (n_c - e_c)}{e_c * (n_t - e_t)})$$
$$\hat{s}^2 = 1/e_t + 1/(n_t - e_t) + 1/e_c + 1/(n_c - e_c)$$

The logarithm of the risk ratio θ and its variance s^2, can be estimated by

$$\hat{\theta} = \log(\frac{e_t/n_t}{e_c/n_c}) \tag{2.3}$$
$$\hat{s^2} = 1/e_t + 1/n_t + 1/e_c + 1/e_t \tag{2.4}$$

Using likelihood theory, one could show that the estimates are maximum likelihood estimates and that one can use the asymptotic normal distribution of the maximum likelihood estimator to calculate a p-value (Held and Sabanés Bové (2014), pg.98).
Thus, with Φ as the cumulative standard normal distribution, we get

$$p = 2 * (1 - \Phi(|\hat{\theta}/\hat{s}|)$$

which summarizes the evidence against θ being zero (i.e. the true risk/odds ratio being 1). Binary and continuous effect measures can be converted into each other. In section 2.6 at the end of this chapter, this is readily described.

2.2.3 Survival Outcomes

Time-to-event data with censoring present have to be analyzed by special means. One frequently used method to take into account right-censoring is the Cox proportional hazards regression model (Cox, 1972). Because the method itself is not applied in this thesis, but only the resulting estimates of the parameters are used, the reader is referred to the extensive literature covering this topic (e.g. Cox and Oakes (1984)). The so-called hazard ratio estimated by cox p.h. regression is the ratio of the instanteneous risk of experiencing the event between the two groups. Because it is a maximum likelihood estimator, one can again use its wald test statistic to test for equal hazards.
Let $\hat{\theta}$ be an estimate of the log hazard ratio and $\hat{s}$ an estimate of the standard error of it. As before

$$p = 2 * (1 - \phi(|\hat{\theta}/\hat{s}|)$$

will give a p-value for the evidence against the null hypothesis.

2.3 Fixed and Random Effects Meta-Analysis

The fixed effects meta-analysis estimator of the pooled treatment effect is a mean of the single treatment effect estimators, weighted by their standard errors Rosenthal and Rubin (1982). Let $w_i = 1/s_i^2$ be the weights, and θ_M be the pooled estimator and s_M^2 its variance. Then

$$\theta_M = \frac{\sum_{i=1}^n w_i \theta_i}{\sum_{i=1}^n w_i} \qquad s_M^2 = \frac{1}{\sum_{i=1}^n w_i} \tag{2.5}$$

This estimator minimizes the variance, an estimate $\hat{\theta}_M$ can be obtained by plugging in the observed treatment effects and standard errors $\hat{\theta}_i, \hat{s}_i^2$. The underlying idea is that we assume $\hat{\theta}_i \sim N(\theta_M, s_i^2)$, θ_M being the true effect, such that all θ_i are disitributed equally.
The random effects model (Whitehead and Whitehead, 1991) assumes instead that

$$\theta_i \sim N(\mu_i, s_i^2) \qquad \mu_i \sim N(\theta_M, \tau^2) \tag{2.6}$$

So marginally, we have θ_i being distributed around a common mean θ_M with additional variance τ^2:

$$\theta_i | \mu_i \sim N(\theta_M, s_i^2 + \tau^2) \tag{2.7}$$

τ^2 is often referred to as a population variance or between-study variance, whereas s_i^2 can be interpreted as sampling error. The pooled treatment effect and its variance is obtained by replacing the weights w_i in equation 2.5 with $w_i = 1/(s_i^2 + \tau^2)$.
The model is clearly superior to the fixed effects model whenever the standard errors of the treatment effects alone are unlikely to fully account for the entire variability between studies. Note that as τ^2 increases, each θ_i will eventually get equal weights, irrespective of sampling error. The estimation of τ^2 has been subject to some debate in the statistical literature. Oftentimes, the method of moment estimator of DerSimonian and Laird (1986) is used. We use the measure of heterogeneity, Q, and divide by C after having subtracted the degrees of freedom $n-1$:

$$Q = \sum_{i=1}^n w_i (y_i - \theta_M)^2 \qquad C = \sum_{i=1}^n w_i - \frac{\sum_{i=1}^n w_i^2}{\sum_{i=1}^n w_i} \tag{2.8}$$

$$\tau^2 = \max(0, \frac{Q-(n-1)}{C}) \tag{2.9}$$

Again, w_i, θ_i... have to be replaced by their estimates in order to get an estimate $\hat{\tau}^2$.
The Paule-Madel estimator (Paule and Mandel, 1982) is considered to have most often better properties than the method of moments estimator (Veroniki *et al.*, 2016). Since we defined $w_i = 1/(s_i^2 + \tau^2)$, it is straightforward that
For any w_i, the variance can be estimated and equated to its expected value:

$$s^2(w_i\theta_i) = \frac{\sum_{i=1}^n w_i(\theta_i - \theta_M)^2}{n-1} \qquad \frac{\sum_{i=1}^n w_i(\theta_i - \theta_M)^2}{n-1} = 1 \tag{2.10}$$

θ_M can be estimated using equation 2.5 and the new weights w_i, thus, the only problem is to estimate τ^2. It can be obtained through an iterative process, using a newly defined function

$$F(\tau^2) = \sum_{i=1}^n w_i(\theta_i - \theta_M)^2 - (n-1)$$

In view of equation 2.10, τ^2 must be such that $F(\tau^2) = 0$. We start with a arbitrary τ^2, iteratively add a term τ_0^2 and update τ^2 until $F(\tau^2 + \tau_0^2)$ is close to zero. Using a truncated taylor series expansion, one can obtain the partial derivative after τ^2, which is a reasonable choice for τ_0^2. Using $\tau^2 + \tau_0^2$ for $\hat{w}_i, \hat{theta}$, we can update $F(\tau^2)$ and check convergence to zero.
The estimation of τ^2 is accompained by uncertainity. A common procedure is to test if it is there is significant heterogeneity between the studies (Borenstein *et al.* (2011), pg. 109). Compute Q, as given in 2.8, based on one of the estimators of τ^2. It is assumed that Q follows a central Chi-squared distribution with $n-1$ degrees of freedom under the null hypothesis of equally distributed effect sizes. Thus, the expected value of Q is $n-1$, and the excess dispersion is $Q-n+1$. The p-value against the null hypothesis of equally distributed effect sizes is, using F as the cumulative distribution function of the Chi-squared distribution with d.f. $= n-1$, $1 - F(Q)$.
An advantage of the τ^2 is that it is directly linked to the variability in the data. Additionally, one can use the I^2 statistic to see what portion the between study variance has of the overall variance. It is computed as

$$I^2 = (Q - n + 1)/Q \tag{2.11}$$

Large values for $\hat{I}^2$ are of great importance when analysing e.g. small study effects.
Importantly, all proposed methods above assume normality and proper estimation of standard errors. This assumptions are not met for very small sample sizes and very few event counts. An alternative in the latter case is the mantel-haenszel method for risk and odds ratios (see e.g. Fleiss *et al.* (2013)).

2.4 Small Study Effects Tests

The tests that will be presented on the following pages are a common way to detect publication bias. Importantly, they are however not interpretable directly as evidence for publication bias, but this is left for the discussion chapter. The tests are also often referred to as funnel plot asymmetry tests, because of the popularity of a recent test that has been used frequently to test and adjust for publication bias, which goes under the name of trim-and-fill Duval and Tweedie (2000). However, it is known for some time that the method has disadvantageous properties, therefore, it will not be discussed here, as well as the funnel plot (the radial plot will be used as an alternative).
A test for small study effects will test in some ways if the size of the estimated treatment effect of a study depends on some measure of its size. An association between study size and treatment effect size can be interpreted as an artifact of publication bias.

2.4.1 Continuous Outcome Tests

For a continuous outcomes that are normally distributed, the sample mean and variance are independent of each other (Schwarzer *et al.* (2015), pg 120). Thus, the estimated standard errors of treatment effects can be used as a proxy for study size, as they should in principle not be tied to the effect size.

Begg and Mazumdar: Rank Correlation Test

Begg (1988) proposed a rank based test to test the null hypothesis of no correlation between effect size and variance. A standardized effect size $\theta_i^\star$ can be computed as in 2.12. $s_i^{2\star}$ is the variance of $\theta_i - \theta_M$ as defined in 2.13, θ_M being the fixed effects pooled treatment effect (2.5.

$$\theta_i^\star = (\theta_i - \theta_M)/s_i^{2\star} \tag{2.12}$$

$$s_i^{2\star} = s_i^2 - 1/\sum_{i=1}^{n} \frac{1}{s_i^2} 1 \tag{2.13}$$

A rank correlation test based on Kendall's tau is then used. First, the pairs are ordered after their ranks based on $s^{2\star}$. Then, for each $s^{2\star}$ rank, the corresponding ranks based on $\theta^{2\star}$ that are larger are counted and summed up to u. The number of ranks based on $\theta^{2\star}$ that are in contrary, smaller, are counted and summed up to l. Then the normalized test statistic Z is given as

$$Z = (u - l)/\sqrt{n(n-1)(2n+5)/18}$$

Thus, large number of concordance between pairs will reflect in large $\hat{u}$ and small $\hat{l}$ and thus lead to a large $\hat{Z}$. The p-value is obtained using the standard normal distribution Φ:

$$p = 2 * (1 - \Phi(|Z|))$$

The changes that have to be made int the case of ties are small and can be found in (Begg, 1988, 410).

Egger's Test: Weighted Linear Regression Test

First, the concept of simple linear regression is introduced. In short, the model assumes a dependent variable y to be a linear function of another explanatory variable x:

$$y = \beta_0 + \beta_1 x + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{2.14}$$

ϵ is the residual noise term that becomes necessary when n pairs (x_i, y_i) are given and there is no exact solution. Then it is common to look for the solution that minimizes the squared residuals, the least-squares solution. Formally,

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}}(\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i) \tag{2.15}$$

Let $\mathbf{X}$ be a matrix whith the explanatory variables x and $\mathbf{y}$ a corresponding vector for all y:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} \\ 1 & x_{22} \\ 1 & x_{32} \\ \vdots & \vdots \\ 1 & x_{n2} \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

Let $\beta = (\beta_0, \beta_1)^\top$. It can be shown that

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{2.16}$$

Is an estimator of β that mimizes the squared residuals. Let $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ and $\hat{\mathbf{r}} = \hat{\mathbf{y}} - \mathbf{y}$. The variance estimates $\hat{\sigma}^2$ and $\hat{\mathbf{s}}_\beta^2$ are

$$\hat{\sigma}^2 = \frac{1}{n-2}\mathbf{r}^\top \hat{\mathbf{r}} \qquad\qquad \hat{\mathbf{s}}^2_\beta = \hat{\sigma}^2(\mathbf{X}^\top \mathbf{X})^{-1} \tag{2.17}$$

The estimate $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1$ the slope of the regression line. Furthermore, in the simple linear regression setting, $\hat{\beta}_0$ can also be obtained by:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\bar{x}, \bar{y}$ denoting the sample means of the corresponding values $x_1, .., x_n$ and $y_1, ..., y_n$. Thus, $\hat{\beta}_0$ is also called global mean. To test whether there is evidence for the intercept β_0 to be unequal to some value β_{H0}, a t-test can be used.

$$p = 2(1 - F(|(\beta_0 - \beta_{H0})/s_{\beta_0}|))$$

where F is the cumulative t distribution with $n-2$ degrees of freedom. p will give the evidence against the null hypothesis $\beta_0 = \beta_{H0}$.
The concept is extendable to weigthed linear regression. Weigthed linear regression may be used if the residuals $\mathbf{r}$ have unequal variances, which is equivalent to ascribe different precision to the observed y. The least squares equation 2.15 is extended to

$$\underset{\beta_0,\beta_1}{\text{argmin}}(\sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_i))$$

with positive weigths w_i that penalize large squared residuals for some i more if w_i is larger compared to other w_i.
Let $\mathbf{W}$ be a $n \times n$ matrix with $\mathbf{W}_{ii} = w_i$, the weigths on the diagonal and zeros on the off-diagonals. The estimates in 2.16 and 2.17 can stil be used if $\mathbf{X}$ is exchanged with $\mathbf{X}^\star = \mathbf{WX}$ and $\mathbf{y}$ with $\mathbf{y}^\star = \mathbf{Wy}$.
Now it is shown how linear regression can be used to test dependency of effect sizes on study sizes. The simplest application was introduced by Egger *et al.* (1997). Let θ/s be the dependent variable y and $1/s$ the explanatory variable x. If plotted, this corresponds to a radial or galbraith plot Galbraith (1988). The linear regression equation as introduced before in 2.14 can be written in two ways:

$$\theta/s = \beta_0 + \beta_1/s + \epsilon, \qquad\qquad \epsilon \sim N(0, \sigma^2) \tag{2.18}$$

2.18 is often provided due to the correspondance to the radial plot. However, it is equivalent to

$$\theta = \beta_0 + \beta_1 s + \epsilon, \qquad\qquad \epsilon \sim N(0, w^{-1}\sigma^2) \tag{2.19}$$

with weigths $w = 1/s^2$. Thus testing β_0 of 2.18 or β_1 is equivalent. The corresponding p-value is then used as evidence for a small study effect. Plugging in $\theta_i/s_i, 1/s_i$ as y_i, x_i into equation 2.16 and 2.17 will give the estimates for $\hat{\beta}_0, \hat{\beta}_1, \hat{s}_{\beta_0}$ and $\hat{s}_{\beta_1}$.

Thompson and Sharp's Test: Weighted Linear Regression Random Effects Test

A method proposed in Thompson and Sharp (1999) allows for between study variance τ^2, as introduced before in section 2.3. It extends the previously seen linear regression approach with $x = 1/s$ and $y = \theta/s$ by introducing weights. The effect size θ_i is assumed to be distributed as

$$\theta_i \sim N(\beta_0 + \beta_1 s_i, s_i^2 + \tau^2) \tag{2.20}$$

τ^2 is estimated as in equation 2.9 (method of moments). The weigths are set as $w_i = 1/\sqrt{s_i^2 + \tau^2})$. After adjusting for the weights as described in 2.4.1, we can proceed analogous to Egger's test. The p-value for $\beta_0 \neq 0$ reflects the evidence for a small study effect.

2.4.2 Dichotomous Outcomes Tests

The issue with dichotomous outcomes is that effect size and variance of effect size are correlated, which can readily be seen in 2.3 and 2.4. For example, a small number of event counts in one or group will inflate the variance and the effect size. Consequently, the tests above will tend to reject the null-hypothesis too often, i.e. report false positives. A number of solutions to this problem are provided in the literature.

Peters Test: Weighted Linear Regression Test

Instead of taking the standard error s as explanatory variable x as in Egger's Test, the inverse of the total sample size is used. Additionally, the variances s_i^2 are used as weights. Thus, the subsequent test procedure is identical to Egger's test. Peters test is a a small modification of Macaskill's test where the explanatory variable is the sample size instead of its inverse.

Harbord's Test: Score based Test

A rank based alternative to Peters test for binary outcomes is Harbord's test (Harbord *et al.*, 2006). It uses a different treatment effect and variance estimate: the score φ of the log-likelihood, evaluated at log odds ratio $\theta_0 = 0$ and its inverse Fisher information s^2. Formally,

$$\varphi = e_t - (e_t - e_c)(e_t + (n_t - e_t))/(n_t + n_c) \tag{2.21}$$

$$s^2 = \frac{(e_t + e_c)(e_t + (n_t - e_t))(e_c + (n_c - e_c))((n_t - e_t) + (n_c - e_c))}{(n_t + n_c)^2(n_t + n_c - 1)} \tag{2.22}$$

It can be shown that they are both good approximations of the log odds ratio and its variance if the real θ is not too far from zero. The standardized estimator r_i/v_i is also known as peto odds ratio. The obtained scores and variances can be used in Egger's test as treatment effects and variances.

Schwarzer's Test: Rank Correlation Test

Schwarzer *et al.* (2007) developed a test for the correlation between $E_t - \mathbb{E}(E_t)$ and the variance of E_t, E_t being a random variable from the non-central hypergeometric distribution, assuming a fixed log odds ratio.

$\mathbb{E}(E_t)$ and variance of E_t are then estimated based on e_t. The standardized cell count deviation

$$(e_t - \mathbb{E}(E_t))/\sqrt{(}s_i^2) \tag{2.23}$$

and the inverse of s_i^2 is then used as before in Begg and Mazumdar's test.

Rücker's Test: Using the Variance Stabilizing Transformation for Binomial Random Variables

The correlation between variance and effect size of dichotomous outcome measures can be abolished by the variance stabilizing transformation for binomial random variables. We use that the arcsine function is the variance stabilizing transformation for a proportion. Let

$$\theta_i = \arcsin e_t/n_t - \arcsin e_c/n_c s_i^2 = 1/4n_t + +/4n_c$$

Then one can optionally apply Begg and Mazumdar's rank correlation test or Thompson and Sharp's test using the newly obtained estimates.

2.5 Small study effect and Publication Bias Adjustment

There are different approaches to correct for small study effects and publication bias. They can mainly be distinguished by their underlying methods: regression based approaches aim to regress the effect to a study with infinte precision (i.e. very small standard error) or to a summary effect, corrected for publication bias. Selection modells aim to simulate different, hypothetical selection processes and attain a approximate treatment effect by sensitivity analysis:

2.5.1 Adjustment by Regression

Rücker *et al.* (2011) use a random effects model to obtain an unbiased estimate. Similarly to regression based tests for small study effects, we have

$$\theta_i = \beta_0 + \beta_1\sqrt{s_i^2 + \tau^2} + \epsilon_i\sqrt{v_i + \tau^2}, \epsilon_i \overset{\text{iid}}{\sim} N(0,1) \tag{2.24}$$

The only difference between Thompson and Sharp's variant is $x = \sqrt{s^2 + \tau^2}$ instead of $x = \sqrt{s^2}$. β_1 represents the bias introduced by small study effects, as can be seen when looking at 2.25

$$\mathbb{E}((\theta_i - \beta_0)/\sqrt{s_i^2}) \rightarrow \beta_1 \text{ if } s_i \rightarrow \infty \tag{2.25}$$

$$\mathbb{E}(theta_i) \rightarrow \beta_0 + \beta_1 \tau \mathrm{if} s_i \rightarrow 0 \tag{2.26}$$

After estimating τ^2, one can estimate β_0 and β_1 as seen before in the simple linear regression framework. Now we have basically two possible estimates at hand:

- β_0 the treatment effect without any influence of study precision with standard error s_{β_0}
- $\beta_0 + \beta_1\tau$ the treatment effect of a hypothetical study with infinite precision, corresponding standard error $s_{\beta_0} + s_{\beta_1}$

2.5.2 Copas Selection Model

A method proposed in Copas and Shi (2001, 2000); Copas and Malley (2008) assumes that the given sample of treatment effects and standard errors is a selected part of a larger random sample. Selection of studies depends on their effect size and variance. Smaller variance is always accompained by larger selection probability.
Let θ_i be the effect size estimate of study i. Then

$$\theta_i \sim N(\mu_i, \sigma_i^2)\mu_i \sim N(\theta, \tau^2) \tag{2.27}$$

which is similar to the random-effects meta-analysis setting. θ is the population mean effect, σ_i^2 the within study variance and τ^2 the between study variance. Equations in 2.27 are termed the *population model.*

The *selection model* is defined as follows. Suppose a selection of studies with reported ($\neq$ estimated) standard errors s (likely different from σ). Only a proportion of the selection will be published, with a defining the overall proportion of published studies and b (assumed to be positive) defining how fast this proportion increases with s becoming smaller. Formally,

$$P(\text{select}|s) = \Phi(a + b/s)$$

The equation can be rewritten as

$$z = a + b/s + \delta$$

with $\delta \sim N(0, 1)$. z is interpreted as the *propensity for selection.* It's sign must be positive in order for the study to be selected. Thus, the larger z, the more likely the study will be selected. Also, a is somewhat like to global retention or selection rate for each study, while b decides about the decline of selection probability with increasing s.

So far, we have, for a study i

$$\theta_i = \mu_i + \sigma_i\epsilon_i \tag{2.28}$$

$$\mu_i \sim N(\theta, \tau^2) \tag{2.29}$$

$$z_i = a + b/s_i + \delta_i \tag{2.30}$$

where (ϵ_i, δ_i) are standard normal residuals. The two models are coupled by introducing a correlation $\rho = cor(\theta_i, z_i)$ by defining (ϵ_i, δ_i) as bivariate standard normals. It follows that, if ρ_i is large and positive and $z_i > 0$, then the estimate of a study i that is selected is likely to have positive ϵ_i and δ_i and thus, the true mean μ is likely to be overestimated.

Let $u_i = a + b/s_i$, $\lambda(u_i)$ the Mill's ratio $\phi(u_i)/\Phi(u_i)$ (ϕ is the standard normal density function and Φ the cdf) and $\tilde{\rho_i} = \sigma/\sqrt{(\tau^2 + \sigma_i^2)}\rho_i$. The probability of a study being selected, given s_i and θ_i, is

$$P(\text{select}|s_i, \theta_i) = \Phi(\frac{u_i + \tilde{\rho_i}((\theta_i - \mu)/\sqrt{(\tau^2 + \sigma_i^2)})}{\sqrt{1 - \tilde{\rho_i^2}}})$$

Which again shows that larger s_i and θ_i lead to a larger selection probability. It can also be shown that the expected value

$$\mathbb{E}(\theta_i|s_i, \text{select}) = \mu + \rho_i\sigma_i\lambda(u_i) \tag{2.31}$$

which shows that the expected value for a study is larger for larger σ.

A likelihood for θ_i, conditional on $z > 0$ can be formulated to estimate the parameters of the model. Regarding a and b, there is no way that they can be estimated because the number of missing studies and their effect sizes is not known. Instead, fixed values for a and b have to be chosen. The nuisance parameter σ_i can be replaced by an estimate if the sample size in the studies is large enough. Since

$$\mathrm{Var}(\theta_i|s_i, z_i > 0) = \sigma_i^2(1 - c_i^2\rho_i^2)$$

with $c^2 = \lambda(u_i)(u_i + \lambda(u_i))$, we can replace σ_i^2 by $\hat{\sigma}_i^2 = \frac{1}{1-c_i^2\rho_i^2}$. Although one has to evaluate the likelihood for fixed pairs (a, b), one can compare the fit of the model to evaluate which one is more suitable: With equation 2.31, one can obtain fitted values of θ_i based on s_i. Also, for two different pairs (a, b), $(a^\star, b^\star)$,

$$\mathbb{E}(\theta_i|z_i > 0, a^\star, b^\star) - \mathbb{E}(\theta_i|z_i > 0, a, b) \approx c^\star + \rho(\lambda(a^\star) - \lambda(a))s_i$$

and that local departures of θ_i can be approximated by adding a linear term in s_i to the expectation of θ_i. Thus, to test a pair (a, b), it is sufficient to test $\beta \neq 0$ in

$$\theta_i = \theta + \beta s_i + \sigma_i\epsilon_i$$

with restriction that $\rho \geq 0$. To test some pair (a, b) against the scenario with no selection, we set $a^\star = \infty$ (or $\rho = 0$). A likelihood ratio test will give a test statistic to test against $H0 =$ no selection:

$$\chi^2 = 2 * (\max_{\theta,\tau,\beta}1\, \tilde{L}(\theta, \tau, \beta) - \max_{\theta,\tau} \tilde{L}(\theta, \tau, 0)) \tag{2.32}$$

with

$$\tilde{L}(\theta, \tau, \beta) = -\frac{1}{2}\sum_{i=1}^{n}[\log(\tau^2 + \sigma_i^2) + \frac{(\theta_i - \theta - \beta s_i)^2}{(\tau^2 + \sigma_i^2)}]$$

χ^2 can be used with a χ^2 distribution with one degree of freedom to obtain a p-value. Note that the test is almost equivalent to Egger's small study effect test with $\tau^2 = 0$. Thus, although the copas selection model models publication bias, it is dependent on the small study effects to find the most suitable pair (a, b).

If one applies the model to a single meta-analysis, a sensitivity analysis is suggested. One can observe how θ and it's confidence intervals change dependent on the underlying selection process. Selection models are in general not recommended for inference (**?**)selection.assessment). Rücker *et al.* (2011) have shown how the method can be implemented in a simulation for inference purposes.

A range of values of (a, b) are applied, and the test for residual small study effect as described in equation 2.32 is applied. If all obtained p-values from the test are above a threshold 0.1, this is interpreted as no evidence, and no need for modelling, and the standard, classical random effects meta-analysis is retained. If none of the p-values is above the threshold, a wider range of values for (a, b) is used. When some p-values are above, and some below the threshold, the pair (a, b) with the smallest number of missing studies is retained (that is, the least intense underlying selection model is chosen).

Currently, there is no test to detect missspecifications in the model itself, the authors themselves have argued that a nonparametric test of the residuals would lack power.

2.6 Transformation between effect sizes

Assumon that binary outcomes result from a dichtomozation of originally continuous random variables, in this case, the logistic distribution, a transformation from typical binary effect measures to Cohen's d can be achieved (Borenstein *et al.* (2011), pg. 47).
Let θ be a log odds ratio and s its standard error. Cohen's d and it's variance s_d^2 is obtained by

$$d = \theta\frac{\sqrt{3}}{\pi} \qquad\qquad s_d^2 = s^2\frac{\sqrt{3}}{\pi}$$

$\frac{\pi}{\sqrt{3}} = 1.81$ is the standard deviation of the logistic distribution $L(\mu, \eta)$ with scale parameter $\eta = 1$, so we just divide the log odds ratio and it's variance throught the standard deviation. The approximation works only well if e_t and e_c are not very small, especially in the case of s_d^2. Plugging in the observed log odds ratio $\hat{\theta}$ and $\hat{s}^2$ will give an estimate of Cohen's d.
Pearsons correlation can be attained by the formulas (Hedges and Olkin (1985), Borenstein *et al.* (2011) pg. 48-49)

$$r = \frac{d}{\sqrt{d^2 + a}} \qquad\qquad a = (n_c + n_t)^2/n_c n_t$$

where a is a correction factor if $n_t \neq n_c$. The variance of r, s_r^2 is computed by

$$s_r^2 = \frac{a^2 s_d^2}{(d^2 + a)^3}$$

Finally, we can get to a fisher's z-scaled correlation z and it's variance v_z by using

$$z = 0.5\ln(\frac{1+r}{1-r})v_z \qquad\qquad = \frac{1}{n-3}$$

Appendix A

Appendix

Maybe some R code here, probably a `sessionInfo()`

Bibliography

Begg, C. B. (1988). Statistical methods in medical research p. armitage and g. berry, blackwell scientific publications, oxford, u.k., 1987. no. of pages: 559. price Â£22.50. *Statistics in Medicine*, **7**, 817–818. 8, 9

Borenstein, M., Hedges, L. V., Higgins, J. P. T., Rothstein, H. R., and Higgins, D. J. P. T. (2011). *Introduction to Meta-Analysis*. John Wiley & Sons, Incorporated, Hoboken, UNKNOWN. 6, 8, 15

Copas, J. and Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis . *Biostatistics*, **1**, 247–262. 12

Copas, J. B. and Malley, P. F. (2008). A robust p-value for treatment effect in meta-analysis with publication bias. *Statistics in Medicine*, **27**, 4267–4278. 12

Copas, J. B. and Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research*, **10**, 251–265. 12

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall. 6

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 187–202. 6

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177 – 188. 7

Duval, S. and Tweedie, R. (2000). Trim and fill: A simple funnel-plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56**, 455–463. 8

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, **315**, 629–634. 10

Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons. 8

Galbraith, R. F. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine*, **7**, 889–894. 10

Harbord, R. M., Egger, M., and Sterne, J. A. C. (2006). A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine*, **25**, 3443–3457. 11

Hedges, L. V. and Olkin, I. (1985). Chapter 11 - combining estimates of correlation coefficients. In Hedges, L. V. and Olkin, I., editors, *Statistical Methods for Meta-Analysis*, 223 – 246. Academic Press, San Diego. 15

Held, L. and Sabanés Bové, D. (2014). Applied statistical inference. *Springer, Berlin Heidelberg, doi*, **10**, 978–3. 6

Kasuya, E. (2001). Mann-whitney u test when variances are unequal. *Animal Behaviour*, **6**, 1247–1249. 6

Paule, R. C. and Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, **87**, 377–385. 7

Rosenthal, R. and Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, **92**, 500–504. 7

Rücker, G., Carpenter, J. R., and Schwarzer, G. (2011). Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal*, **53**, 351–368. 12, 14

Schwarzer, G., Antes, G., and Schumacher, M. (2007). A test for publication bias in meta-analysis with sparse binary data. *Statistics in Medicine*, **26**, 721–733. 11

Schwarzer, G., Carpenter, J. R., and Rücker, G. (2015). *Meta-analysis with R*, volume 4724. Springer. 8

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, **18**, 2693–2708. 11

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., and Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, **7**, 55–79. 7

Whitehead, A. and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, **10**, 1665–1677. 7