Published in final edited form as:
J Evol Biol. 2011 August ; 24(8): 1836–1841. doi:10.1111/j.1420-9101.2011.02297.x.

Optimally weighted *Z*-test is a powerful method for combining probabilities in meta-analysis

Dmitri V. Zaykin[1]

[1] National Institute of Environmental Health Sciences, National Institutes of Health

Abstract

The inverse normal and Fisher's methods are two common approaches for combining *P*-values. Whitlock demonstrated that a weighted version of the inverse normal method, or "weighted *Z*-test" is superior to Fisher's method for combining *P*-values for one-sided *T*-tests. The problem with Fisher's method is that it does not take advantage of weighting and loses power to the weighted *Z*-test when studies are differently sized. This issue was recently revisited by Chen who observed that Lancaster's variation of Fisher's method had higher power than the weighted *Z*-test. Nevertheless, the weighted *Z*-test has comparable power to Lancaster's method when its weights are set to square roots of sample sizes. Power can be further improved when additional information is available. Although there is no single approach that is the best in every situation, the weighted *Z*-test enjoys certain properties that make it an appealing choice as a combination method for meta analysis.

Keywords

combining *P*-values; meta-analysis

Introduction

Evolutionary biologists have long used meta-analytic approaches to combine information from multiple studies. When raw data cannot be pooled across studies, meta analysis based on *P*-values presents a convenient approach that can be nearly as powerful as that based on combining data. Many popular *P*-value combination methods take the same general form, where *P*-value for the *i*-th study, p_i, is transformed by some function H, possibly utilizing study-specific weights, w_i. Next, a sum is taken, and the combined *P* -value is computed using the distribution of the resulting statistic, $T = \Sigma w_i H(p_i)$. For example, Stouffer's (also known as "inverse normal") method (Stouffer et al., 1949) takes H to be the inverse normal distribution function. Lipták's method (Lipták, 1958) is Stouffer's method with weights; this method is commonly referred to as the weighted *Z*-test. Fisher's method (Fisher, 1932) sets $H(p_i) = -2 \ln(p_i)$. The binomial test (Wilkinson, 1951) counts the number of *P* -values that are below a threshold α, in which case H is the indicator function, $H(p_i) = I(p_i \leq \alpha)$. Truncated *P*-value methods (Zaykin et al., 2002) add up *P*-values that fall below a threshold α by setting $H(p_i) = \Sigma I(p_i \leq \alpha) \ln(p_i)$. Combined *P* -value can be used in support for a common hypothesis tested in all studies, and a series of non-significant results may collectively suggest significance.

Correspondence: Dmitri Zaykin, National Institute of Environmental Health Sciences, 111 T.W. Alexander Drive, South Bldg (101); Mail Drop: A3-03, RTP, NC 27709, zaykind@niehs.nih.gov, phone: (919) 541-0096, fax: (919) 541-4311.

Carefully chosen weights can, in general, improve power of combination methods. A motivation for the weighting may follow from the fact that different studies might be differently powered, and that should be reflected by the weighting. Consider the combined *P*-value of the weighted *Z*-test:

$$p_Z = 1 - \Phi\left(\frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}\right) \quad (1)$$

where $Z_i = \Phi^{-1}(1 - p_i)$; p_i is a *P* -value for the *i*-th study of *k* studies in total, w_i are weights, and Φ, Φ^{-1} denote the standard normal cumulative distribution function and its inverse. Lipták suggested that the weights in this method "*should be chosen proportional to the 'expected' difference between the null hypothesis and the real situation and inversely proportional to the standard deviation of the statistic used in the i-th experiment*" and further suggested that when nothing else is available but the sample sizes of the studies (n_i), then the square root of n_i can be used as a weight (Lipták, 1958). Won et al. verified Lipták's claim more formally by showing that his test has optimal power when weights are set to the expected difference (i.e. the effect size) over the known or the estimated standard error (Won et al., 2009). This method of weighting requires knowledge of anticipated effect sizes for all combined studies, which is rarely available. Weightings by the estimated standard error or by the square root of sample size are more feasible in practice.

When different samples are taken from similar populations, a model that assumes a common effect size and direction among samples is appropriate. The ideal approach in this case is to pool raw data from all samples and to conduct a single statistical test. Whitlock considered such a test with its *P*-value and evaluated how well a combined *P*-value approximates this "true" *P*-value (Whitlock, 2005). He evaluated Fisher's method for combining *P*-values (Fisher, 1932) as well as the unweighted and weighted *Z*-tests, using one-sided *P*-values. Indeed, Whitlock found via simulation experiments that a weighted version of the combination *Z*-test outperformed both Fisher's and Stouffer's methods. Nevertheless, weighted versions of Fisher's method exist and it had remained unclear whether the power of a weighted version of Fisher's method may be as powerful as that of the weighted *Z*-test. This issue was recently taken on by Chen who found that Lancaster's generalization of Fisher's test was more powerful than the weighted *Z*-test (Chen, 2011). In Chen's application, *P*-values were transformed to chi-square variables by an inverse chi-square transformation with the degrees of freedom equal to the sample size of the study, i.e. Lancaster's statistic is $T = \sum \left[\chi^2_{(n_i)}\right]^{-1}(1 - p_i)$ with the distribution $T \sim \chi^2_{(\sum n_i)}$.

Both Whitlock and Chen used non-optimal weights for the weighted *Z*-test, setting them to the sample sizes of the studies. The original Whitlock's conclusions are valid, but the weights need to be adjusted according to suggestions by Lipták and Won et al. In Whitlock's setup, samples that corresponded to different studies were drawn from the same population. In this setup, the *T*-test based on pooled raw data can be viewed as an "ideal" test. In this case, optimal weights for the weighted *Z* method are given by the square root of the sample sizes, $\sqrt{n_i}$. These weights are optimal in the sense that the combined *P* -value approximates the *P* value of the test based on raw data. This can be seen from writing out a *Z* statistic based on pooled raw data in terms of statistics for the individual studies. The pooled data statistic is $Z_{\text{total}} = \sqrt{n_T}\overline{T}/\widehat{S}_T$, where $\overline{T}$ is the sample average for the total sample of size n_T and $\hat{S}_T$ is the sample standard deviation. Suppose that we split the sample into two parts of sizes

n_X, n_Y and calculate sample means ($\overline{X}$, $\overline{Y}$) and standard deviations ($\hat{S}_X$, $\hat{S}_Y$) separately for these two samples. We can write the pooled statistic in terms of the two means as

$$Z_{\text{total}} = \frac{n_X \overline{X}}{\sqrt{n_T}\widehat{S}_T} + \frac{n_Y \overline{Y}}{\sqrt{n_T}\widehat{S}_T}$$

while the weighted statistic that combines information from the two samples is

$$Z_w = \frac{w_X \frac{\sqrt{n_X}\overline{X}}{\widehat{S}_X} + w_Y \frac{\sqrt{n_Y}\overline{Y}}{\widehat{S}_Y}}{\sqrt{w_X^2 + w_Y^2}}$$

The pieces $\frac{\sqrt{n_X}\overline{X}}{\widehat{S}_X}$ and $\frac{\sqrt{n_Y}\overline{Y}}{\widehat{S}_Y}$ can be recovered from *P*-values for the two samples by the inverse normal transformation. This statistic is the weighted *Z*-test for combining *P*-values. We can see that Z_w approximates Z_{total} when the weights w_X, w_Y are set to $\sqrt{n_X}$, $\sqrt{n_Y}$. The same argument holds for more than two samples. Regarding Lancaster's method, Chen noted cautiously that setting degrees of freedom to the sample size of the *i*-th study "may not be optimal". It is an optimal weighting however for his simulation setup, where samples are obtained from the same population. This follows from the fact that the chi-square distribution for the *i*-th statistic approaches a normal distribution with the variance $2n_i$: when the degrees of freedom are set to n_i, the variance of the *i*-th statistic in Lancaster's method is proportional to the variance of the corresponding term for the optimally weighted *Z*-test. Thus, power advantage of Lancaster's method over the weighted *Z* method observed by Chen was at least to some degree due to the usage of non-optimal weights for the *Z* method. As I will verify by simulation experiments, power of the optimally weighted *Z* method at conventional 1% and 5% levels is very similar to that of Lancaster's method.

Chen chose Lancaster's method in favor of an extension of Fisher's test where weighted inverse chi-square-transformed *P*-values are added, for the reason that "the distribution of the sum of weighted χ^2 is usually unknown". Several algorithms for obtaining this distribution have been published however, and are freely available. Duchesne and Lafaye De Micheaux recently described an R package that implements several approximations to that distribution as well as "exact" algorithms with guaranteed, user-controlled precision (Duchesne & Lafaye De Micheaux, 2010). The weighted Fisher's test is a direct χ^2-based analogue of the weighted *Z*-test. Therefore, I included this method into comparisons. Specifically, the weighted Fisher's test (the weighted χ^2 test) is based on the distribution of the following statistic:

$$F_w = \sum_{i=1}^{k} w_i \left[\chi^2_{(2)}\right]^{-1}(1 - p_i) \quad (2)$$

where $\left[\chi^2_{(2)}\right]^{-1}$ is the inverse cumulative chi-square distribution function with two degrees of freedom.

Methods

For simulation experiments I followed the setup of Chen and Whitlock. I assumed a *T*-test for the null hypothesis H_0: $\mu > 0$ and values of μ from 0 to 0.1 with an increment of 0.01. For eight studies with sample sizes n_i of 10,20,40,80,160,320,640, and 1280, random samples were obtained assuming a normal distribution with the mean μ and the variance of one. As in Chen, power values were computed for two significance levels, $\alpha = 0.01$ and $\alpha = 0.05$. Weightings by $\sqrt{n_i}$, by the inverse of the estimated standard error ($1/\widehat{\mathrm{SE}_i}$), and by the standardized effect size, ($\mu/\widehat{\mathrm{SE}_i}$) were considered. The number of simulations was 30,000. Tukey's plots (Tukey, 1977) showing correspondence of combined and "true" *P*-values (i.e. obtained from a statistic on pooled data) were obtained for $\mu = 0$ and $\mu = 0.05$. In Tukey's plots, $(X+Y)/2$ is plotted against $Y - X$. Large spread on the plot indicates discrepancy between the *X* and *Y* values. Combined *P*-value for the weighted Fisher's method (Equation 2) was obtained with a function from R package CompQuadForm (Duchesne & Lafaye De Micheaux, 2010) that implements Farebrother's "exact" method (Farebrother, 1984). In addition, I considered three scenarios with study heterogeneity. In the first scenario, μ value for the *i*-th study was randomly drawn without replacement from the vector of values (0.01, 0.02, …, 0.1) for each simulation run. In the second scenario, μ was assumed fixed (0.07), and the standard deviation value for the *i*-th study (σ_i) at every simulation step was drawn without replacement from eight values that were equally spaced, starting from $^3/_4$ to $2\ ^3/_4$. In the third scenario, both, μ_i and σ_i were randomly drawn for each simulation run.

Results

Tables 1 and 2 present power values for the studied tests. Table 1 that followed the setup of Whitlock and Chen shows that the weighted *Z* test with weights $\sqrt{n_i}$, $1/\widehat{\mathrm{SE}_i}$, Lancaster's method, and the test based on pooled data all have nearly identical power. The weighted Fisher's test has a slightly lower power. Table 2 shows power values for heterogeneity scenarios as well as type-I error rates for the case $\mu = 0$ but with a random, study-specific variance. The total *T* test is no longer most powerful in this case, due to heterogeneity of effects. Weighting by either $\sqrt{n_i}$ or by $1/\widehat{\mathrm{SE}_i}$ delivers the same improvement in power when only the means are heterogeneous between studies. When there is heterogeneity of the variances, weighting by $1/\widehat{\mathrm{SE}_i}$ yields a power advantage over weighting by $\sqrt{n_i}$. Power is the highest when standardized effects ($\mu/\widehat{\mathrm{SE}_i}$) as used as weights. Correlations between the true and the combined *P*-values were found to be at least 99% for all values of μ for Lancaster's and the weighted *Z* methods. The corresponding correlation for the weighted Fisher's method was lower, ranging from about 91% to 94% depending on the value of μ. Tukey's plots in Figure 1 show a good correspondence of *P*-values for the pooled data test with *P*-values for Lancaster's and the $\sqrt{n_i}$–weighted *Z* methods. Lancaster's method forms a more "snowy" cloud and the weighted *Z* method *P*-values are somewhat closer to the true values.

Discussion

Meta-analysis of *P*-values generally benefits from weighting. When samples are obtained from the same or similar populations, as in the model studied by Whitlock and Chen, the optimal weights for the *Z*-test are given by $\sqrt{n_i}$. In this case, the weighted *Z*-test, Lancaster's test and the test based on pooled data provide very similar power. This is expected, because Lancaster's method approaches the weighted *Z* method asymptotically, as $\min(n_i)$ increases. When there is heterogeneity of variances, but the true mean is the same across studies, weighting by $1/\widehat{\mathrm{SE}_i}$ is optimal, but the gain in power is not great, compared to weighting by

$\sqrt{n_i}$ (0.784 vs. 0.743 at α=5%). To an extent, power increase is small because of the large range of sample sizes. A constant sample size of n=289 would have given the powers of 0.801 vs. 0.743 respectively, for the same assumed heterogeneity of variances, max (σ^2)/min (σ^2)=13.4. When there is heterogeneity of means, Z-test that uses standardized effect sizes as weights has the largest power (Lipták, 1958; Won et al., 2009), however an application of this test requires the knowledge of μ. Note that this value needs to be pre-specified: plugging in an estimate $\hat{\mu}$ obtained from the same data that was used to compute P-values would invalidate the combination test.

In this study, one-sided P-values were assumed. Such P-values are appropriate for meta-analytic combination of P-values from several studies. Two-sided P-values are generally inappropriate, because they are oblivious to the effect direction. Two-sided P-values from two studies in which the effect direction is flipped can both be small nevertheless, resulting in an inappropriately small combined P-value. On the other hand, combined result of corresponding one-sided P-values will properly reflect cancellation of the pooled effect that would have been observed if raw data from the two studies were combined.

Despite the fact that the mechanics of the meta-analytic process involves manipulation of one-sided P-values, it is often the case that the final result needs to be a two-sided P-value. For example, when allele frequencies are compared between two groups of individuals classified based on the presence or absence of a trait, the null hypothesis is usually that the frequency is the same, and the alternative hypothesis does not specify a particular effect direction. The weighted Z-test provides an important advantage in dealing with this situation, due to symmetry of the normal transformation. There are two possible one-sided combined P-values for each assumed effect direction, but with the weighted Z method, the combined P-value for the first assumed direction is the same distance from $^1/_2$ as the combined P-value for the second assumed direction. Therefore, one can arbitrarily assume either one of the two directions when computing one-sided P-values, and obtain a combined one-sided P-value, $p_{\text{one-sided}}$. The two-sided combined P -value is the same regardless of the assumed direction:

$$p_{\text{two-sided}} = \begin{cases} 2\,p_{\text{one-sided}} & \text{if } p_{\text{one-sided}} < 1/2 \\ 2\,(1 - p_{\text{one-sided}}) & \text{otherwise} \end{cases} \tag{3}$$

What if available individual P-values are all two-sided? Often, studies report P-values that correspond to statistics such as $|T|$ and $|Z|$, or its squared value, i.e. the one degree of freedom chi-square. These individual P-values can be converted to one-sided before combining as follows:

$$p_{\text{one-sided}} = \begin{cases} p_{\text{two-sided}}/2; & \text{if effect direction} > 0 \\ 1 - p_{\text{two-sided}}/2; & \text{otherwise} \end{cases}$$

Once again, the assumed effect direction can be chosen arbitrarily. For example, in testing for association of an allele with a trait at a biallelic locus A/a, we can arbitrarily choose one of the alleles, e.g. allele A. Then the “effect direction” for i-th study is positive if there is positive correlation of that allele with the presence of the trait in that study. Once these one-sided P-values are combined, the result can be converted back to two-sided by Equation (3).

Another advantage of the weighted Z test is that it can be easily extended to account for the case of correlated statistics between studies. For the test to be valid under independence, we need an assumption that the set of $\{Z_i\}$ jointly follows a multivariate normal distribution

under the null hypothesis. If $\mathrm{cor}(Z_i, Z_j) = r_{ij}$, the modification amounts to replacing the denominator in Equation (1) with $\sqrt{\sum w_i^2 + 2\sum_{i<j} w_i w_j r_{ij}}$. The multivariate normal assumption is often justified asymptotically, and in certain situations the correlations $\{r_{ij}\}$ are known. For example, when each Z_i is a result of comparing group i of sample size n_i to a common "control" group of sample size n_0 by the two-sample T -test, then $r_{ij} = \sqrt{[1/(1+n_0/n_i)]\,[1/(1+n_0/n_j)]}$ (Dunnett, 1955). In principle, a variation of the weighted Fisher's method can be extended to this situation, if we can assume that chi-square statistics formed from individual P-values can be represented by squares of underlying multivariate normal variables with correlations r_{ij}. However, required computations are more involved. First, the two degree of freedom chi-square transformation in Equation (2) would have to be replaced with the one degree of freedom transformation. Then one would need to compute eigenvalues of $\mathrm{diag}(\sqrt{\mathbf{w}})\,(\mathbf{R} \circ \mathbf{R})\,\mathrm{diag}(\sqrt{\mathbf{w}})^T$, where $\mathbf{w}$ is the vector of weights and $\mathbf{R} \circ \mathbf{R}$ is the matrix of squared correlations. Finally, to compute the combined P-value, one can use the fact that the weighted sum of these correlated chi-squares can be represented by the sum of independent weighted chi-squares with weights given by the above eigenvalues (Box, 1954). Thus, one can use the observed weighted sum of correlated chi-squares with weights substituted by the eigenvalues as an input to a routine for computing the cumulative distribution of the sum of independent weighted chi-squares.

Although there is no single method for combining P-values that is most powerful in all situations, a meta analytic setup considered by Whitlock and extended here to include study heterogeneity is quite general, because many forms of one-sided statistics approach a normal distribution asymptotically. Therefore, the $\sqrt{n_i}$– or $1/\widehat{\mathrm{SE}_i}$–weighted Z-test for combining one-sided P-values can be recommended in most situations.

In this study, the weighted Fisher's method showed slightly smaller power values compared to other methods in this study. If absolute values or squares of T-statistics for each study were assumed instead, as in calculation of two-sided tests, the weighted Fisher's would have yielded higher power values than either Lancaster's or the weighted Z methods. As already noted, combining individual two-sided P-values is generally not appropriate in meta-analysis, where the same hypothesis is tested in all studies. Combination of two sided P-values is more appropriate when individual tests are concerned with separate hypotheses. Small combined P-value in that case can be interpreted as evidence that one or more individual null hypotheses are false. Owing to the virtue of being sensitive to small P-values, the weighted Fisher's method would provide good power, especially in those situations where there is pronounced heterogeneity of effect sizes between studies.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. I wish to express my appreciation for comments and suggestions by Professors Michael Whitlock and Allen Moore, and by an anonymous reviewer.

References

Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the One-Way classification. The Annals of Mathematical Statistics. 1954; 25:290–302.

Chen Z. Is the weighted z-test the best method for combining probabilities from independent tests? Journal of Evolutionary Biology. 2011; 24:926–930. [PubMed: 21401770]

Duchesne P, Lafaye De Micheaux P. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. Computational Statistics & Data Analysis. 2010; 54:858–862.

Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc. 1955; 50:1096–1121.

Farebrother RW. Algorithm AS 204: The distribution of a positive linear combination of χ^2 random variables. Applied Statistics. 1984; 33:332–339.

Fisher, R. Statistical methods for research workers. Oliver and Boyd; Edinburgh: 1932.

Lipták T. On the combination of independent tests. Magyar Tud Akad Mat Kutato Int Közl. 1958; 3:171–196.

Stouffer, S.; DeVinney, L.; Suchmen, E. The American soldier: Adjustment during army life. Vol. 1. Princeton University Press; Princeton, US: 1949.

Tukey, J. Exploratory data analysis. Addison-Wesley; Boston, Massachusetts, US: 1977.

Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. Journal of Evolutionary Biology. 2005; 18:1368–1373. [PubMed: 16135132]

Wilkinson B. A statistical consideration in psychological research. Psychological Bulletin. 1951; 48:156–158. [PubMed: 14834286]

Won S, Morris N, Lu Q, Elston R. Choosing an optimal method to combine P-values. Statistics in medicine. 2009; 28:1537–1553. [PubMed: 19266501]

Zaykin D, Zhivotovsky L, Westfall P, Weir B. Truncated product method for combining P-values. Genetic Epidemiology. 2002; 22:170–185. [PubMed: 11788962]

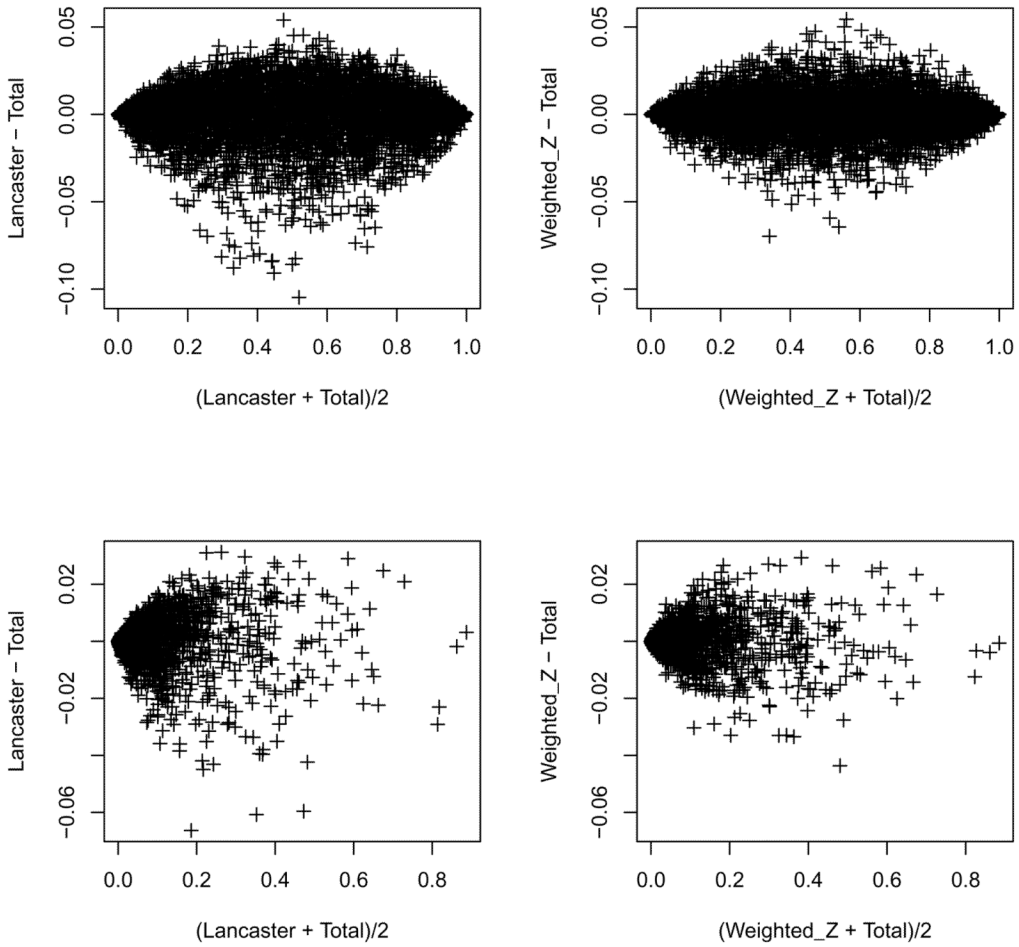

Figure 1.

Tukey's plots of *P*-values for Lancaster's and the $\sqrt{n_i}$–weighted *Z*-test vs. the total *T* test. Top row: $\mu = 0$. Bottom row: $\mu = 0.05$.

Table 1

Power assuming a common μ value for all samples

Method	μ										
	α	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**0.10**
Total T	0.01	0.034	0.093	0.209	0.382	0.580	0.751	0.882	0.955	0.986	0.997
	0.05	0.129	0.262	0.450	0.650	0.807	0.917	0.970	0.991	0.998	1.000
Lancaster's	0.01	0.034	0.093	0.208	0.382	0.579	0.750	0.882	0.955	0.986	0.997
	0.05	0.129	0.261	0.449	0.649	0.806	0.915	0.970	0.991	0.998	1.000
Weighted Z (by $\sqrt{n}$)	0.01	0.034	0.093	0.208	0.383	0.579	0.750	0.882	0.955	0.986	0.997
	0.05	0.129	0.261	0.450	0.649	0.807	0.916	0.970	0.991	0.998	1.000
Weighted Z (by $1/\widehat{\mathrm{SE}}$)	0.01	0.034	0.094	0.209	0.384	0.579	0.750	0.883	0.954	0.986	0.997
	0.05	0.129	0.262	0.451	0.649	0.807	0.915	0.970	0.991	0.998	1.000
Weighted χ^2 (by $\sqrt{n}$)	0.01	0.031	0.082	0.183	0.339	0.530	0.702	0.848	0.936	0.978	0.994
	0.05	0.121	0.242	0.420	0.611	0.774	0.892	0.957	0.987	0.997	0.999
Weighted χ^2 (by $1/\widehat{\mathrm{SE}}$)	0.01	0.032	0.084	0.184	0.341	0.532	0.705	0.849	0.936	0.978	0.994
	0.05	0.122	0.244	0.423	0.614	0.775	0.893	0.957	0.986	0.997	0.999

J Evol Biol. Author manuscript; available in PMC 2012 August 1.

J Evol Biol. Author manuscript; available in PMC 2012 August 1.

Table 2

Type-I error and power assuming heterogeneous μ and σ^2 values

Method	α	Type-I error	random μ	random σ^2	random μ, σ^2
Total T	0.01	0.010	0.634	0.370	0.248
	0.05	0.049	0.812	0.618	0.461
Weighted Z (by $\sqrt{n}$)	0.01	0.010	0.634	0.523	0.351
	0.05	0.049	0.812	0.743	0.568
Weighted Z (by $1/\widehat{\mathrm{SE}}$)	0.01	0.010	0.634	0.584	0.394
	0.05	0.049	0.812	0.784	0.600
Weighted Z (by $\mu/\widehat{\mathrm{SE}}$)	0.01	0.010	0.719	0.584	0.443
	0.05	0.049	0.872	0.784	0.656
Lancaster's	0.01	0.010	0.637	0.525	0.356
	0.05	0.049	0.814	0.745	0.571
Weighted χ^2 (by $\sqrt{n}$)	0.01	0.010	0.627	0.504	0.357
	0.05	0.049	0.810	0.722	0.563
Weighted χ^2 (by $1/\widehat{\mathrm{SE}}$)	0.01	0.010	0.629	0.546	0.385
	0.05	0.049	0.811	0.755	0.591