Evidence for Alternative Hypotheses

Stephan Morgenthaler and Robert G. Staudte

Abstract Most researchers want evidence for the direction of an effect, not evidence against a point null hypothesis. Such evidence is ideally on a scale that is easily interpretable, with an accompanying standard error. Further, the evidence from identical experiments should be repeatable, and evidence from independent experiments should be easily combined, such as required in meta-analysis. Such a measure of evidence exists and has been shown to be closely related to the Kullback-Leibler symmetrized distance between null and alternative hypotheses for exponential families. Here we provide more examples of the latter phenomenon, for distributions lying outside the class of exponential families, including the non-central chi-squared family with unknown non-centrality parameter.

1 Introduction

Statisticians are trained to avoid 'lying with statistics,' that is, to avoid deceiving others and themselves about what the data say about questions or hypotheses. At the most fundamental level, they are battling against the power of *one* number to influence thinking, rather than two numbers. Telling someone that 'smoking doubles the risk of lung cancer' is a powerful message, likely to be accepted as a fact. But reporting the 'two', with a standard error, or reporting a confidence interval for the relative risk is likely to have far less impact. It is as if there were less reliability in the message with the greater information, no doubt because the second number reminds us of the imprecision in the first. Statisticians are not immune from this

Stephan Morgenthaler
École Polytechnique Fédérale de Lausanne, EPFL – FSB – STAP, Station 8, 1015 Lausanne, Switzerland; e-mail: stephan.morgenthaler@epfl.ch

Robert G. Staudte
Department of Mathematics and Statistics, La Trobe University, Melbourne, Victoria, Australia, 3086; e-mail: r.staudte@latrobe.edu.au

human fallibility. We often quote a p-value against a null hypothesis, or a posterior probability for a hypothesis or a likelihood ratio for comparing two hypotheses, as if they were important numerical facts, to be taken at face-value, without further question. Evans [4] in his comments on [10], makes the same point: 'Some quantification concerning the uncertainty inherent in what the likelihood ratio is saying seems to be a part of any acceptable theory of statistical inference. In other words, such a quantification is part of the summary of statistical evidence.' We agree with Evans and thus require that any measure of statistical evidence be a statistic reported with a standard deviation or other measure of uncertainty.

Another example of an incomplete message occurs frequently in the meta-analytic literature. Results are derived for the case of known weights, and then estimates of the weights are substituted in the ensuing formulae, as if no theory were needed to account for the second estimation. This works for very large sample sizes, but not for those usually encountered in practice and results in optimistically small confidence intervals, inflated coverages and many published false claims, [8], e.g. Thus we require that any measure of evidence found for individual studies of the same effect should be easy to combine to obtain an overall evidence for this effect, and the combination of evidence must be based on a sound theory (see [13] for a discussion of meta analysis). In the remainder of this section we motivate and define statistical evidence on our preferred calibration scale in which one function, called the Key Inferential Function, contains all the information required for inference.

For the sake of simplicity of presentation we restrict attention to one-sided alternatives $\theta > \theta_0$ to the null hypothesis $\theta = \theta_0$ (or $\theta \leq \theta_0$); evidence for two-sided alternatives is presented in detail in Section 17.4, p. 134 of [6]. Our third requirement is that the expected evidence in favor of $\theta > \theta_0$ should be increasing with θ and have value 0 at $\theta = \theta_0$.

Fourth, if the parameter of interest θ is estimable by a $\hat{\theta}_n$ based on n observations, with standard error $\mathrm{SE}[\hat{\theta}_n]$ of order $1/\sqrt{n}$, then the evidence for an alternative hypothesis $\theta > \theta_0$ should grow at the rate $\sqrt{n}$. This means it will require 9 times as much work to obtain 3 times as much evidence for an alternative hypothesis.

Fifth, evidence should be replicable in the sense that if an experimenter obtains a certain amount of evidence for a hypothesis, then an independent repetition of the experiment should lead to a similar result, up to sampling error. While this is true for the p-value under the null hypothesis, when the null hypothesis does not hold the variation under repetition may come as a surprise due to the highly skewed distribution of the p-value. These five motivating factors lead us to illustrate what is achievable for the simplest possible model in the next section, a model that forms the basis for all that follows.

Prototypical Example

We will now consider an example that is often discussed in elementary statistics courses. In this example each experiment produces an independent realization of a random variable $X \sim \mathbf{N}(\mu, \sigma_0^2)$, where σ_0^2 is known. This is the prototypical normal

translation model, which we would like to use as a 'universal model' for other testing problems. We want to quantify the evidence based on n experiments against the null hypothesis $\mu = \mu_0$ and in favor of the alternative $\mu > \mu_0$. Letting $\bar{X}_n$ denote the sample mean, the usual test statistic is $S_n = \bar{X}_n - \mu_0$. The corresponding evidence is the standardized version of S_n, that is, $T_n = \sqrt{n}(\bar{X}_n - \mu_0)/\sigma_0$, which is distributed as $\mathbf{N}(\sqrt{n}(\mu - \mu_0)/\sigma_0, 1)$. The way an evidence is constructed means that the expected evidence is the function of the parameters that carries all the information. In the normal shift model, this is $\sqrt{n}(\mu - \mu_0)/\sigma_0$. Because T_n has variance one, evidence can be reported as $T_n \pm 1$, indicating that it has error, with the subtext that this standard error means the same thing to everyone, because all students of statistics recognize a standard normal distribution. This standard error of 1 also becomes the unit for a calibration scale for evidence: if one observes $T_n = 3$, one knows that one has observed a result 3 times its own standard error. If one obtains $T_n = -2$, one has evidence +2 for the opposite alternative hypothesis $\mu < \mu_0$, again with standard normal unit error.

The statistical evidence $T_n = \sqrt{n}(\bar{X}_n - \mu_0)/\sigma_0$ is monotone increasing in $\sqrt{n}$ for each fixed μ; and, for each fixed sample size n, the expected evidence grows from 0 as $\mu > \mu_0$ increases. This evidence is also replicable in the sense that given $T_n = t$ and an independent $T_n^* \sim \mathbf{N}(\sqrt{n}(\mu - \mu_0)/\sigma_0, 1)$ the optimal predictor $\mathrm{E}[T_n^* | T_n = t]$ is simply t, and this estimator of the expected evidence has standard error 1.

When combining evidence from independent studies, given $T_{n_1} \sim N(\tau_1, 1)$ and $T_{n_2} \sim N(\tau_2, 1)$, it is easy to think of combinations of T_{n_1}, T_{n_2} which remain on the same calibration scale. An effective combination is given in Section 1.2.

1.1 Desirable properties of statistical evidence

Most statisticians, including us, would prefer an axiomatic approach to statistical evidence, but we provide an operational one. That is, guided by the above example, we state what properties we would like a measure of evidence to have, and then in specific problems show there are indeed statistics which come close to satisfying them. *The fact that it is an approximate theory in no way reduces its usefulness.* Normal approximations via the Central Limit Theorem are ubiquitous in statistics, because they are useful in computing approximate p-values and confidence intervals. Similarly they are useful in providing evidence for alternative hypotheses.

Let θ represent an unknown real parameter for which it is desired to test $\theta = \theta_0$ against $\theta > \theta_0$, and let S_n be a test statistic based on n observations which rejects H_0 for large values of S_n. We want a measure of one-sided evidence $T_n = T_n(S_n)$ to satisfy:

E_1. The evidence T_n for a one-sided alternative is monotone increasing in S_n;
E_2. the distribution of T_n is normal for all values of the unknown parameters;
E_3. the variance $\mathrm{Var}[T_n] = 1$ for all values of the unknown parameters; and
E_4. the expected evidence $\tau(\theta) = \mathrm{E}_\theta[T_n]$ is increasing in θ from $\tau(\theta_0) = 0$.

In E_2 we require that the evidence *always* be unit normal, not only under the null hypothesis. As a consequence, the evidence proposed here carries much more information than results that are only true under the null hypothesis. For the prototypical model and $T_n(\bar{X}_n) = \sqrt{n}(\bar{X}_n - \theta_0)/\sigma_0$ all of the above properties hold exactly. Property E_1 is essential if the evidence is to remain a test statistic. In general, properties $E_2 - E_4$ will hold only approximately, but to a surprising degree, even for small sample sizes, provided one can find a *variance stabilizing transformation* (VST), of the test statistic S_n, $T_n = h_n(S_n) - h_n(\mathrm{E}_{\theta_0}[S_n])$ such that $\mathrm{Var}_\theta[T_n] \doteq 1$ for θ of interest. From now on, the symbol $\doteq$ signifies an approximate equality up to an error of smaller order in n. Since the variance of S_n is usually of order n^{-1}, the VST can usually be chosen as $h_n(\cdot) = \sqrt{n}\,h(\cdot)$.

Kulinskaya, Morgenthaler and Staudte ([6], denoted KMS in the following) propose a measure of evidence in favor of alternative hypotheses that is based on a transformation of the usual test statistic to a normal translation family with unit variance, and provide numerous applications of it to standard problems of meta-analysis. Our purpose here is to explain in more detail why we advocate this particular definition. Connections with other measures of evidence, such as the p-value and Bayes factor, are given in [9].

It turns out that the *expected KMS evidence*, when dealing with a sample of size n instead of a single observation, is equal to a product of two terms, the square root of n and a quantity $\mathscr{K}$ whose value indicates the difficulty in distinguishing the null density f_{θ_0} from an alternative density f_{θ_1}. This second term is the key to understanding and implementing inferential procedures (see 1.2 for details).

We restrict attention to a real-valued parametric family $f_\theta(x)$, where the testing problem of interest is $\theta = \theta_0$ against $\theta > \theta_0$. The elements of a traditional test are the test statistic S_n and its distribution under the null. To obtain a measure of evidence one needs a monotone transformation $T_n = h_n(S_n)$, which stabilizes the variance and is such that the distribution of T_n is approximately normal for all parameter values θ, *not only for the null value* θ_0.

When the observation x is a realization of $X \sim f_{\theta_0}$, the likelihood ratio statistic on average favors f_{θ_0}, which means that $\mathrm{E}_{\theta_0}[\log(f_{\theta_0}(X)/f_\theta(X))] > 0$. This is a good measure of the difficulty in distinguishing f_{θ_0} from f_θ based on data from f_{θ_0}. It turns out that the symmetrized version of this quantity, the Kullback-Leibler Divergence, is closely linked to the function $h_n(\cdot)$.

Beginning with Fisher in [5], many statisticians have investigated 'normalizing' a family of distributions through a transformation which often simultaneously stabilizes the variance, see the Wald Memorial Lecture by Efron [3]. As he points out, the purpose of transforming a test statistic so that its distribution is a normal translation family is both aesthetic (to gain insight) and practical (to easily obtain a confidence interval for an unknown parameter). To these desirable properties we would add that this calibration scale is ideally suited for meta-analysis, because it allows for cancelation of evidence from conflicting studies, and facilitates combination of evidence obtained from several studies. Concerning this last point, the established theory of meta-analysis (see [1] or [12]), is a large-sample theory that is not very reliable for small sample sizes. Its implementation depends on estimators of weights and these

estimators can be highly variable even for moderate sample sizes. By using variance stabilization first, researchers can apply the meta-analytic theory with much more confidence because, after transformation, no weights need to be estimated.

There is a constructive method for finding potential VSTs, see, for example, p. 32 [2] or Chapter 17 of [6]. These transformations are monotone increasing, so satisfy property E_1. They are defined only up to an additive constant, which may be chosen so that T satisfies property E_4. Variance stabilized statistics are often approximately normally distributed, and when they are so, the potential evidence T also 'satisfies' E_2. The degree of satisfaction can be measured by simulation studies that show the VST leads to more accurate coverage of confidence intervals and more accurate estimates of power functions than the usual Central-Limit based approximations of the form $(S_n - \mathrm{E}_{\theta_0}[S_n])/\sqrt{\mathrm{Var}_{\theta_0}[S_n]}$. [3] provides a constructive method for finding normalizing transformations.

1.2 Key Inferential Function

Suppose that one has in hand a measure of evidence T_n satisfying $E_1 - E_4$, at least asymptotically. In that case the expectation $\tau(\theta) = \mathrm{E}_\theta[T_n]$ summarizes the complete information. If we found T_n by application of a VST, that is, $T_n = h_n(S_n) - h_n(\mathrm{E}_{\theta_0}[S_n])$, then we can deduce $\tau(\theta) \doteq h_n(\mathrm{E}_\theta[S_n]) - h_n(\mathrm{E}_{\theta_0}[S_n])$, which can usually be written as $\tau(\theta) \doteq \sqrt{n}\left(h(\mathrm{E}_\theta[S_n]) - h(\mathrm{E}_{\theta_0}[S_n])\right)$.

Definition 1. Let T_n be a statistical evidence with $\tau(\theta) = \mathrm{E}_\theta[T_n] \doteq \sqrt{n}\,\mathscr{K}_{\theta_0}(\theta)$. Then $\mathscr{K}_{\theta_0}$ is called the *Key Inferential Function* or simply the *Key* for this statistical model and boundary value θ_0.

In the case of the normal shift model as given in the prototypical example, we found $\mathscr{K}_{\mu_0}(\mu) = (\mu - \mu_0)/\sigma_0$, which is often called the standardized effect and denoted by the symbol δ. In the case of a VST $h_n(\cdot) = \sqrt{n}\,h(\cdot)$, we have $\mathscr{K}_{\theta_0}(\theta) = h(\mathrm{E}_\theta[S_n]) - h(\mathrm{E}_{\theta_0}[S_n])$. This last expression is simply a centered version of the VST, where the centering assures the equality $\mathscr{K}_{\theta_0}(\theta_0) = 0$.

The *Key* contains all the essential information, and knowing it enables one to solve many routine statistical problems, such as

K_1. *Choosing sample sizes:* For testing $\theta = \theta_0$ against $\theta > \theta_0$ using a sample of n observations the expected evidence is $\tau(\theta) = \sqrt{n}\,\mathscr{K}_{\theta_0}(\theta)$ for each θ. To attain a desired expected evidence τ_1 against alternative θ_1 one can choose n_1 to be the smallest integer greater than or equal to $[\tau_1/\mathscr{K}_{\theta_0}(\theta_1)]^2$.

For the prototypical model, this means $n_1 \geq \{\tau_1/\delta_1\}^2$, where $\delta_1 = (\mu_1 - \mu_0)/\sigma_0$. Also, for this model the test statistic is $T_n = \sqrt{n}\,(\bar{X}_n - \mu_0)/\sigma_0) \sim N(\tau_1, 1)$, where $\tau_1 = \sqrt{n}\,\delta_1$. Hence the power $1 - \beta(\mu_1)$ of the level α test for μ_1 is exactly $1 - \beta(\mu_1) = P_{\mu_1}(T_n \geq z_{1-\alpha}) = \Phi(\tau_1 - z_{1-\alpha})$; that is, $\tau_1 = z_{1-\alpha} + z_{1-\beta(\mu_1)}$. Now substituting this expression for τ_1 into the lower bound for n_1 gives the well known expression $n_1 \geq \{\tau_1/\delta_1\}^2 = \sigma_0^2\{z_{1-\alpha} + z_{1-\beta(\mu_1)}\}^2/(\mu_1 - \mu_0)^2$.

K_2. *Power calculations:* A Neyman-Pearson level α test based on T_n has power $1-\beta(\theta)$ against alternative θ given by

$$1-\beta(\theta) = \doteq \Phi(\sqrt{n}\,\mathscr{K}_{\theta_0}(\theta) - z_{1-\alpha}) \tag{1}$$

or

$$\sqrt{n}\,\mathscr{K}_{\theta_0}(\theta) = z_{1-\alpha} + z_{1-\beta(\theta)}\,. \tag{2}$$

Formula (1) often leads to more accurate power approximations than standard asymptotics, see [6], Chapter 22. It follows that accurate choice of sample size to obtain power at a given level is possible. Formula (2) shows that the VST expected evidence is more basic than level and power: it can be partitioned into the sum of the probits of the false positive and false negative error rates.

K_3. *Confidence intervals:* A $100(1-\alpha)\%$ confidence interval for θ is given by

$$\left[\mathscr{K}_{\theta_0}^{-1}\left(\frac{T_n - z_{1-\alpha/2}}{\sqrt{n}}\right),\ \mathscr{K}_{\theta_0}^{-1}\left(\frac{T_n + z_{1-\alpha/2}}{\sqrt{n}}\right)\right], \tag{3}$$

where $\mathscr{K}^{-1}$ is the inverse function to $\mathscr{K}$.

For the prototypical model the *Key* is $\mathscr{K}_{\mu_0}(\mu) = (\mu-\mu_0)/\sigma_0 = \delta$, so $\mathscr{K}_{\mu_0}^{-1}(\kappa) = \sigma_0\kappa + \mu_0$. Substituting $T_n = \sqrt{n}\,(\bar{X}_n - \mu_0)/\sigma_0)$ into (3) produces the confidence interval $[\bar{X}_n - z_{1-\alpha/2}\sigma_0/\sqrt{n}\,,\, \bar{X}_n + z_{1-\alpha/2}\sigma_0/\sqrt{n}\,]$.

K_4. *Meta-analysis for the fixed effects model:* Given independent $T_1,\ldots,T_K$, where $T_k = T_{n_k} \sim N(\tau_k, 1)$ with $\tau_k = \sqrt{n_k}\,\mathscr{K}_{\theta_0}(\theta)$, each evidences for $\theta > \theta_0$, let

$$T_{1:K} = \frac{\sqrt{n_1}\,T_1 + \cdots + \sqrt{n_K}\,T_K}{\sqrt{N_K}}\,, \tag{4}$$

where $N_K = \sum_k n_k$. Then $T_{1:K} \sim N(\tau_{1:K}, 1)$, with $\tau_{1:K} = \sqrt{N_K}\,\mathscr{K}_{\theta_0}(\theta)$, is the combined evidence for $\theta > \theta_0$, and a $100(1-\alpha)\%$ confidence interval for θ based on all the evidence is found by replacing the T_n of (3) by $T_{1:K}$.

For the prototypical model, where $T_k = \sqrt{n_k}(\bar{X}_k - \mu_0)/\sigma_0 \sim N(\tau_k, 1)$ with $\tau_k = \sqrt{n_k}\,\mathscr{K}_{\mu_0}(\mu)$ and $\mathscr{K}_{\mu_0}(\mu) = (\mu-\mu_0)/\sigma_0$, one has $T_{1:K} = \sqrt{N_K}\,(\bar{\bar{X}} - \mu_0)/\sigma_0$. Here, $\bar{\bar{X}}$ is the mean of all $N_K = \sum_k n_k$ observations.

Note that if the initial statistical model is reparameterised in terms of $\eta = \eta(\theta)$, where $\eta(\cdot)$ is a strictly increasing function, then the *Key* $\mathscr{K}_{\eta_0}(\eta)$ becomes the composition of $\mathscr{K}_{\theta_0}(\theta)$ with the inverse reparametrization $\theta = \theta(\eta)$, that is, $\mathscr{K}_{\eta_0}(\eta) = \mathscr{K}_{\theta(\eta_0)}(\theta(\eta))$. The transformation to the 'right parameter' $\eta = \mathscr{K}_{\theta_0}(\theta)$, for example, leads to $\mathscr{K}_{\eta_0}(\eta) = \eta$, where $\eta_0 = 0 = \mathscr{K}_{\theta_0}(\theta_0)$.

For all the above reasons the *Key* appears to contain all the information required for inference in one-parameter families, and this claim is supported by the material in the next Section 2. In it we describe the very strong link between the *Key* and the Kullback-Leibler Divergence for exponential families. In Section 3 we illustrate many of the above results for the non-central chi-squared family, which is not an

exponential family. In Section 4 we summarize the results and describe areas for future research.

2 Connection to the Kullback-Leibler Divergence

[7] is a well-written and highly informative book whose principal topic is the following measure of information

$$I(\theta_0 : \theta_1) = \mathbf{E}\left[\log\left(\frac{f_{\theta_0}(X)}{f_{\theta_1}(X)}\right)\right], \quad \text{where } X \sim f_{\theta_0}.$$

This quantity is the average value of the log likelihood ratio when choosing between the model densities f_{θ_0} and f_{θ_1} with data X that is generated by f_{θ_0}. The logarithm of the likelihood ratio $\log(f_{\theta_0}(x)/f_{\theta_1}(x))$ is taken as the information in an observation $X = x$ for discrimination in favor of $X \sim f_{\theta_0}$ against $X \sim f_{\theta_1}$ (p.5,[7]). A variety of strong arguments give backing to this choice.

Definition 2. The symmetrized information, defined as $J(\theta_0, \theta_1) = I(\theta_0 : \theta_1) + I(\theta_1 : \theta_0)$ is called the *Kullback-Leibler Divergence (KLD)* (see p. 6, [7]).

Kullback's terminology has been modified over the years, and now $I(\theta_0 : \theta_1)$ is often called the *divergence* or *directed divergence* and $J(\theta_0, \theta_1)$ the *symmetrized divergence*. When the likelihood ratio test is performed with n independent observations, both I and J for discriminating will be multiplied by n. Thus in most of the examples and theory to follow we can omit the sample size.

2.1 Example 1. Normal model

We begin with a return to the prototypical model in which there are no surprises, but the generality soon becomes clear. If f_{μ_0} and f_{μ_1} are normal densities with equal variances σ_0^2, but unequal means μ_0 and μ_1, the Kullback-Leibler Information is

$$I(\mu_0 : \mu_1) = \mathrm{E}\left[\frac{1}{2}\left(\frac{(X-\mu_1)^2}{\sigma_0^2} - \frac{(X-\mu_0)^2}{\sigma_0^2}\right)\right], \quad \text{where } X \sim f_0.$$

Therefore $I(\mu_0 : \mu_1) = \frac{1}{2}(1 + (\mu_1 - \mu_0)^2/\sigma_0^2 - 1)$ and $J(\mu_0, \mu_1) = (\mu_1 - \mu_0)^2/\sigma_0^2 = \delta^2$. The Kullback-Leibler Divergence is equal to the square of the standardized effect δ. The information for discrimination is thus equal to the square of the Key Inferential Function for the z test of the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, namely $\mathscr{K}_{\mu_0}(\mu_1) = \delta$, found in Section 1.

The above example can be extended to the case of evidence for alternative $\theta > \theta_0$ to the null $\theta = \theta_0$, for which the *Key* is $\mathscr{K}_{\theta_0}(\theta)$, where we now drop the subscript

on the parameter in the alternative $\theta > \theta_0$. We also write $J_{\theta_0}(\theta)$ for $J(\theta_0, \theta)$ to emphasize that θ_0 is fixed and θ is any value in the alternative $\theta > \theta_0$. The Kullback-Leibler Divergence (KLD) between the models $\mathbf{N}(0,1)$ and $\mathbf{N}(\mathscr{K}_{\theta_0}(\theta),1)$ is by the previous example equal to $\mathscr{K}^2_{\theta_0}(\theta)$. This suggests that we can find the *Key* and the VST $h(\cdot)$ by computing the KLD, because

$$\mathscr{K}_{\theta_0}(\theta) \approx \sqrt{J_{\theta_0}(\theta)}\,\mathrm{sgn}(\theta-\theta_0)\,. \tag{5}$$

Common examples for which this approximation is excellent for θ in a large neighborhood of the null value θ_0 are the Poisson, exponential, binomial, and the correlation coefficient of bivariate normal. It is also true for the non-central t, see [9], and the non-central chi-square models, see Section 3.

2.2 Result for exponential families

Let X have density of the form $f(x|\eta) = g(x)\exp\{\eta x - k(\eta)\}$ for x in an interval not depending on η. These densities for X are called an exponential family with natural parameter η; see [11] for background material. We further assume that $Var_\eta[X] > 0$ for all η. We want to compare the Kullback-Leibler Symmetrized Divergence with the square of the Key Inferential Function for this class of models. As a Corollary, we will compare the *Key* itself with the signed square root of the divergence.

The derivatives of the function k give the cumulants of X; so that

$$\begin{aligned} \mu = \mathrm{E}_\eta[X] &= \kappa_1(\eta) = k'(\eta) \\ \sigma^2 = \mathrm{Var}_\eta[X] &= \kappa_2(\eta) = k''(\eta) \\ \mathrm{E}_\eta[(X-\mu)^3] &= \kappa_3(\eta) = k'''(\eta)\,. \end{aligned} \tag{6}$$

Now $\mu = k'(\eta)$ has positive derivative, and therefore a monotone increasing inverse $\eta = (k')^{-1}(\mu)$ so all the cumulants of X can be written as functions of μ. For example, $\sigma^2(\mu) = k'' \circ (k')^{-1}(\mu)$.

The Kullback-Leibler Information about $f(\cdot|\eta)$ when $f(\cdot|\eta_0)$ is the density of X is

$$I(\eta_0 : \eta) = \mathrm{E}_{\eta_0}[\ln(f(X|\eta_0)/f(X|\eta))] = (\eta_0-\eta)k'(\eta_0) - k(\eta_0) + k(\eta)$$

Therefore the Divergence is

$$\begin{aligned} J_{\eta_0}(\eta) &= (\eta-\eta_0)\{k'(\eta)-k'(\eta_0)\} \\ &= \{(k')^{-1}(\mu) - (k')^{-1}(\mu_0)\}(\mu-\mu_0) = J_{\mu_0}(\mu)\,. \end{aligned}$$

If a VST $h(X)$ for X exists which has variance $\mathrm{Var}[h(X)] \doteq 1$, it must satisfy $h'(\mu) = 1/\sigma(\mu)$, and the *Key* for testing $\mu = \mu_0$ against $\mu > \mu_0$ is defined by $\mathscr{K}_{\mu_0}(\mu) = h(\mu) - h(\mu_0)$.

Proposition 1. *Suppose the model is a one-parameter exponential family and let* $J_{\mu_0}(\mu)$ *denote the Kullback-Leibler Divergence, whereas* $\mathscr{K}_{\mu_0}(\mu)$ *is the Key Inferential Function. It follows that*

$$J_{\mu_0}(\mu) = \mathscr{K}^2_{\mu_0}(\mu)\left\{1 + C_2\,(\mu-\mu_0)^2/2! + O(|\mu-\mu_0|^3)\right\}$$

and

$$sign(\mu-\mu_0)\sqrt{J_{\mu_0}(\mu)} = \mathscr{K}_{\mu_0}(\mu)\left\{1 + \frac{1}{2}C_2\,(\mu-\mu_0)^2/2! + O(|\mu-\mu_0|^3)\right\},$$

where $C_2 = \kappa_3^2(\mu_0)/\{24\,\sigma^8(\mu_0)\}$.

A proof is given in [9].

For contiguous alternatives $\theta_n = \theta_0 + O(1/\sqrt{n})$, the relative error in the approximation is of order $O(1/n)$. Thus, the approximation remains useful for alternatives that are much further removed from the null value than the contiguous ones.

The procedure based on variance stabilization is applicable beyond the context of exponential families. The basic idea of approximating a test problem by a normal translation family is not new and it is well-known that many hypothesis testing procedures, which reject for large values of S_n, take this form for large sample sizes n and contiguous alternatives. This is true in the sense that the power of the level α test of $\theta = \theta_0$ against the alternatives $\theta > \theta_0$ is approximately equal to $\Phi(z_\alpha + \sqrt{n}\,e(\theta_0)\,(\theta-\theta_0))$, where z_α denotes the α quantile of the standard normal distribution and $\sqrt{n}\,e(\theta_0) = \mu'(\theta_0)/\sigma(\theta_0) > 0$ describes the efficacy of the test statistic, where $\mu(\theta)$ and $\sigma^2(\theta)$ are the mean and variance of S_n. For the variance stabilized test statistic $T_n = h_n(S_n)$, the simpler formula $\Phi(z_\alpha + \sqrt{n}\,\mathscr{K}_{\theta_0}(\theta))$ is obtained and as we have seen, this gives a good approximation beyond contiguous alternatives. In order that these two formulae agree in a neighborhood of θ_0, it must be true that $\frac{d}{d\theta}\mathscr{K}_{\theta_0}(\theta)$, evaluated at the null value θ_0, is equal to $\mu'(\theta_0)/\sigma(\theta_0)$. Because the *VST* satisfies $\frac{d}{d\mu}h(\theta_0) = 1/\sigma(\theta_0)$, this is indeed the case.

2.3 Example 2. Poisson model

Let $X \sim$Poisson(λ) and find the evidence for $\lambda > \lambda_0$ when the null hypothesis is $\lambda \le \lambda_0$. An elementary calculation gives $I(\lambda_0 : \lambda) = \lambda - \lambda_0 + \lambda_0 \log(\lambda_0/\lambda)$, which implies that $J_{\lambda_0}(\lambda) = (\lambda-\lambda_0)\log(\lambda/\lambda_0)$. The classical VST for the Poisson model leads to $\mathscr{K}_{\lambda_0}(\lambda) = \sqrt{4\lambda} - \sqrt{4\lambda_0}$. The graphs of $\mathscr{K}_{\lambda_0}(\lambda)$ and $\sqrt{J_{\lambda_0}(\lambda)}\,\mathrm{sgn}(\lambda-\lambda_0)$ are in agreement in a relatively large neighborhood of λ_0, regardless of its value. To check this, consider the parametrization $\lambda = \lambda_0 + (\lambda-\lambda_0) = \lambda_0 + \Delta$ for which we have $J_{\lambda_0}(\lambda) = \Delta\log(1+\Delta/\lambda_0) = (\Delta^2/\lambda_0)(1-\Delta/(2\lambda_0))$, while $\mathscr{K}_{\lambda_0}(\lambda) = 2\left(\sqrt{\lambda_0+\Delta} - \sqrt{\lambda_0}\right) = \Delta/\sqrt{\lambda_0} - \Delta^2/(4\lambda_0^{3/2})$. The leading term of the signed root of J and of the *Key* is $\Delta/\sqrt{\lambda_0}$, which is the standardization obtained by dividing

the raw effect $\lambda - \lambda_0$ by the standard error at the null hypothesis. We leave it to the reader to check that the next order term also is in agreement. The classical VST suggests that when $\lambda_0 = 1$, the correct parameter to use for testing and evaluating evidence is $\eta = 2(\sqrt{\lambda} - 1)$, while the KLD gives $\sqrt{(\lambda - 1)\log(\lambda)}$.

3 Non-central chi-squared family

In this section we illustrate some of the results from Sections 1 and 2 in the context of the chi-squared family with known degrees of freedom and unknown non-centrality parameter. This model is not an exponential family.

3.1 Comparing the KLD with the Key

Let $X \sim \chi^2_\nu(\lambda)$ have the non-central chi-squared distribution with ν degrees of freedom and non-centrality parameter λ. In most applications ν is known and λ is unknown. It is not possible to compute the Kullback-Leibler symmetrized divergence (KLD) between $\chi^2_\nu(\lambda_0)$ and $\chi^2_\nu(\lambda_1)$ analytically, but because of the well-known VST, we think that it has to be

$$J(\lambda_0;\lambda_1) \doteq \left(\sqrt{\lambda_1 + \nu/2} - \sqrt{\lambda_0 + \nu/2}\right)^2 . \tag{7}$$

The approximation (7) is confirmed by computational results for many choices of ν, λ_0 and λ_1, some of which are presented in Figure 1. But first we show the motivation for the conjecture by finding the *Key* for the evidence in X when testing $\lambda \le \lambda_0$ against $\lambda > \lambda_0$.Using the fact that $\mathrm{E}[X] = \nu + \lambda$ and $\mathrm{Var}[X] = 2\nu + 4\lambda$, one can write $\mathrm{Var}[X] = g(\mathrm{E}[X])$, where $g(t) = 4t - 2\nu$. Its inverse square root has indefinite integral

$$h - \nu(x) = \int^x \frac{dt}{\sqrt{4t - 2\nu}} = \sqrt{x - \nu/2} + c\,. \tag{8}$$

Thus by the standard method (p. 32 of Bickel and Doksum, 1977), $h_\nu(x)$ is a potential VST for X. It is only defined for $x > \nu/2$, but this is not a practical restriction because

$$P_{\nu,\lambda}(X \le \nu/2) \le P_{\nu,0}(X \le \nu/2) \approx \Phi\left(-\frac{\sqrt{\nu}}{2\sqrt{2}}\right) , \tag{9}$$

which is negligible even for moderate ν. The approximation of $\mathrm{E}[h(X)]$ by $\sqrt{\mathrm{E}[X] - \nu/2} + c$ leads to (7).

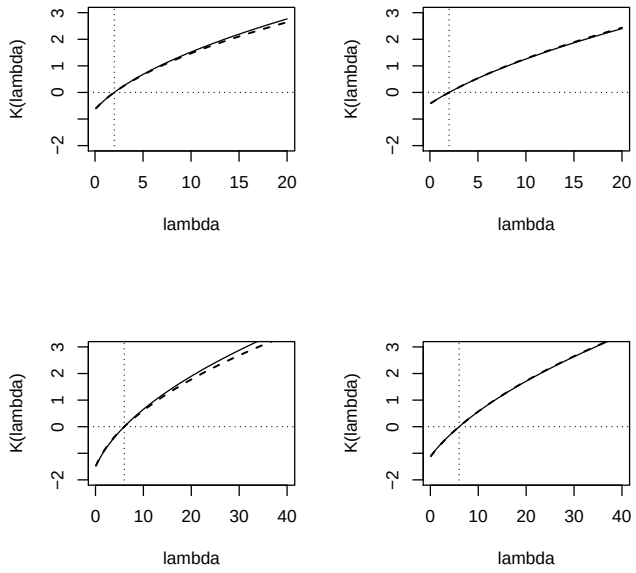

Fig. 1 In all plots the solid line depicts the graph of $\mathscr{K}_{\lambda_0}(\lambda)$, for the $\chi^2_\nu(\lambda)$ model, when testing $\lambda \leq \lambda_0$ against $\lambda > \lambda_0$. The dashed line that approximates it is the signed square root of the Kullback-Leibler symmetrized divergence $\sqrt{J_{\lambda_0}(\lambda)}\,\mathrm{sgn}(\lambda-\lambda_0)$. The latter is computed by Monte Carlo integration on R. For the two left hand plots $\nu = 3$ and in the upper plot $\lambda_0 = 2$ while in the bottom plot $\lambda_0 = 6$. For the two right hand plots $\nu = 9$ and in the upper plot $\lambda_0 = 2$ while in the bottom plot $\lambda_0 = 6$. The dotted vertical lines mark the null hypothesis.

3.2 Tests for the non-centrality parameter

Given $X_1,\ldots,X_n$ i.i.d. with $X_i \sim \chi^2_\nu(\lambda)$, it is desired to test the null $\lambda = \lambda_0$ against $\lambda > \lambda_0$ using as test statistic the sample mean $\bar{X}_n$. Any VST is derived as above to be $h_n(\bar{X}_n) = \sqrt{n}\sqrt{\bar{X}_n - \nu/2} + c$. To convert this into evidence T_n for $\lambda > \lambda_0$ we need to choose c so that $\mathrm{E}[T_n] = \mathrm{E}[h_n(\bar{X}_n)]$ is monotone increasing in λ with value 0 at the boundary $\lambda = \lambda_0$. To a first approximation, $\mathrm{E}[\sqrt{\bar{X}_n - \nu/2}] = \sqrt{\lambda + \nu/2}$ so we choose $c = -\sqrt{n}\sqrt{\lambda_0 + \nu/2}$. Then

$$\mathrm{E}[T_n] \doteq \sqrt{n}\left[\sqrt{\lambda+\nu/2} - \sqrt{\lambda_0+\nu/2}\right]. \tag{10}$$

It remains to check that T_n is approximately normal with variance near 1 and this is left to the reader. Other important results are that the evidence grows with the square root of the sample size and the *Key* function is monotone increasing in λ from 0 at

the null. The *Key* function evidently is $\mathscr{K}_{\lambda_0}(\lambda) = \sqrt{\lambda+\nu/2} - \sqrt{\lambda_0+\nu/2}$. Now it is apparent, in view of Proposition 1, how the conjecture (7) arises, even though the non-central chi-squared distribution is not an exponential family.

Figure 1 shows some examples of the approximation (7). Even for $\nu = 3$ (left-hand plots) the approximation is good near the null; and the approximations appear to improve with ν. This means that we can use the simple expression $\mathscr{K}_{\lambda_0}(\lambda) = \sqrt{\lambda+\nu/2} - \sqrt{\lambda_0+\nu/2}$ for the *Key* to carry out inference for λ as described in Section 1. Further, we know that the *Key* is a good approximation to the signed square root of the KLD between null and alternative hypothesized distributions, at least for a large neighborhood of λ_0.

While the above ideas are straightforward, we do not always have n independent observations on a chi-squared family; rather the non-central chi-squared distribution arises through a consideration of K groups, as described in the next subsection.

3.3 Between group sum of squares (for known variance)

For each group $k = 1,\ldots,K$ let $\mathbf{X}_k' = [X_{k1}, X_{k2}, \ldots, X_{k,n_k}]$ denote a sample of n_k observations, each with distribution $N(\mu_k, 1)$. Also assume the elements of $\mathbf{X}' = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$ are independent. Further introduce the total sample size $N = \sum_k n_k$, the sample proportions $q_k = n_k/N$, the kth sample mean $\bar{X}_k$, the overall sample mean $\bar{X} = \sum_k q_k \bar{X}_k$, its expectation $\mu = \sum_k q_k \mu_k$ and the parameter $\lambda = N\sum_k q_k(\mu_k - \mu)^2$. Then the between group sum of squares $Y = N\sum_k q_k(\bar{X}_k - \bar{X})^2 \sim \chi^2_\nu(\lambda)$, where $\nu = K-1$, see Section 22.1, [6]. The ratio $\theta = \lambda/N = \sum_k q_k(\mu_k - \mu)^2$ depends only on the *relative* sample sizes q_k, and measures the variability of the group means μ_k using a weighted sum of squared deviations from the weighted mean μ , with weights q_k.

Let the test statistic be $S = Y/N$. The transformation to evidence for $\theta > \theta_0$ is then $T = \sqrt{N}\left[\sqrt{S - \nu/(2N)} - \sqrt{\theta_0 + \nu/(2N)}\right]$. Further, introduce the parameter $r = \nu/N = (K-1)/N$; the mean and variance of S in this notation are $\mathrm{E}[S] = \theta + r$ and $\mathrm{Var}[S] = (4\theta + 2r)/N$. The expected evidence for $\theta \geq \theta_0$ become $\mathrm{E}_\theta[T] \doteq \sqrt{N}\,\mathscr{K}_{\theta_0,N}(\theta)$, with the *Key* given by

$$\mathscr{K}_{\theta_0,N}(\theta) \doteq \sqrt{\theta + r/2} - \sqrt{\theta_0 + r/2} - \frac{1}{2N\sqrt{\theta + r/2}} \ . \tag{11}$$

This shows that the expected evidence is monotone increasing in θ for $\theta > \theta_0$, and is approximately 0 at $\theta = \theta_0$. For fixed θ it grows with $\sqrt{N}$. Also, for fixed θ, if $r = (K-1)/N$ remains fixed with increasing N, the correction term becomes negligible and the *Key* is essentially the first two terms of (11). If $K = o(N)$ as $N \to \infty$, then $r \to 0$ and the Key approaches $\mathscr{K}_{\theta_0,+\infty}(\theta) = \sqrt{\theta} - \sqrt{\theta_0}$.

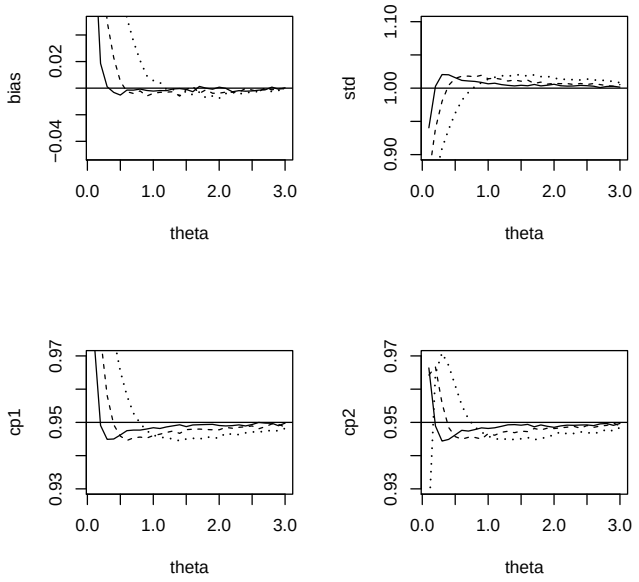

Fig. 2 In the first row of plots above are shown the empirical biases and standard deviation of T for $\nu = K-1 = 4$ degrees of freedom in Example 2 of Section 3. The results correspond to $N = 10$ (dotted line), $N = 20$ (dashed line), and $N = 40$ (solid line). The second row of plots gives the empirical coverage probabilities of nominal 95% upper confidence bounds and 95% confidence intervals.

Confidence intervals for the non-centrality parameter.

To obtain the confidence bounds of Equation (3) we need to solve for $\theta = \mathscr{K}^{-1}_{\theta_0,N}(u)$ Setting $c = -\sqrt{\theta_0 + r/2}$ we start with

$$u = \mathscr{K}_{\theta_0,N}(\theta) = \sqrt{\theta + r/2} + c - \frac{1}{2N\sqrt{\theta + r/2}} \,. \tag{12}$$

Solving this quadratic in θ yields

$$\mathscr{K}^{-1}_{\theta_0,N}(u) = \frac{1}{2}\left[\frac{1}{N} + \{(u-c)^2 - r\} + \left\{(u-c)^4 + \frac{2(u-c)^2}{N}\right\}^{1/2}\right] . \tag{13}$$

Evaluating this function at $u_\pm = (T \pm z_{0.975})/\sqrt{N}$, for $T = \sqrt{N}\left[\sqrt{S - \nu/(2N)} + c\right]$ yields the 95% confidence interval for θ in terms of the test statistic $S = Y/N = \sum_k q_k(\bar{X}_k - \bar{X})^2$. For convenience we note that $u_\pm - c = \sqrt{S - r/2} \pm z_{0.975}/\sqrt{N}$.

The performance of T and confidence intervals for θ based on it were examined by generating 100,000 simulations of $Y = NS \sim \chi^2_\nu(\lambda)$ for various choices of $\nu = K-1$ and N, and then computing the average bias $T - \sqrt{N}\mathscr{K}_{\theta_0,N}(\theta)$ (which is free of θ_0), the average standard deviation SD$[T]$, the one-sided 95% confidence bound empirical coverage, and finally the two-sided 95% confidence interval empirical coverage probabilities. These results are plotted as a function of θ over the range $[0,3]$ in Figure 2.

In the above derivation of confidence intervals we included a bias term in the *Key* to see if the resulting confidence intervals had better coverage than when we used the simpler the simpler *Key* $\mathscr{K}_{\theta_0}(\theta) = \sqrt{\theta + r/2} - \sqrt{\theta_0 + r/2}$. However, one only loses a little in accuracy of coverage probabilities and the derivation of the confidence interval is much quicker by the standard method K3 of Section 1.2.

4 Conclusions and further research problems

We have shown that it is often possible and practical to define an evidence T in favor of alternatives. This statistic is based on the idea of variance stabilization and the mean function of this evidence is closely related to the Kullback-Leibler divergence (KLD). Investigating the generality of this result merits further research.

In general, it may be said that the KLD gives insights into a variety of inferential questions and deserves renewed attention by statisticians. In the following we give two other examples that show the power of the KLD in revealing underlying structure. When the densities to be compared are $f_i(x) = f(x/\sigma_i)/\sigma_i$, one has $\mathrm{KLD}(\sigma_1,\sigma_2) = \mathrm{KLD}(1,\sigma_2/\sigma_1)$ — the ratio of the scales is the essential parameter. To be more precise, we have to compute the KLD. If the underlying density is normal, one obtains $\mathrm{KLD}(\sigma_1,\sigma_2) = \frac{1}{2}(\sigma_2/\sigma_1 - \sigma_1/\sigma_2)^2$. Reparametrizing to $\sigma_2 = (1+\Delta)\sigma_1$, we have $\frac{1}{2}(1+\Delta-1/(1+\Delta))^2$ for the value of the KLD. This expands into $\frac{1}{2}\left(1+\Delta-[1-\Delta+\Delta^2+O(\Delta^3)]\right)^2 = \frac{1}{2}\left(2\Delta-\Delta^2+O(\Delta^3)\right)^2$. The square root for $\Delta > 0$ leads to the *Key* $(\sqrt{2}\Delta - \Delta^2/\sqrt{2} + O(\Delta^3)$, which is up to this order the same as $\sqrt{2}\log(1+\Delta) = \sqrt{2}(\log(\sigma_2) - \log(\sigma_1))$. Thus, the transformed parameter obtained through the signed root of the KLD is simply the logarithm and furthermore, the test statistic is based on the difference. This is, of course, simply related to the fact that if the observed random variable $Y = \sigma X_0$, then $\log(Y) = \log(X_0) + \log(\sigma)$, which transforms the model into location-form.

Another example concerns robust, heavy-tailed models. When comparing $f_i(x) = f((x-\mu_i)/\sigma_0)/\sigma_0$, it is easy to show that the KLD only depends on $\delta = (\mu_2 - \mu_1)/\sigma$. As we have seen in our prototypical example, the KLD has value δ^2 for the normal shift model. What happens, if one moves to a heavy-tailed density? Figure 3 shows the case of the Cauchy density. It turns out that the amount of information available for small δ remains linear in δ and a loss of information only occurs for large values. Thus, with appropriate estimators of δ, no loss of information due to heavy-tails occurs. The loss is only due to the difficulty in estimating δ. As robust

theory shows, it is possible to construct compromise estimators that exploit this underlying information successfully for a wide range of tail behaviors. A similar loss of information for large values of δ occurs in the central Student-t model with unknown scale σ and a smallish number of degrees of freedom.

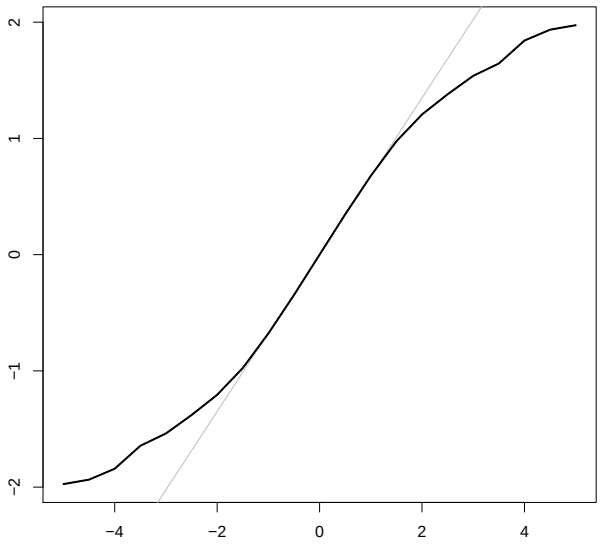

Fig. 3 The dark curve shows the signed root of the KLD for two standard Cauchy densities with a translational shift between them. The values were computed by Monte Carlo simulation. The value of the shift is indicated on the x-axis. The grey line has a slope equal to the ratio of the normal upper quartile divided by the Cauchy upper quartile, which can serve as an estimator of the scale change when switching the standard normal to the standard Cauchy density. For small shifts, there is but a tiny difference between the straight line and the root of the Cauchy KLD. For large shifts, the Cauchy KLD grows at a slower pace and turns out to be sub-linear.

Even though we have only considered cases, where the underlying parameter takes real values, extensions to multidimensional parameters are possible and this problem is open to further investigation. It would also be of interest to consider examples where the evidence is multidimensional.

References

1. Becker, B.J. (1997). P-values, Combination of. In: Kotz, S. (eds.) Encyclopedia of Statistical Sciences, Update Volume 1, pp. 448–453. Wiley, New York.
2. Bickel, P.J. and Doksum, K.A. (1977) Mathematical Statistics: Basic ideas and selected topics. Holden–Day, San Francisco.
3. Efron, B. (1982). Transformation Theory: How Normal is a Family of Distributions? The Annals of Statistics. **10**, 323–339.
4. Evans, M. (2000). Comment: On the probability of observing misleading statistical evidence. Journal of the American Statistical Association. **95**, 768–769.
5. Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. **10**, 507–521.
6. Kulinskaya, E., Morgenthaler, S. and Staudte, R.G. (2008). Meta Analysis: a Guide to Calibrating and Combining Statistical Evidence. John Wiley & Sons, West Sussex, England.
7. Kullback, S. (1968). Information theory and statistics. Dover, Mineola, New York.
8. Malzahn, U. and Bohning, D. and Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardized difference used in meta-analysis. Biometrika. **87**, 619–632.
9. Morgenthaler, S., Staudte, R.G. (2012). Advantages of Variance Stabilization. Scandinavian Journal of Statistics, doi: 10.1111/j.1467-9469.2011.00768.x.
10. Royall, R. (2000). On the probability of observing misleading statistical evidence. Journal of the American Statistical Association. **95**, 760–768.
11. Severini, T.A. (2000). Likelihood Methods in Statistics. Oxford University Press.
12. Thompson, S.G. (1998). Meta Analysis of Clinical Trials. In: Armitage, P. and Colton, T. (eds.) Encyclopedia of Biotatistics, pp. 2570–2579. Wiley, London.
13. Wellmann, J. (2012). Meta-analysis of trials with binary outcomes. In: Festschrift für Ursula Gather, Springer, Berlin.