International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 11(Special Issue), 2011, pp 59-74
© 2011 Dept. of Statistics, Univ. of Rajshahi, Bangladesh

On Some Aspects of Inference about Effect Sizes

Guido Knapp
Department of Statistics
TU Dortmund University
44221 Dortmund, Germany

Dihua Xu
Department of Mathematics and Statistics
University of Maryland, Baltimore County
Baltimore, MD 21250, USA

Philip L. H. Yu
Department of Statistics and Actuarial Science
University of Hong Kong
Hong Kong, China

[Received December 16, 2010; Revised February 3, 2011; Accepted March 1, 2011]

Abstract

We consider the problem of testing the null hypothesis of equality of several normal means when the variances while sharing the same functional forms are distinct functions of the means. This formulation arises in the context of using the standardized mean difference as a measure of effect size. Under the same setup, we also address the problem of drawing inference about the common normal mean when the null hypothesis holds. Both exact and approximate solutions as well as a Bayesian solution are developed. Applications are indicated in the context of effect size estimates with two data sets.

Keywords and Phrases: Effect size, Latent variable, Likelihood ratio test, Markov Chain Monte Carlo.

AMS Classification: Primary 62F03; Secondary 62F15.

1 Introduction

In statistical meta-analysis problems, the notion of effect size is very basic and it is often necessary to perform appropriate tests for the equality of population effect sizes arising out of several studies before performing a meta-analysis or pooling of evidence or data synthesis (Hedges and Olkin, 1985; Hartung, Knapp, and Sinha, 2008). Indeed, a next reasonable step after the homogeneity hypothesis of the equality of population effect sizes is accepted is to draw suitable inference about the common effect size. When the population effect size (θ) is based on standardized mean difference with respect to a control and a treatment, namely

$$\theta = (\mu_1 - \mu_2)/\sigma, \tag{1}$$

where μ_1, μ_2 and σ^2 denote, respectively, the control mean, the treatment mean and the (assumed) common variance, information about θ is typically obtained from random samples from the two populations. Thus, if $\bar{X}_1$, $\bar{X}_2$ and S^2 denote, respectively, the two sample means and the pooled sample variance based on a sample of size n_1 from control population and size n_2 from the treatment population, θ is routinely estimated by (Cohen's d, Glass's Δ, Hedges's g)

$$\hat{\theta} = (\bar{X}_1 - \bar{X}_2)/S_{pooled} \tag{2}$$

Under the assumption of normality and independence of the two samples, the large sample distribution of $\hat{\theta}$ can be approximated by

$$\hat{\theta} \sim N[\theta, a^2(b^2 + \theta^2)] \tag{3}$$

where a and b are functions of the sample sizes n_1 and n_2. Estimates of θ of the form $\hat{\theta}$ given above along with some slight variations are popularly known as Cohen's d, Glass's Δ and Hedges's g. For expressions for a and b and further details about effect sizes, we refer to the excellent texts by Hedges and Olkin (1985) and Hartung, Knapp, and Sinha (2008). Statistical inference associated with equation (3) above is the focus of this paper.

Assume that there are k independent studies dealing with the same control and treatment, and all are targeted towards the same common goal of providing information about the basic effect size θ. This means that several studies are independently performed to compare the *same* pair of control and treatment effects in order to provide information about the presumably common effect size θ. Of course, a priori, we cannot assume that the population effects sizes, say $\theta_1, \cdots, \theta_k$, arising out of the k studies are the same. A standard approach in statistical meta-analysis is two-fold:

1. Test $H_0 : \theta_1 = \cdots = \theta_k$ versus $H_1 : \theta_i$'s unequal
2. Assuming H_0 holds, draw suitable inference about the common population effect size θ.

To solve the above problems in a natural way, independent estimates of the effect size parameters are derived from the underlying k studies, resulting in

$$\hat{\theta}_i \sim N[\theta_i, a_i^2(b_i^2 + \theta_i^2)], i = 1, \cdots, k. \tag{4}$$

The important point to note here is that usually the constants a_i's and b_i's will vary with the studies, sometimes quite wildly. For example, when Cohen's d is used as $\hat{\theta}$, $a_i^2 = (n_1+n_2)/[2(n_1+n_2-2)^2]$ and $b_i^2 = [2(n_1+n_2)(n_1+n_2-2)][n_1 n_2]$, and there is no reason to believe that these design constants would remain the same across studies. In fact, in one of the examples (example 2) analyzed in this paper, these constants vary substantially.

We address the first problem of testing the equality of the population effect sizes in Section 2, and the second problem of drawing appropriate inference about the assumed common effect size in Section 3 where we have provided both the frequentist and the Bayesian solutions. Two illustrative examples are worked out to explain the suggested methods.

We conclude this section with the observation that when the distribution of $\hat{\theta}$ defined in (3) is taken as non-central t rather than normal, the corresponding inference problem is related to the non-centrality parameter of the t distribution, the non-centrality parameter being a multiple of θ. Under the condition of equality of the multipliers of θ_i's arising out of k studies, which makes the equality of θ_i's equivalent to the equality of non-centrality parameters, the first problem (test of H_0) has been discussed in Miwa (1994, 1996) for $k = 2$, and in Nagata et al. (2003) for a general k.

2 Test of $H_0 : \theta_1 = \cdots = \theta_k$ versus $H_1 : \theta_i$'s unequal

In this section we propose several tests of H_0 based on the model (4). We should mention that when b_i's are all equal, a simple variance-stabilizing transformation can be used to easily derive a chi-square test of H_0 (Hedges and Olkin, 1985).

2.1 Likelihood ratio test and modified likelihood ratio test

Here we describe the likelihood ratio test (LRT). Under H_0, denoting the common effect size by θ, it is easy to verify that the maximum likelihood estimate (MLE) of θ is obtained by minimizing the expression $Q_0(\theta)$ given by

$$Q_0(\theta) = \sum_{i=1}^{k} \frac{(\hat{\theta}_i - \theta)^2}{a_i^2(b_i^2 + \theta^2)} + \sum_{i=1}^{k} \ln(b_i^2 + \theta^2). \tag{5}$$

The unrestricted maximum likelihood estimate of θ_i, $i = 1, \cdots, k$ is obtained by minimizing the expression $Q_1(\theta_i)$ given by

$$Q_1(\theta_i) = \frac{(\hat{\theta}_i - \theta_i)^2}{a_i^2(b_i^2 + \theta_i^2)} + \ln(b_i^2 + \theta_i^2). \tag{6}$$

It is tempting to infer that the unrestricted MLE of θ_i is $\hat{\theta}_i$! However, this is *not* the case although they can be quite close (see the applications in a later section). An explicit solution of (5) and even (6) is difficult to obtain, however it is indeed possible to provide a numerical solution once we plug in the data, namely, values of $\hat{\theta}_i$'s and the (design) constants a_i's and b_i's. Consequently, it is indeed possible to numerically compute the value of the LRT statistic. One can also use $\hat{\theta}_i$ as the approximate unrestricted MLE of θ_i, and compute the resultant LRT, which we will call *modified* LRT. Obviously, the exact null distribution of the LRT (or modified LRT) statistic would be quite complicated, and one may take recourse to simulation to approximate the cut-off point, and hence carry out the test. The null distribution of the LRT (or modified LRT) being dependent on the unknown common ES θ, an extensive simulation may be necessary. To study the power of the LRT, one again has to depend entirely on simulation. Some limited simulation results are reported in this paper in order to compare the performance of the LRT and modified LRT with those of the other tests suggested below.

2.2 A new test

In this subsection we provide a new test of H_0. The key idea here is to express the underlying model (3) as a marginal distribution of a suitable joint distribution via a *latent* variable X. This is done by introducing the conditional and marginal models as

$$\hat{\theta}|X \sim N[\theta(1 + aX), a^2b^2], \;\; X \sim N[0, 1]. \tag{7}$$

Referring to (4), with the introduction of k independent latent variables $X_1, \cdots, X_k$, we now derive a test of H_0 based on $(\hat{\theta}_1, \cdots, \hat{\theta}_k)$, conditional on $\mathbf{X} = (X_1, \cdots, X_k)$ as follows. Recall that our model is now

$$\hat{\theta}_i|X_i \sim N[\theta_i(1 + a_iX_i), a_i^2b_i^2], \;\; X_i \sim N[0, 1], \;\; i = 1, \cdots, k. \tag{8}$$

Hence, conditionally given $\mathbf{X}$, since a_i's and b_i's are known positive constants, the test of H_0, namely, the equality of θ_i's is fairly routine, leading to the familiar chi-square test based on

$$\chi^2(\mathbf{X}) = \sum_{i=1}^{k} \Big[\hat{\theta}_i - (1 + a_iX_i)\frac{\sum_{j=1}^{k} \frac{\hat{\theta}_j(1+a_jX_j)}{a_j^2b_j^2}}{\sum_{j=1}^{k} \frac{(1+a_jX_j)^2}{a_j^2b_j^2}}\Big]^2 \Big/ a_i^2b_i^2 \tag{9}$$

which can be simplified as

$$\chi^2(\mathbf{X}) = \sum_{i=1}^{k} \frac{\hat{\theta}_i^2}{a_i^2 b_i^2} - \frac{\left[\sum_{i=1}^{k} \frac{\hat{\theta}_i(1+a_i X_i)}{a_i^2 b_i^2}\right]^2}{\sum_{i=1}^{k} \frac{(1+a_i X_i)^2}{a_i^2 b_i^2}}. \tag{10}$$

It is well known that $\chi^2(\mathbf{X})$ has a central χ^2 distribution with $(k-1)$ $d.f.$ under H_0, for any fixed $\mathbf{X}$. Since the null conditional distribution of $\chi^2(\mathbf{X})$ is independent of $\mathbf{X}$, this is also the unconditional distribution of $\chi^2(\mathbf{X})$, irrespective of the marginal distribution of $\mathbf{X}$. One can then reject H_0 at level α when $\chi^2(\mathbf{X})$ exceeds $\chi^2_{\alpha;k-1}$ for *any* fixed vector $\mathbf{X}$, thus providing an exact test of H_0!

One wonders what would be some meaningful choices of $\mathbf{X}$ leading to some interesting tests of H_0. Rather than concentrating on a single value of $\mathbf{X}$, we propose just one test of H_0 by the following algorithm:

1. Generate independent $X_i \sim N[0,1]$, $i = 1, \cdots, k$.
2. Compute $\chi^2(X)$.
3. Repeat step 1 and 2 m times leading to $\chi^2(X)_1, \chi^2(X)_2, \cdots, \chi^2(X)_m$.
4. Take the average: $\bar{\chi}^2(X) = \sum_{i=1}^{m} \chi^2(X)_i/m$.

Of course, the *null* distribution of this test statistic is not central χ^2 any more, and is, in fact, quite complicated. To estimate the null distribution, we use the parametric bootstrap approach using the MLE under H_0 to generate the data according to model (4) under H_0 and then compute $\bar{\chi}^2(X)$. We reject H_0 at level α, if the observed value of the test statistic is larger than the $100(1-\alpha)\%$-bootstrap-quantile of $\bar{\chi}^2(X)$.

We note in passing that quite interestingly,

$$X_i|\hat{\theta}_i \sim N[\frac{\theta(\hat{\theta}_i - \theta)}{a_i(b_i^2 + \theta^2)}, \frac{b_i^2}{b_i^2 + \theta^2}]. \tag{11}$$

Remark. In some applications it may happen that $b_1, \cdots, b_k$, arising out of the k studies, are all equal $(= b)$, reducing our basic model (4) to

$$\hat{\theta}_i \sim N[\theta_i, a_i^2(b^2 + \theta_i^2)], \quad i = 1, \cdots, k \tag{12}$$

Since $a_i^2 = O(n_i^{-1})$, following a standard variance-stabilizing argument, we can work with the modified effect sizes and their asymptotic distributions given by

$$\hat{\theta}_i^* = \ln(\hat{\theta}_i + \sqrt{\hat{\theta}_i^2 + b^2}) \sim N[\ln(\theta_i + \sqrt{\theta_i^2 + b^2}), a_i^2]. \tag{13}$$

It is then obvious that a test for $H_0 : \theta_1 = \cdots = \theta_k$, which is equivalent to $H_0^* : \tilde{\theta}_1 = \cdots = \tilde{\theta}_k$, $\tilde{\theta}_i = \ln(\theta_i + \sqrt{\theta_i^2 + b^2})$, $i = 1, \cdots, k$, is obtained by using the test statistic

$$\chi_0^2 = \sum_{i=1}^{k} (\hat{\theta}_i^* - \bar{\theta}^*)^2 / a_i^2, \;\; \bar{\theta}^* = \frac{\sum_{i=1}^{k} \hat{\theta}_i^* / a_i^2}{\sum_{i=1}^{k} 1/a_i^2}, \tag{14}$$

the statistic χ_0^2 having a null asymptotic chi-square distribution with $(k-1)$ df.

2.3 Applications

We have proposed several tests of the homogeneity hypothesis H_0 in the previous section. In this subsection, we apply these statistical procedures to two well known data sets borrowed from Abrams and Sanso (1998), and Kirsch et al. (2008). All throughout, we have used Hedges's estimate g as the effect size measure, and the resultant values of a_i and b_i.

We recall that Hedges's g (Hartung, Knapp, and Sinha, 2008) is defined as

$$g = \frac{\bar{X}_1 - \bar{X}_2}{S^*},$$

where the standardized quantity S^* is also the pooled sample standard deviation defined as $S^* = \sqrt{S^{*2}}$ with

$$S^{*2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

The variances of the estimate of θ, in large samples, is given by the following:

$$\sigma^2(g) = \text{var}(g) \approx \frac{n_1 + n_2}{n_1 n_2} + \frac{\theta^2}{2(n_1 + n_2 - 2)}.$$

Therefore,

$$a^2 = \frac{1}{2(n_1 + n_2 - 2)}, \;\; b^2 = \frac{2(n_1 + n_2)(n_1 + n_2 - 2)}{n_1 n_2}.$$

2.3.1 Dentifrice Data

The data set is taken from Abrams and Sanso (1998) and concerns a previously published meta-analysis which was conducted of all randomized controlled trials comparing sodium monofluorophosphate (SMFP) to sodium fluoride (NaF) dentifrices (toothpastes) in the prevention of caries; see Johnson (1993). The outcome in each trial was the change from baseline in the decayed missing (due to caries) filled surface (DMFS) dental index at three years follow-up. Of 12 studies identified as meeting the inclusion criteria, 9 considered a straight comparison of NaF and SMFP. Table 1 displays the data from these 9 studies in terms of mean change in DMFS index for each treatment.

Table 1: Randomized evidence comparing sodium fluoride (NaF) with sodium monofluorophosphate (SMFP) dentrifices in terms of differences from baseline in DMFS dental index

Study	N	NaF Mean	SD	N	SMFP Mean	SD
1	134	5.96	4.24	113	6.82	4.72
2	175	4.74	4.64	151	5.07	5.38
3	137	2.04	2.59	140	2.51	3.22
4	184	2.70	2.32	179	3.20	2.46
5	174	6.09	4.86	169	5.81	5.14
6	754	4.72	5.33	736	4.76	5.29
7	209	10.10	8.10	209	10.90	7.90
8	1151	2.82	3.05	1122	3.01	3.32
9	679	3.88	4.85	673	4.37	5.37

In Table 2 we display the values of $\hat{\theta}_i$'s, a_i's, b_i's and the unrestricted MLEs of θ_i's. Note that the design constant b_i's are close together so that the test statistic χ_0^2 from (14) may be a valid alternative test.

Table 3 shows the values of exact LRT, modified LRT, and the two suggested test statistics $\bar{\chi}^2(X)$ and χ_0^2. The test statistic $\bar{\chi}^2(X)$ is based on the average of $m = 1,000$ individual test statistics $\bar{\chi}^2(X)$ and the p-value is calculated according to the parametric bootstrap approach outlined in Section 2.2, where the bootstrap sample size is $t = 300$. In the test statistic χ_0^2, we replace the design constant b by the mean of the b_i^2's. It is interesting to observe that all tests lead to the acceptance of H_0. Note that the MLE of θ under H_0 is -0.0688.

Table 2: Values of design constants of dentifrice data

Study	$\hat{\theta}$	a^2	b^2	$\hat{\theta}_{mle}$
1	-0.1926	0.0020	7.9930	-0.1922
2	-0.0660	0.0015	7.9942	-0.0659
3	-0.1607	0.0018	7.9432	-0.1604
4	-0.2092	0.0014	7.9574	-0.2089
5	0.0560	0.0015	7.9550	0.0559
6	-0.0075	0.0003	7.9904	-0.0075
7	-0.1000	0.0012	7.9617	-0.0999
8	-0.0596	0.0002	7.9943	-0.0596
9	-0.0958	0.0004	7.9883	-0.0958

Table 3: Test results of dentifrice data

Tests	Test Statistics	P-value	Conclusion
LRT	6.4259	0.5950	Accept H_0
mLRT	6.4259	0.5950	Accept H_0
$\bar{\chi}^2(X)$	6.4432	0.5950	Accept H_0
χ_0^2	6.4282	0.5994	Accept H_0

2.3.2 Antidepressant Data

The data set used here were obtained by Kirsch et al. (2008) from FDA following Freedom of Information Act and deal with changes (drug-placebo differences) in the severity of depression in very severe depressant patients. Kirsch et al. (2008) carried out a meta-analysis in order to study the relationship between baseline severity and antidepressant efficacy. Our focus here is to assess the homogeneity of the population effect sizes. The relevant data from 35 studies for our purpose are given in Table 4. For each study, we have the mean change from baseline $\bar{X}_i$, $i = D, P$, the standardized effect size d_i, $i = D, P$, calculated as $d_i = \bar{X}_i/sd_i$ and sd_i is the standard deviation in the ith group, as well as the sample size N_i in each group. We readily obtain $sd_i = \bar{X}_i/d_i$, $i = D, P$ and can then easily compute Hedges's g for the comparison of drug and placebo group. The estimated effect sizes for each study $\hat{\theta}_i = g_i$ are given in Table 5, where also the values of the design constants a^2 and b^2 as well as the unrestricted MLE's of θ_i can be found.

In Table 6, the results of the homogeneity tests are given. The test statistic and p-value of $\bar{\chi}^2(X)$ is calculated following the same lines as in the previous example. Just for illustrative purpose, we also give the result for χ_0^2, replacing again b by the mean of the b_i^2's, though the design constants b_i^2 substantially vary. Again, all four tests come to the same conclusion, this time to reject the null hypothesis. Note that the MLE of θ under H_0 is -0.3141 in this example.

2.4 Some simulation results

In a small simulation study, we investigate the properties of the tests discussed in Section 2.1 and 2.2. We consider the LRT , mLRT, $\bar{\chi}^2(X)$, and χ_0^2 tests like in the examples above. Additionally, we also include the parametric bootstrap versions of both the likelihood ratio tests, denoted by LRTb and mLRTn, following the same ideas as described in Section 2.2 for the $\bar{\chi}^2(X)$ test. Since we are confident that the tests will work well if the design constants b_i^2's are less variable like in the first example in Section 2.3, we concentrate on sample size designs with highly variable constants b_i^2's.

In scenario 1, we consider five experiments with sample sizes $(n_{1i}, n_{2i}) = (40, 160)$,

Table 4: Studies of antidepressant medications

ID	Drug			Placebo			ID	Drug			Placebo		
	X_D	d_D	N_D	X_P	d_P	N_P		X_D	d_D	N_D	X_P	d_P	N_P
1	12.50	1.44	22	5.50	0.63	24	19	10.00	1.34	80	8.90	1.20	78
2	7.20	0.83	18	8.80	1.03	24	20	13.50	1.67	24	10.50	1.30	24
3	11.00	1.15	181	8.40	0.88	163	21	12.30	1.28	51	6.80	0.70	53
4	5.89	1.02	299	5.82	1.05	56	22	10.90	1.23	36	5.80	0.66	34
5	8.82	1.13	297	5.69	0.72	48	23	9.70	0.93	33	7.20	0.69	33
6	11.20	1.37	231	6.70	0.82	92	24	12.70	1.87	36	7.60	1.12	38
7	13.90	1.77	64	9.45	1.20	78	25	10.80	1.60	40	4.70	0.69	38
8	11.90	1.16	65	8.88	0.87	75	26	8.00	1.14	40	6.20	0.88	40
9	10.10	1.27	69	9.89	1.24	79	27	9.90	1.18	41	10.00	1.19	42
10	11.00	1.34	227	9.49	1.15	75	28	10.40	1.33	37	6.70	0.86	37
11	14.20	1.45	46	4.80	0.43	47	29	10.00	0.99	40	4.10	0.41	42
12	9.57	1.15	101	8.00	0.92	52	30	9.10	1.11	39	3.00	0.37	37
13	8.90	1.17	153	8.90	1.17	77	31	9.10	1.28	403	8.20	1.14	51
14	11.40	1.41	156	9.50	1.17	75	32	6.00	0.97	19	6.20	0.83	22
15	10.00	1.31	74	9.84	1.27	70	33	9.10	1.23	19	6.70	0.86	10
16	12.30	1.42	175	9.80	1.11	47	34	8.80	0.80	20	4.50	0.49	21
17	10.80	1.36	57	8.20	1.03	57	35	13.10	1.20	13	10.90	0.99	12
18	12.00	1.51	86	8.00	1.01	90							

$(70, 130)$, $(100, 100)$, $(130, 70)$, and $(160, 40)$ leading to the design constants $b_i^2 =$ 12.375, 8.7033, 7.92, 8.7033, and 12.375.

In scenario 2, we consider ten experiments with sample sizes $(n_{1i}, n_{2i}) = (42, 57)$, $(45, 57)$, $(33, 33)$, $(43, 48)$, $(57, 43)$, $(37, 67)$, $(53, 52)$, $(34, 34)$, $(26, 60)$, and $(61, 62)$ leading to the design constants $b_i^2 = 8.0226$, 7.9532, 7.7576, 7.8479, 7.9967, 8.5583, 7.8483, 7.7647, 9.2615, and 7.8704.

In scenario 3, we consider 25 experiments with sample sizes $(n_{1i}, n_{2i}) = (16, 14)$, $(26, 25)$, $(27, 15)$, $(30, 30)$, $(14, 10)$, $(18, 30)$, $(24, 30)$, $(29, 12)$, $(26, 17)$, $(26, 25)$, $(18, 10)$, $(11, 22)$, $(19, 23)$, $(19, 23)$, $(25, 21)$, $(15, 25)$, $(30, 14)$, $(28, 30)$, $(20, 10)$, $(28, 22)$, $(21, 15)$, $(29, 23)$, $(10, 10)$, $(14, 21)$, and $(28, 27)$ leading to design constants b_i^2 between 7.2 and 9.1897

Note that in all the three scenarios, the total sample size is $N = \sum n_{1i} + n_{2i} = 1000$.

In the three scenarios, the true common effect sizes are chosen as $\theta = 0$ and $\theta = 1$. The first parameter value reflects no difference between the two groups, while the second on stands for strong superiority of the treatment group compared to the control group.

Each estimated actual size of the six homogeneity tests is based on 10,000 simulation runs. The nominal level of the tests is $\alpha = 0.05$. The results for the estimated

Table 5: Values of design constants of antidepressant medications data

ID	$\hat{\theta}$	a^2	b^2	$\hat{\theta}_{mle}$	ID	$\hat{\theta}$	a^2	b^2	$\hat{\theta}_{mle}$
1	-0.8040	0.0114	7.6667	-0.7950	19	-0.1478	0.0032	7.9000	-0.1474
2	0.1861	0.0125	7.7778	0.1838	20	-0.3713	0.0109	7.6667	-0.3673
3	-0.2721	0.0015	7.9753	-0.2717	21	-0.5692	0.0049	7.8491	-0.5664
4	-0.0122	0.0014	14.9683	-0.0122	22	-0.5778	0.0074	7.7778	-0.5736
5	-0.4003	0.0015	16.6014	-0.3997	23	-0.2396	0.0078	7.7576	-0.2378
6	-0.5505	0.0016	9.7575	-0.5497	24	-0.7513	0.0069	7.7895	-0.7461
7	-0.5658	0.0036	7.9647	-0.5638	25	-0.8997	0.0066	7.8000	-0.8938
8	-0.2952	0.0036	7.9262	-0.2941	26	-0.2560	0.0064	7.8000	-0.2544
9	-0.0264	0.0034	7.9281	-0.0263	27	0.0119	0.0062	7.8084	0.0118
10	-0.1837	0.0017	10.6432	-0.1834	28	-0.4740	0.0069	7.7838	-0.4708
11	-0.8946	0.0055	7.8289	-0.8897	29	-0.5871	0.0062	7.8095	-0.5835
12	-0.1858	0.0033	8.7978	-0.1852	30	-0.7481	0.0068	7.7949	-0.7430
13	0.0000	0.0022	8.9025	-0.0000	31	-0.1264	0.0011	19.9687	-0.1263
14	-0.2347	0.0022	9.0426	-0.2342	32	0.0290	0.0128	7.6507	0.0286
15	-0.0208	0.0035	7.8950	-0.0207	33	-0.3187	0.0185	8.2421	-0.3129
16	-0.2875	0.0023	11.8760	-0.2868	34	-0.4253	0.0128	7.6143	-0.4200
17	-0.3270	0.0045	7.8596	-0.3255	35	-0.2007	0.0217	7.3718	-0.1964
18	-0.5042	0.0029	7.9132	-0.5027					

actual sizes are displayed in Table 7.

Generally, all the six homogeneity tests maintain the nominal level. The actual size of LRT and mLRT are nearly identical so that the effort of calculating the unrestricted MLE does not yield a real advantage. Thus, the use of mLRT can be recommended. LRT and mLRT maintain the nominal level quite well, the bootstrap versions of both test tend to be a little bit liberal. The $\bar{\chi}^2(X)$ behave similar to the bootstrap likelihood ratio test.

Despite the variability of the design constants b_i^2 in all three scenarios, the test χ_0^2 yields surprisingly good results; the estimated actual size of χ_0^2 is close to the estimated

Table 6: Test results of antidepressant medications data

Tests	Test Statistics	P-value	Conclusion
LRT	62.5495	0.0017	Reject H_0
mLRT	62.5442	0.0017	Reject H_0
$\bar{\chi}^2(X)$	64.0876	0.0016	Reject H_0
χ_0^2	61.5898	0.0026	Reject H_0

Table 7: Estimated actual sizes of six homogeneity tests given a nominal level of $\alpha = 0.05$

Scenario	θ	LRT	LRTb	mLRT	mLRTb	$\bar{\chi}^2(X)$	χ_0^2
1	0	0.0457	0.0491	0.0457	0.0491	0.0487	0.0478
	1	0.0502	0.0545	0.0502	0.0545	0.0547	0.0523
2	0	0.0483	0.0526	0.0483	0.0526	0.0523	0.0474
	1	0.0482	0.0537	0.0484	0.0537	0.0535	0.0481
3	0	0.0472	0.0566	0.0471	0.0566	0.0565	0.0449
	1	0.0460	0.0530	0.0456	0.0530	0.0520	0.0453

actual sizes of LRT and mLRT. Note that under model (13), the test based on χ_0^2 is the most powerful invariant test (Lehmann, 1986). It turns out that even when the assumption of model (13) is not strictly satisfied, the χ^2 test based on this model performs rather well.

There is no obvious dependence of the size of the test with the true underlying common efect size. In scenario 2 and 3, the differences can be clearly explained via Monte Carlo error. The largest differences occur in scenario 1, the scenario with the largest range of design constants b_i^2.

Further results of our simulation study reveal that the power of the six homogeneity tests are rather similar. We omit the details here.

3 Inference about common effect size θ

In this section we discuss statistical inference about the common effect size θ, assuming that H_0 holds, based on the independent effect size estimates $\hat{\theta}_i$, distributed as

$$\hat{\theta}_i \sim N[\theta, a_i^2(b_i^2 + \theta^2)], \;\; i = 1, \cdots, k. \tag{15}$$

3.1 Frequentist solution

As a point estimate of θ, we propose the MLE $\hat{\theta}_{mle}(H_0)$ which has been described earlier. Of course, one can also use the simple weighted unbiased estimate $\hat{\theta}_w$ and also the asymptotically unbiased estimate $\tilde{\theta}_w$ defined as

$$\hat{\theta}_w = \frac{\sum_{j=1}^k \frac{\hat{\theta}_j}{a_j^2 b_j^2}}{\sum_{j=1}^k \frac{1}{a_j^2 b_j^2}}. \tag{16}$$

and

$$\tilde{\theta}_w = \frac{\sum_{i=1}^{k} \frac{\hat{\theta}_i}{a_i^2(b_i^2+\hat{\theta}_i^2)}}{\sum_{i=1}^{k} \frac{1}{a_i^2(b_i^2+\hat{\theta}_i^2)}}. \tag{17}$$

The second estimate mentioned above is the traditional estimated variance inverse-weighted estimate of the common effect size (see Hartung, Knapp, and Sinha, 2008). To compare the three estimates asymptotically, we note that the asymptotic variance of the MLE of θ can be derived from the following calculations.

A. $\frac{\partial \ln L(\theta|\text{data})}{\partial \theta} = -\sum_{i=1}^{k} \frac{\theta}{b_i^2+\theta^2} + \sum_{i=1}^{k} \frac{(\hat{\theta}_i-\theta)}{(b_i^2+\theta^2)a_i^2} + \sum_{i=1}^{k} \frac{(\hat{\theta}_i-\theta)^2\theta}{(b_i^2+\theta^2)^2a_i^2}$

B. $\frac{\partial^2 \ln L(\theta|\text{data})}{\partial \theta^2} = -\sum_{i=1}^{k} \frac{1}{b_i^2+\theta^2} + \sum_{i=1}^{k} \frac{2\theta^2}{(b_i^2+\theta^2)^2} - \sum_{i=1}^{k} \frac{1}{(b_i^2+\theta^2)a_i^2}$
$+ \sum_{i=1}^{k} \frac{(\hat{\theta}_i-\theta)^2}{(b_i^2+\theta^2)^2a_i^2} - \sum_{i=1}^{k} \frac{4\theta^2(\hat{\theta}_i-\theta)^2}{(b_i^2+\theta^2)^3a_i^2}$ + other terms whose mean is 0

C. $E[-\frac{\partial^2 \ln L(\theta|\text{data})}{\partial \theta^2}] = \sum_{i=1}^{k} \frac{1}{(b_i^2+\theta^2)a_i^2} + \sum_{i=1}^{k} \frac{2\theta^2}{(b_i^2+\theta^2)^2}$ leading to

$$\text{var}(\hat{\theta}_{mle}) \sim \frac{1}{\sum_{i=1}^{k} \frac{1}{(b_i^2+\theta^2)a_i^2} + \sum_{i=1}^{k} \frac{2\theta^2}{(b_i^2+\theta^2)^2}} \tag{18}$$

On the other hand, a direct computation yields

$$\text{var}(\hat{\theta}_w) = \frac{1}{\sum_{j=1}^{k} \frac{1}{a_j^2 b_j^2}} + \theta^2 \times \frac{\sum_{j=1}^{k} \frac{1}{a_j^2 b_j^4}}{[\sum_{j=1}^{k} \frac{1}{a_j^2 b_j^2}]^2} \tag{19}$$

and

$$\text{var}(\tilde{\theta}_w) \sim \frac{1}{\sum_{j=1}^{k} \frac{1}{a_j^2(b_j^2+\theta^2)}} \tag{20}$$

Hence, it follows that the MLE of θ has a smaller asymptotic variance than $\tilde{\theta}_w$. A comparison between the MLE of θ and $\hat{\theta}_w$ naturally depends on the unknown value of the common effect size θ. It is easy to show that the relative efficiency of $\hat{\theta}_w$ relative to $\hat{\theta}_{mle}$ lies between the bounds

$$1 \leq RE \leq \frac{[\sum_{i=1}^{k} \frac{1}{a_i^2} + 2k][\sum_{i=1}^{k} \frac{1}{a_i^2 b_i^4}]}{[\sum_{i=1}^{k} \frac{1}{a_i^2 b_i^2}]^2} \tag{21}$$

Interestingly enough, the upper bound of RE is nearly 1 in most applications, implying that the traditional estimated variance inverse-weighted common effect size estimate which is readily computable can be used in practice.

Large sample tests for $H_0^{\dagger} : \theta = 0$, a common null hypothesis value in the context of standard meta-analysis, as well as large sample confidence intervals for θ can be based on $\hat{\theta}_{mle}(H_0)$, $\hat{\theta}_w$ and $\tilde{\theta}_w$ in the usual fashion, based on their standardized versions,

replacing θ in their asymptotic variances by their respective estimates. The LRT statistic for testing $H_0^\dagger : \theta = 0$ on the other hand can be readily computed as

$$\begin{aligned} -2\ln(LRT) &= \sum_{i=1}^{k} \frac{\hat{\theta}_i^2}{a_i^2 b_i^2} + \sum_{i=1}^{k} \ln(b_i^2) - \sum_{i=1}^{k} \frac{(\hat{\theta}_i - \hat{\theta}_{mle}(H_0))^2}{a_i^2(b_i^2 + \hat{\theta}_{mle}(H_0)^2)} \\ &\quad - \sum_{i=1}^{k} \ln(b_i^2 + \hat{\theta}_{mle}(H_0)^2), \end{aligned} \tag{22}$$

and the test rejects $H_0^\dagger$ when $-2\ln(LRT) \geq \chi^2_{1;\alpha}$.

Remark. When $b_1 = \cdots = b_k$, since testing $H_0^\dagger : \theta = 0$ is equivalent to testing $H_0^\star : \theta^* = g(\theta) = \ln b$, an optimum test is based on

$$z = (\bar{\theta}^* - \theta^*)\sqrt{\sum_{i=1}^{k} 1/a_i^2}$$

where $\bar{\theta}^*$ is defined in equation (14).

3.2 Bayesian solution

We now turn our attention to a Bayesian solution of the above inferential problem about θ. Assume we have a flat prior for the common effect size θ, which implies that (15) is also the posterior distribution of θ, given the data (sample effect size estimates). Inspite of the fact that this posterior distribution is rather complicated, that is,

$$\pi(\theta|\text{data}) \propto \prod_{i=1}^{k} \left(b_i^2 + \theta^2\right)^{-1/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{k} \frac{(\hat{\theta}_i - \theta)^2}{a_i^2(b_i^2 + \theta^2)}\right) \tag{23}$$

we can readily use the familiar MCMC algorithms implemented in the software packages WinBUGS or OpenBUGS to generate samples of θ values and hence compute the posterior mean and credible interval of θ. It may also be possible to use (8) and (11) recursively to generate posterior distribution of θ. However, our approach on (23) is direct. We present in the next section the results for the Dentifrice data used in Section 2.3.

3.3 An application

Table 8 provides the estimates and their associated 95% confidence intervals for the Dentifrice data for which the null hypothesis of a common effect size is accepted. The MLE estimate under H_0 as well as $\hat{\theta}_w$ and $\tilde{\theta}_w$ are nearly identical in this example as

well as the 95%-confidence intervals. In all three cases, zero is not included in the intervals. Inference under the Bayes solution is based on the posterior mean and the credible interval. The posterior distribution of θ is graphically displayed in Figure 1. The 95% credible interval does not include zero. The programming code for the Bayes analysis is provided in the Appendix.

Applying the LRT from (22) to the data set, we obtain the value of the test statistic as $LRT = 0.0150$ which leads to rejection of $H_0^\dagger$ since the cut-off point is $\exp(-\chi^2_{1;\alpha}/2) = 0.1489$.

For this data set, the b_i's being nearly equal, the mean of the b_i^2's is 7.9752 with standard deviation 0.0205. So, we have used the estimate based on the transformed version of $\hat{\theta}^*$ and the test based on z with constant $b^2 = 7.9752$. The estimate $\hat{\theta}^*$ is 1.014 leading to $z = -2.8976$ and two-sided p-value of 0.0038, again leading to rejection of $H_0^\dagger$.

Based on the above results, we conclude that there is a significant difference between the control and the treatment effects for Dentifrice.

Table 8: Estimates and 95% intervals in the Dentifrice example

$\hat{\theta}_{mle}(H_0)$	$\hat{\theta}_w$	$\tilde{\theta}_w$	Bayes
-0.0689	-0.0688	-0.0687	-0.0625
(-0.1154, -0.0223)	(-0.1155, -0.0223)	(-0.1154, -0.0222)	(-0.1141, -0.0103)

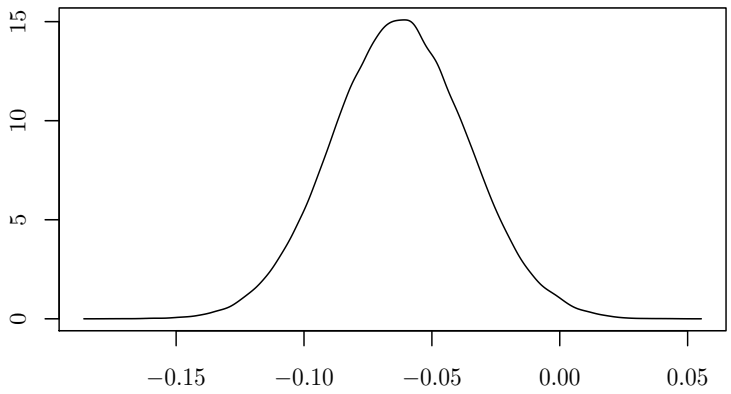

Figure 1: Posterior distribution of effect size θ in the Dentrifrice example

Acknowledgment

We would like to thank Prof. Bimal Sinha for his valuable comments during this research project.

References

[1] Abrams, K. and Sanso, B. (1998). Approximate Bayesian inference for random effects meta-analysis, *Statistics in Medicine*, **17**, 201-218.

[2] Hartung, J., Knapp, G., and Sinha, Bimal K. (2008). *Statistical Meta-Analysis with Applications*, Wiley, New York.

[3] Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*, Academic Press, Boston.

[4] Kirsch, I, Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., and Johnson, B. T. (2008). Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration, *PLOS medicine*, **5**, 0260-0268.

[5] Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2^{nd} Edition, Springer, New York.

[6] Li, Y., Shi, L., and Roth, H. D. (1994). The bias of the commonly used estimate of variance in meta-analysis, *Communications in Statistics—Theory and Methods*, **23**, 1063-1085.

[7] Miwa, T. (1994). Statistical inference on non-centrality parameters and Taguchi's SN ratios, *Proceedings on International Conference, Statistics in Industry, Science and Technology*, 66-71.

[8] Miwa, T. (1996). A variance-stabilizing transformation of non-central F distribution, *Proceedings of 18th Symposium, Japanese Society of Applied Statistics*, 81-85.

[9] Nagata, Y., Miyakawa, M., and Yokozawa, T. (2003). A test of the equality of several SN ratios for the systems with dynamic characteristics, *Journal of the Japanese Society for Quality Control*, **33**, 83-92.

Appendix: Programming code for the Bayes analysis

The R package BRugs has been used for the Bayes analysis. The programming code reads as follows

```
library(BRugs)
# Check the model
modelCheck("effect_size_model.txt")
# Load the data
modelData("dentifrice_data.txt")
# Compile and number of chains
modelCompile(numChains=2)
# Set seed for reproducing the results
modelSetSeed(2923)
# Initial value of chain 1
# Here: Initial value 0
modelInits("initial_values_chain1.txt", chainNum=1)
# Initial value of chain 2
# Here: Initial value 0.5
modelInits("initial_values_chain2.txt", chainNum=2)
# Set the parameter
samplesSet("theta")
# Update the model
modelUpdate(100000)
# Statistcs; set the length of the burn-in period
# The length of the burn-in period is 5000 here
samplesStats("*", beg=5000)
# Plot the posterior density
plotDensity("theta", xlab=expression(theta), lwd=2)
```

The code for the model is:

```
model { for(i in 1:N){
     g[i] ~ dnorm(mu[i], tau[i])
     mu[i] <-  theta
     tau[i] <- 1 / (a[i] * ( b[i] +pow( theta, 2 ) ) ) }
theta ~dnorm(0, 1.0E-10)
}
```