AppliedMath

Review

Two *P* or Not Two *P*: Mendel Random Variables in Combining Fake and Genuine *p*-Values

M. Fátima Brilhante [1,2,*], M. Ivette Gomes [2,3,4,5], Sandra Mendonça [2,6], Dinis Pestana [2,3,5] and Rui Santos [2,7]

[1] Departamento de Matemática e Estatística, Faculdade de Ciências e Tecnologia, Universidade dos Açores, Rua da Mãe de Deus, 9500-321 Ponta Delgada, Portugal
[2] Centro de Estatística e Aplicações, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal; migomes@ciencias.ulisboa.pt (M.I.G.); sandram@staff.uma.pt (S.M.); ddpestana@ciencias.ulisboa.pt (D.P.); rui.santos@ipleiria.pt (R.S.)
[3] Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
[4] Academia das Ciências de Lisboa, Rua da Academia das Ciências 19, 1249-122 Lisboa, Portugal
[5] Instituto de Investigação Científica Bento da Rocha Cabral, Calçada Bento da Rocha Cabral 14, 1250-012 Lisboa, Portugal
[6] Departamento de Matemática—FCEE, Campus Universitário da Penteada, Universidade da Madeira, 9020-105 Funchal, Portugal
[7] Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, Apartado 4133, 2411-901 Leiria, Portugal
* Correspondence: maria.fa.brilhante@uac.pt

Abstract: The classical tests for combining p-values use suitable statistics $T(P_1, \ldots, P_n)$, which are based on the assumption that the observed p-values are genuine, i.e., under null hypotheses, are observations from independent and identically distributed Uniform$(0,1)$ random variables $P_1, \ldots, P_n$. However, the phenomenon known as publication bias, which generally results from the publication of studies that reject null hypotheses of no effect or no difference, can tempt researchers to replicate their experiments, generally no more than once, with the aim of obtaining "better" p-values and reporting the smallest of the two observed p-values, to increase the chances of their work being published. However, when such "fake p-values" exist, they tamper with the statistic $T(P_1, \ldots, P_n)$ because they are observations from a Beta$(1,2)$ distribution. If present, the right model for the random variables P_k is described as a tilted Uniform distribution, also called a Mendel distribution, since it was underlying Fisher's critique of Mendel's work. Therefore, methods for combining genuine p-values are reviewed, and it is shown how quantiles of classical combining test statistics, allowing a small number of fake p-values, can be used to make an informed decision when jointly combining fake (from Two P) and genuine (from not Two P) p-values.

Keywords: combined p-values; fake p-values; Mendel random variables

MSC: 62A01; 62P10

check for updates

Citation: Brilhante, M.F.; Gomes, M.I.; Mendonça, S.; Pestana, D.; Santos, R. Two *P* or Not Two *P*: Mendel Random Variables in Combining Fake and Genuine *p*-Values. *AppliedMath* **2024**, *4*, 1128–1142. https://doi.org/10.3390/appliedmath4030060

Received: 1 August 2024
Revised: 27 August 2024
Accepted: 3 September 2024
Published: 5 September 2024

1. Introduction

The concept of p-value is generally credited to Pearson [1], although it was implicitly used much earlier by Arbuthnot [2] in 1710. Defined as the probability of obtaining, under a null hypothesis, a result that is as extreme or more extreme than the one observed, it was considered to be an informal index to assess the discrepancy between the data and the hypothesis under investigation. The use of p-values gained popularity with Sir Ronald Fisher [3,4], and about their use, Fisher [5] states that "A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this [P = 0.05] level of significance". Therefore, the question of reproducibility of results was naturally raised (cf. Greenwald et al. [6], or Colquhoun [7]), which in turn demanded the

p-values collected from replicated experiments to be summarized into a combined p-value. In 1931, Tippett [8], a co-worker of Fisher, performed the first meta-analysis of p-values, and in 1932, Fisher himself [9] suggested a method for combining p-values.

The classical combined test procedures assume that the observed p-values, $p_1, \dots, p_n$, are, under null hypotheses H_{0k}, $k = 1, \dots, n$, of no difference or no effect, observations from independent random variables $P_k \sim \text{Uniform}(0,1)$, which is an immediate consequence of the probability integral transform theorem. It is then said that a p_k from $P_k \sim \text{Uniform}(0,1)$ is a genuine (or a true) p-value.

Section 2 describes some classical methods for combining p-values, using either their values directly, for example through order statistics or Pythagorean means, or using basic transformations of standard uniform random variables, such as $-\ln P_k$ and $\Phi^{-1}(P_k)$, where Φ^{-1} is the inverse of the standard Gaussian cumulative distribution function, or the logit function $\ln\left(\frac{P_k}{1-P_k}\right)$. For additional p-values combinations, see Brilhante et al. [10].

Although today there is an intense debate on whether significance testing, and therefore the use of p-values, is an acceptable scientific research tool; see, for instance, the editorials in *The American Statistician* vol. 70 (Wasserstein and Lazar [11]) and vol. 73 (Wasserstein et al. [12]) on the topic, traditionally low p-values were a valid passport for being published. This has created a so-called file drawer problem due to publication bias. As with other techniques used in meta-analysis, publication bias can easily lead to false conclusions. In fact, the set of available p-values comes mainly from studies considered worthy of publication because the observed p-values were small, presumably indicating significant results. Thus, the assumption that the p_k's are observations from independent $\text{Uniform}(0,1)$ random variables is quite questionable since generally they are a set of low order statistics, given that p-values greater than 0.05 have less chances of being published.

One way of assessing publication bias is by computing the number of non-significant p-values that would be needed to reverse the decision to reject the overall hypothesis based on a set of available p-values. For example, Jin et al. [13] and Lin and Chu [14] give interesting overviews of how to deal with publication bias. Givens et al. [15] also provide a deep insight into publication bias in meta-analysis, namely using data-augmentation techniques.

Publication bias is also the cause of poor scientific practices, in some cases even fraud, especially when the replication of experiments is carried out with the intent of, hopefully, obtaining more favorable p-values to increase the chances of publishing. While replicating experiments is legitimate and recommended to establish consistent results, replicating with the purpose of reporting the smallest of the observed p-values is an unacceptable scientific practice. If this is indeed the case, the reported "fake" p-value, being the minimum of $\ell > 1$ independent standard uniform random variables, is $\text{Beta}(1,\ell)$-distributed. However, replicating experiments has a cost, either monetary or timewise, and if in the replication of an experiment only once, both p-values obtained are greater than 0.05, then what appears to be the wisest decision is not to continue replicating the experiment; otherwise, the smallest of the two p-values is reported, or none at all. In fact, what seems realistic to consider is either $\ell = 2$, and therefore a nuisance "fake two p-value" is reported, or $\ell = 1$, i.e., a "genuine", not the minimum of "two p-value", is disclosed.

In Fisher's [16] comments about Mendel's work, he conjectured that "the data of most, if not all, of the experiments have been falsified to agree closely with Mendel's expectations". Fisher made it quite clear that he suspected that Mendel's "too good to be true" results were carefully chosen to support the hereditary theory that Mendel wanted to prove. Due to this historical background, in Section 3, we shall call Mendel distribution the model that is a mixture of a $\text{Beta}(1,2)$ (or $\text{Beta}(2,1)$) distribution and a $\text{Uniform}(0,1)$ distribution, thus representing a mixture of "fake two p-value" and "genuine not two p-value". We briefly explain how an extension of Deng and George's [17] characterization of the standard uniform distribution using a Mendel random variable instead of a uniform random variable can be considered to test the uniformity of a set of p-values or determine if it is contaminated with fake p-values.

In Section 4, an example is given to illustrate how to use the critical values from the tables in Brilhante et al.'s [10] supplementary materials for jointly combining genuine and fake p-values using classical combining methods. The example shows that a thorough comparison should always be made, since most likely there is no reliable information that rules out the existence of fake p-values that have resulted from bad scientific practices, and therefore it is important to acknowledge their potential effects when performing a meta-analysis of p-values.

In Section 5, further developments for combining p-values are reviewed, with a very brief reference to the recent research field on e-values. Finally, Section 6 reinforces the recommendation that when extending the usual combined tests to include genuine and fake p-values, they should be compared with each other in terms of the conclusions drawn for an informed final decision.

2. An Overview of Classical Combined Tests for p-Values

Let us assume that the p-values p_k are known for testing H_{0k} versus H_{Ak}, $k = 1, \ldots, n$, in n independent studies on some common issue, and that the objective is to decide on the overall hypothesis H_0^*: all the H_{0k} are true versus H_A^*: some of the H_{Ak} are true. As there are many different ways in which H_0^* can be false, selecting the right test is generally unfeasible. On the other hand, combining the available p_k's so that a function $T(p_1, \ldots, p_n)$ is the observed value of a random variable with a known sampling distribution under H_0^* is a simple problem, since under H_0^*, $(p_1, \ldots, p_n)$ is the observed value of a random sample $(P_1, \ldots, P_n)$ from a $\text{Uniform}(0,1)$ distribution. In fact, several different and reasonable combined testing procedures are often used with suitable functions of the p_k's. Moreover, it should be guaranteed that a combined procedure is monotone, in the sense that if one set of p-values $(p_1, \ldots, p_n)$ leads to the rejection of the overall null hypothesis H_0^*, then any set of component-wise smaller p-values $(p'_1, \ldots, p'_n)$, i.e., $p'_k \le p_k$, $k = 1, \ldots, n$, must also lead to its rejection.

Tippett [8] used the statistic

$$T_T(P_1, \ldots, P_n) = \min\{P_1, \ldots, P_n\} = P_{1:n}\,.$$

From the fact that $P_{1:n}|H_0^* \sim \text{Beta}(1, n)$, the criterion for rejecting H_0^* at a significance level α is $p_{1:n} < 1 - (1-\alpha)^{1/n}$. Tippett's method is a special case of Wilkinson's method [18], which recommends that H_0^* should be rejected when some observed order statistic $p_{k:n} < c$. As $P_{k:n}|H_0^* \sim \text{Beta}(k, n+1-k)$, the cut-of-point c to reject H_0^* is the solution of

$$\int_0^c x^{k-1}(1-x)^{n-k}\mathrm{d}x = \alpha\, B(k, n+1-k)\,,$$

where $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\mathrm{d}x$, $p, q, > 0$, is the Beta function.

Simes [19], on the other hand, gives an interesting development of Wilkinson's method: Let $P_{1:n}, \ldots, P_{n:n}$ be the ordered p-values for testing the overall hypothesis H_0^*, which should be rejected at a significance level α if $P_{j:n} \le j\alpha/n$ for any $j = 1, \ldots, n$.

Another way of constructing combined p-values is to use functions of standard uniform random variables. Fisher [9] suggested the use of the statistic

$$T_F(P_1, \ldots, P_n) = -2\sum_{k=1}^{n} \ln P_k\,,$$

since $-2\ln P_k \sim \chi_2^2$ when $P_k \sim \text{Uniform}(0,1)$, $k = 1, \ldots, n$. As $-2\sum_{k=1}^{n} \ln P_k | H_0^* \sim \chi_{2n}^2$, the criterion for rejecting H_0^* at a significance level α is $-2\sum_{k=1}^{n} \ln p_k > \chi^2_{2n,1-\alpha}$, with $\chi^2_{m,p}$ denoting the p-th quantile of the chi-square distribution with m degrees of freedom.

Tippett's method illustrates the direct use of standard uniform random variables, while Fisher's method shows the use of transformed standard uniform random variables. Moreover, Fisher's method is often the most efficient way of making use of all the information available, whereas Tippett's method disregards almost all available information. Therefore, these two methods can be viewed as two extreme cases.

Combining p-values using functions of their sums or products, namely their arithmetic mean or their geometric mean, is also feasible but less appealing than Fisher's chi-square transformation method. Edgington [20] suggested the use of the arithmetic mean as a test statistic, i.e.,

$$T_E(P_1,\dots,P_n)=\overline{P}_n=\frac{1}{n}\sum_{k=1}^{n}P_k\,,$$

but it has a very cumbersome probability density function, defined as

$$f_{\overline{P}_n}(x)=\frac{n}{\Gamma(n)}\left[\sum_{j=0}^{\lfloor nx\rfloor}(-1)^j\binom{n}{j}\left(\max\{0,nx-j\}\right)^{n-1}\right]\mathbb{I}_{[0,1)}(x),$$

with $\lfloor x\rfloor$ being the largest integer not greater than x and $\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x$, $\alpha>0$, Euler's Gamma function. However, if n is large, an approximation based on the central limit theorem can be used to perform an overall test on H_0^* versus H_A^*, but it is not consistent, in the sense that it can fail to reject the overall test's null hypothesis, even though the results of some of the individual tests are extremely significant.

Pearson's [21] proposal for combining p-values is based on their product, i.e., on the statistic

$$T_P(P_1,\dots,P_n)=\prod_{k=1}^{n}P_k\,,$$

which under H_0^* has a probability density function

$$f_{T_P}(x)=\frac{(-\ln x)^{n-1}}{\Gamma(n)}\mathbb{I}_{(0,1)}(x)\,.$$

In other words, $\prod_{k=1}^{n}P_k|H_0^*\sim\mathrm{BetaBoop}(1,1,1,n)$ (see Brilhante et al. [22] for more details on BetaBoop random variables). Consequently, the geometric mean

$$\mathcal{G}_n=T_{\mathcal{G}_n}(P_1,\dots,P_n)=\left(\prod_{k=1}^{n}P_k\right)^{1/n}$$

has a cumulative distribution function

$$F_{\mathcal{G}_n}(x)=\frac{\Gamma(n,\,-n\,\ln x)}{\Gamma(n)}\,\mathbb{I}_{(0,1)}(x)+\mathbb{I}_{[1,\infty)}(x)\,,$$

where $\Gamma(\alpha,z)=\int_z^\infty x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x$, $\alpha,z>0$, is the upper incomplete Gamma function. The critical quantiles $g_{n,1-\alpha}$ of $\mathcal{G}_n$ can easily be computed from the critical quantiles $g^*_{n,1-\alpha}$ of $\mathcal{G}_n^n=T_P(P_1,\dots,P_n)$, where $\int_0^{g^*_{n,1-\alpha}}\frac{(-\ln x)^{n-1}}{\Gamma(n)}\,\mathrm{d}x=1-\alpha$, since $g_{n,1-\alpha}=(g^*_{n,1-\alpha})^{1/n}$.

Note, however, that using products of standard uniform random variables or adding their exponential logarithms provides essentially the same information, as recognized by Pearson [21] in his final remark, and hence, it is more convenient to use Fisher's statistic.

In 1934, Pearson [23] considered that in a bilateral framework, it would be more appropriate to use the statistic

$$T_{P^*}(P_1,\dots,P_n)=\min\left\{\prod_{k=1}^{n}P_k,\prod_{k=1}^{n}(1-P_k)\right\}.$$

Owen [24] suggested a simple modified version of $T_{P^*}(P_1, \dots, P_n)$, namely the statistic

$$T_O(P_1, \dots, P_n) = \max\left\{-2\sum_{k=1}^{n} \ln P_k, -2\sum_{k=1}^{n} \ln(1-P_k)\right\},$$

for which he recommends a Bonferroni correction to establish lower and upper bounds for the computation of probabilities. Another alternative to $T_{P^*}(P_1, \dots, P_n)$ is Pearson's [23] minimum of geometric means statistic,

$$T_{\min\{\mathcal{G}_n,\mathcal{G}_n^*\}}(P_1, \dots, P_n) = \min\left\{\left(\prod_{k=1}^{n} P_k\right)^{1/n}, \left(\prod_{k=1}^{n}(1-P_k)\right)^{1/n}\right\}.$$

Also, concerning the use of transformed p-values, Stouffer et al. [25] used as a test statistic

$$T_S(P_1, \dots, P_n) = \sum_{k=1}^{n} \frac{\Phi^{-1}(P_k)}{\sqrt{n}}.$$

Since $T_S(P_1, \dots, P_n)|H_0^* \sim \text{Gaussian}(0,1)$, the criterion for rejecting H_0^* at a significance level α is $\left|\sum_{k=1}^{n} \frac{\Phi^{-1}(P_k)}{\sqrt{n}}\right| > z_{1-\alpha}$, with z_p denoting the p-th quantile of the standard Gaussian distribution.

A further simple transformation based on the standard uniform random variables P_k and $1-P_k$ is the logit transformation $\ln\left(\frac{P_k}{1-P_k}\right) \sim \text{Logistic}(0,1)$, which was used by Mudholkar and George [26] to construct the combined test statistic

$$T_{MG}(P_1, \dots, P_n) = -\sum_{k=1}^{n} \ln\left(\frac{P_k}{1-P_k}\right).$$

Using the approximation

$$-\left(\frac{n\,\pi^2(5n+2)}{3(5n+4)}\right)^{-1/2} \sum_{k=1}^{n} \ln\left(\frac{P_k}{1-P_k}\right) \approx t_{5n+4},$$

H_0^* should be rejected at a significance level α if

$$\left|\left(\frac{n\,\pi^2(5n+2)}{3(5n+4)}\right)^{-1/2} \sum_{k=1}^{n} \ln\left(\frac{p_k}{1-p_k}\right)\right| > t_{5n+4,\,1-\alpha},$$

with $t_{m,p}$ denoting the p-th quantile of Student's t distribution with m degrees of freedom.

On the other hand, Birnbaum [27] has shown that every monotone combined test procedure is admissible, i.e., provides a most powerful test against some alternative hypothesis for combining a collection of tests, and therefore is optimal for a combined testing situation whose goal is to harmonize possibly conflicting evidence or to pool inconclusive evidence. In the context of social sciences, Mosteller and Bush [28] recommend Stouffer's method, but Littell and Folks [29,30] have shown that under mild conditions, Fisher's method is optimal for combining independent tests.

The thorough comparison performed by Loughin [31] shows that the normal combining function performs quite well in problems where the evidence against the combined null hypothesis is spread among more than a small fraction of the individual tests. However, when the total evidence is weak, Fisher's method is the best choice, especially when the evidence is at least moderately strong, and it is concentrated in a relatively small fraction of the individual tests. Mudholkar and George's [26] logistic combined test manages to provide a compromise between the two previous cases. Additionally, when the total evidence against the combined null hypothesis is concentrated on one or on a few tests to be combined, Tippett's combining function is useful.

3. Fake *p*-Values and Mendel Random Variables

An important issue that should be addressed before combining p-values is whether they are genuine or not. The overall alternative hypothesis H_A^* states that some of the individual H_{Ak} are true, and so a meta-decision on H_0^* implicitly assumes that some of the P_k's may not have a uniform distribution, cf. Hartung et al. [32] (pp. 81–84), and Kulinskaya et al. [33] (pp. 117–119). In fact, the uniformity of the P_k's is solely the consequence of assuming that the null hypothesis is true, but this questionable assumption led Tsui and Weerahandi [34] to introduce the concept of generalized p-values. See Weerahandi [35], Hung et al. [36] and Brilhante [37], and references therein, on the concepts of generalized and random p-values.

Moreover, the assumption $P_k|H_0 \sim \text{Uniform}(0,1)$, $k = 1, \dots, n$, can be unrealistic. As a matter of fact, when an observed p-value is not highly significant or significant, there is a possibility that the experiment will be repeated in the hope of obtaining a "better" p-value to increase the likelihood of the research being published. However, the scientific malpractice of trying to obtain better p-values to comply with research teams' expectations, which in some cases can be labeled as a fraudulent practice, can lead to disclosing results that are "too good to be true", as Fisher [16] observed in his appraisal of Mendel's work. Consult Pires and Branco [38] and Franklin [39] for more information on the famous Mendel-Fisher controversy.

If a reported p_k is the "best" of ℓ_k observed p-values of ℓ_k independent replications of an experiment, i.e., is the minimum of ℓ_k independent Uniform(0, 1) random variables, then $P_k \sim \text{Beta}(1, \ell_k)$, which has a probability density function $f_{P_k}(x; \ell_k) = \ell_k (1-x)^{\ell_k - 1} \mathbb{I}_{(0,1)}(x)$. Therefore, $-2\ell_k \ln(1 - P_k) \sim \chi_2^2$. This also holds true for the case $\ell_k = 1$, i.e., for genuine p-values, since $-2\ln P_k \overset{\mathrm{d}}{=} -2\ln(1-P_k) \sim \chi_2^2$ when $P_k \sim \text{Uniform}(0,1)$. So, the changes needed in Fisher's statistic are $T_F^*(P_1, \dots, P_n) = -2\sum_{k=1}^n \ell_k \ln(1-P_k)$, which under H_0^* is also χ_{2n}^2-distributed. However, the main problem here is that there is no information on whether some of the p-values are "fake ones", and if they do exist, which ones are and what are the corresponding values of ℓ_k.

Please note that what makes the most sense is to consider either $\ell_k = 1$ or $\ell_k = 2$ because it would be a complete waste of time and resources to continue replicating an experiment if non-significant p-values keep showing up, especially if there is the (wrong) belief that a p-value is only "a good one" if it is significant. It is, therefore, assumed that $\ell_k = 1$ when a genuine p-value is reported, regardless of whether it is significant or not. However, when some researchers are dissatisfied with obtaining non-significant p-values for their (first) results, they may decide not to report them and abandon their research, or repeat the experiment once ($\ell_k = 2$). In the latter case, one of the following scenarios takes place:

(a) the second p-value is significant, and hence it is the one reported (fake p-value);
(b) the second p-value is also not significant and consequently, either the smallest of the two observed p-values is reported (fake p-value), or none is reported and the research stops.

From the above, if $\ell_k = 2$, then clearly the right model for P_k is a mixture of the minimum of two independent Uniform$(0,1)$ random variables (or a Beta$(1,2)$ random variable) and a Uniform$(0,1)$ random variable, i.e., with probability density function

$$f_{P_k}(x; \mathfrak{p}) = (\mathfrak{p}\, 2(1-x) + (1-\mathfrak{p}))\mathbb{I}_{(0,1)}(x)\,,$$

where $0 \le \mathfrak{p} \le 1$, and which can be reparameterized as

$$f_{P_k}(x; m) = \left(m(1-x) + \left(1 - \frac{m}{2}\right)\right)\mathbb{I}_{(0,1)}(x)\,, \tag{1}$$

with $m = 2\mathfrak{p}$, $m \in [0,2]$. Therefore, in Equation (1), $\frac{m}{2}$ is the probability of a p-value being a fake p-value.

What is interesting to notice is that if the probability density function of the standard uniform distribution is tilted using the point $\left(\frac{1}{2}, 1\right)$ as a pole, then for $m \in [-2,2]$, the right-hand side of Equation (1) is still a probability density function, more specifically, the probability density function of a Mendel random variable $X_m \sim \text{Mendel}(m)$.

From Equation (1), it is straightforward to see that $X_0 \sim \text{Uniform}(0,1)$, $X_2 \sim \text{Beta}(1,2)$, and $X_{-2} \sim \text{Beta}(2,1)$, i.e., the maximum of two independent standard uniform random variables. Moreover, if $m \in (-2,0)$, then the Mendel distribution is a mixture of standard uniform distribution, with weight $1 + \frac{m}{2}$, and a $\text{Beta}(2,1)$ distribution, while if $m \in (0,2)$, it is a mixture of standard uniform distribution, with weight $1 - \frac{m}{2}$, and a $\text{Beta}(1,2)$ distribution. So, the probability density function of $X_m \sim \text{Mendel}(m)$, $m \in [-2,2]$, can be expressed in the form

$$f_{X_m}(x;m) = \frac{|m|}{2} f_{P_{i:2}}(x) + \left(1 - \frac{|m|}{2}\right) f_P(x)\,,$$

with $i = 1$ if $m \in (0,2]$, or $i = 2$ if $m \in [-2,0)$, and where $P_{1:2}$ and $P_{2:2}$ denote, respectively, the minimum and maximum of two independent standard uniform random variables, and $P \sim \text{Uniform}(0,1)$.

An interesting fact related to the Mendel distribution is that if X and Y are independent random variables, both with support $[0,1]$, and with $X \sim \text{Mendel}(m)$, then

$$V = \min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right) \sim \text{Mendel}((2\mathbb{E}[Y]-1)m)\,,$$

which generalizes Deng and George's [17] characterization of the standard uniform distribution when $X \sim \text{Uniform}(0,1)$ (see Theorem 1 in Brilhante et al. [10]). Furthermore, if $X \sim \text{Uniform}(0,1)$, then V and Y are independent random variables.

In particular, if X and Y are independent such that $X \sim \text{Mendel}(m_1)$ and $Y \sim \text{Mendel}(m_2)$, then

$$V = \min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right) \sim \text{Mendel}\left(\frac{m_1 m_2}{6}\right).$$

On the other hand, if $X \sim \text{Mendel}(m)$ and $Y \sim \text{Beta}(n,1)$ are independent, then

$$V = \min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right) \sim \text{Mendel}\left(m\,\frac{n-1}{n+1}\right), \tag{2}$$

while if $X \sim \text{Mendel}(m)$ and $Y \sim \text{Beta}(1,n)$ are independent, then

$$V = \min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right) \sim \text{Mendel}\left(m\,\frac{1-n}{1+n}\right). \tag{3}$$

Please note that Equations (2) or (3) can be used to test whether a sample of p-values $(p_1, \ldots, p_n)$ are observations from a $\text{Uniform}(0,1)$, a $\text{Mendel}(2)$, or a $\text{Mendel}(-2)$ distribution, being very useful to increase the test's power when the sample size is small (see Gomes et al. [40] and Brilhante et al. [41] for more details). For this purpose, setting $x_k = p_k$ and generating y_k, then $v_k = \min\left(\frac{x_k}{y_k}, \frac{1-x_k}{1-y_k}\right)$ is obtained, and therefore to test, for instance, the uniformity of the sample $(p_1, \ldots, p_n)$, one tests the uniformity of the pseudo-random sample $(p_1, \ldots, p_n, v_1, \ldots, v_n)$.

4. Combining Genuine and Fake p-Values

It is generally impossible to know whether there are or not fake p-values among the set of p-values to be combined. Therefore, a realistic approach is to examine possible scenarios and assess how the probable existence of fake p-values in a sample can affect the decision on the overall hypothesis H_0^*. For this purpose, tables with critical quantiles for p-values' combination methods that take into account the existence of fake p-values in a sample, most likely in a very small number, can be useful to give an overall picture.

Such tables are given in Brilhante et al.'s [10] supplementary materials for the most commonly used combined test statistics, where it is assumed that among the n ($n = 3, \ldots, 28$) p-values to be combined there are at most $n_f = 0, 1, \ldots, \max\{3, \lfloor n/3 \rfloor\}$ fake ones. The usefulness of the tables is illustrated with Example 1.

Example 1. *For the set of $n = 13$ p-values obtained in studies on depressive effects of a weekly 1mg dose of semaglutide*

$$\begin{array}{ccccccc} 0.0571 & 0.5954 & 0.0249 & 0.4793 & 0.1792 & 0.2917 & 0.6783 \\ 0.0554 & 0.2805 & 0.8137 & 0.2824 & 0.3338 & 0.1923 & \end{array}$$

the observed values for the combined test statistics are:

$$\begin{aligned} &T_F(0.0571, \ldots, 0.0.1923) = 39.0602 \\ &T_S(0.0571, \ldots, 0.1923) = -2.0842 \\ &T_{MG}(0.0571, \ldots, 0.1923) = 13.1940 \\ &T_{\mathcal{G}_{13}}(0.0571, \ldots, 0.1923) = 0.2226 \\ &T_{\min\{\mathcal{G}_{13}, \mathcal{G}_{13}^*\}}(0.0571, \ldots, 0.1923) = 0.2226 \\ &T_E(0.0571, \ldots, 0.1923) = 0.3280 \\ &T_T(0.0571, \ldots, 0.1923) = 0.0249 \end{aligned}$$

The quantiles for $n = 13$ are extracted from the tables in [10] (without the standard errors) for the following methods: Fisher (Table 1), Stouffer (Table 2), Mudholkar and George (Table 3), Pearson's geometric mean (Table 4), Pearson's minimum of geometric means (Table 5), Edgington's arithmetic mean (Table 6) and Tippett (Table 7).

The quantiles that lead to the rejection of H_0^* are highlighted for each method, thus showing for which significance level $\alpha \in \{0.005, 0.01, 0.025, 0.05, 0.1\}$ this happens.

Fisher's Statistic $T_F = 39.0602$

Table 1. Estimated quantiles of T_F with n_f fake p-values.

n	n_f	0.900	0.950	0.975	0.990	0.995
13	0	35.5632	38.8851	41.9232	45.6417	48.2899
13	1	36.6548	40.0294	43.0852	46.7821	49.4752
13	2	37.7241	41.1053	44.1240	47.9461	50.6022
13	3	38.8119	42.1576	45.2735	49.0533	51.7268
13	4	39.9069	43.2759	46.2994	50.1729	52.9071

Stouffer et al.'s Statistic $T_S = -2.0842$

Table 2. Estimated quantiles of T_S with n_f fake p-values.

n	n_f	0.900	0.950	0.975	0.990	0.995
13	0	1.2815	1.6448	1.9600	2.3264	2.5758
13	1	1.1087	1.4720	1.7844	2.1391	2.3767
13	2	0.9350	1.2924	1.6009	1.9524	2.1995
13	3	0.7620	1.1079	1.4117	1.7670	2.0049
13	4	0.5908	0.9312	1.2345	1.5756	1.8255

Mudholkar and George's Statistic $T_{MG} = 13.1940$

Table 3. Estimated quantiles of T_{MG} with n_f fake p-values.

n	n_f	**0.900**	**0.950**	**0.975**	**0.990**	**0.995**
13	0	8.3859	10.7850	12.8627	15.3892	17.1365
13	1	9.2840	11.6589	13.7337	16.2667	17.9027
13	2	10.1682	12.5187	14.5952	17.1134	18.8075
13	3	11.0523	13.3512	15.4344	17.9587	19.5983
13	4	11.9532	14.2848	16.2954	18.7587	20.4252

Pearson's Geometric Mean Statistic $T_{\mathcal{G}_n} = 0.2226$

Table 4. Estimated quantiles of $T_{\mathcal{G}_n}$ with n_f fake p-values.

n	n_f	**0.005**	**0.010**	**0.025**	**0.050**	**0.100**
13	0	0.15609	0.17283	0.19940	0.22412	0.25466
13	1	0.14919	0.16544	0.19070	0.21448	0.24420
13	2	0.14287	0.15820	0.18323	0.20578	0.23436
13	3	0.13684	0.15162	0.17531	0.19762	0.22476
13	4	0.13075	0.14522	0.16853	0.18930	0.21549

Pearson's Minimum of Geometric Means Statistic $T_{\min\{\mathcal{G}_n,\mathcal{G}_n^*\}} = 0.2226$

Table 5. Estimated quantiles of $T_{\min\{\mathcal{G}_n,\mathcal{G}_n^*\}}$ with n_f fake p-values.

n	n_f	**0.005**	**0.010**	**0.025**	**0.050**	**0.100**
13	0	0.14144	0.15578	0.17882	0.19940	0.22388
13	1	0.14177	0.15608	0.17876	0.19939	0.22400
13	2	0.13916	0.15370	0.17667	0.19710	0.22212
13	3	0.13536	0.14960	0.17235	0.19326	0.21799
13	4	0.13019	0.14438	0.16717	0.18746	0.21216

Edgington's Arithmetic Mean Statistic $T_E = 0.3280$

Table 6. Estimated quantiles of T_E with n_f fake p-values.

n	n_f	**0.005**	**0.010**	**0.025**	**0.050**	**0.100**
13	0	0.29609	0.31496	0.34333	0.36774	0.39629
13	1	0.28659	0.30475	0.33258	0.35682	0.38486
13	2	0.27780	0.29548	0.32267	0.34616	0.37356
13	3	0.26796	0.28591	0.31225	0.33557	0.36253
13	4	0.25868	0.27644	0.30163	0.32471	0.35112

Tippett's Minimum Statistic $T_T = 0.0249$

Table 7. Quantiles of T_T with n_f fake p-values.

n	n_f	**0.005**	**0.010**	**0.025**	**0.050**	**0.100**
13	0	0.00039	0.00077	0.00195	0.00394	0.00807
13	1	0.00036	0.00072	0.00181	0.00366	0.00750
13	2	0.00033	0.00067	0.00169	0.00341	0.00700
13	3	0.00031	0.00063	0.00158	0.00320	0.00656
13	4	0.00029	0.00059	0.00149	0.00301	0.00618

For this example, Fisher's method shows some stability when it comes to deciding on H_0^*, even when a small number of fake p-values can exist in the sample, and thus it seems robust to a prior choice of a significance level α (usually 0.05). The same can be said of Pearson's geometric mean method, which is, in fact, equivalent to Fisher's method. The runner-up is Mudholkar and George's method, which in traditional contexts has shown to be a compromise between Fisher's and Stouffer's methods. Please note that Stouffer's method, recommended in the social sciences, looks less reliable in this case. Clearly, Tippett's method should be avoided, despite being the simplest of them all and having a very uncomplicated sampling distribution for its statistic, even when n_f fake p-values exist, since $P_{1:n}|H_0^* \sim \text{Beta}(1, n+n_f)$.

This example reinforces, to some extent, the general belief that Fisher's combined test (or Pearson's equivalent geometric mean test) should be used, even in a wider context of jointly combining genuine and fake p-values. However, a more in-depth study is needed to support such a conclusion, but this is beyond the scope of this review paper.

5. Further Developments in Combining *p*-Values

There are many other modifications and generalizations of the classical test statistics for combining genuine P-values than those discussed in Section 2.

Fisher's statistic is the most widely used for combining p-values and has therefore been the subject of several generalizations, namely weighted versions. The discussion of conceptual advantages of weighting p-values, for instance, to improve the power of the combination method, goes as far as Good [42]. In regard to the weighted combination of independent probabilities, see also Bhoj [43]. As for the combination of dependent and weighted p-values, these are intertwined topics. Aside from the references Chuang and Shih [44], Hou [45], Makambi [46], and Yang [47], cf. for instance Alves and Yu [48].

Lancaster [49] generalized Fisher's method by transforming p-values using the chi-squared distribution with d_k degrees of freedom,

$$T_L(P_1,\ldots,P_n) = \sum_{k=1}^{n} F^{-1}_{\chi^2_{d_k}}(1-P_k),$$

where $F^{-1}_{\chi^2_{d_k}}$ is the inverse of the chi-square cumulative distribution function with d_k degrees of freedom, so that in an independent setup, $T_L|H_0^* \sim \chi^2_{\sum_{k=1}^n d_k}$. Chen's [50] numerical comparisons indicate that Lancaster's statistic T_L has a higher power than the traditional combination rules described in Section 2. Dai et al. [51] combined dependent P-values using approximations to the distribution of T_L, obtaining higher Bahadur efficiency than with a weighted version of the z-test.

Hou and Yang [52] developed a weighted version of Lancaster's statistic, namely

$$T_{HY}(P_1,\ldots,P_n) = \sum_{k=1}^{n} w_k\, F^{-1}_{\chi^2_{d_k}}(1-P_k).$$

Regardless of whether $P_1,\ldots,P_n$ are independent or not, $T_{HY} \approx c\chi^2_f$, and by equating expectations and variances, i.e., $\mathbb{E}(T_{HY}) = \mathbb{E}(c\chi^2_f) = cf$ and $\text{Var}(T_{HY}) = \text{Var}(c\chi^2_f) = 2c^2 f$, the parameter c can be estimated considering that

$$c = \frac{\text{Var}(T_{HY})}{2\mathbb{E}(T_{HY})} = \frac{\sum\limits_{k=1}^{n} w_k^2 d_k + \sum\limits_{k=1}^{n}\sum\limits_{j<k} w_k w_j \,\text{Cov}\left(F^{-1}_{\chi^2_{d_k}}(1-P_k), F^{-1}_{\chi^2_{d_j}}(1-P_j)\right)}{\sum\limits_{k=1}^{n} w_k d_k},$$

and the parameter f by considering

$$f = \frac{2\left[\mathbb{E}(T_{HY})\right]^2}{\mathrm{Var}(T_{HY})} = \frac{\left(\sum\limits_{k=1}^{n} w_k d_k\right)^2}{\sum\limits_{k=1}^{n} w_k^2 d_k + \sum\limits_{k=1}^{n}\sum\limits_{j<k} w_k w_j \,\mathrm{Cov}\left(F^{-1}_{\chi^2_{d_k}}(1-P_k), F^{-1}_{\chi^2_{d_j}}(1-P_j)\right)}.$$

It then follows that the $(1-\alpha)$100-th percentile of the distribution of $T_{HY}(P_1,\ldots,P_n)$ can be approximated by $cF^{-1}_{\chi^2_f}(1-\alpha)$.

Zhang and Wu [53] investigated a general family of Fisher's type of statistics referred to as the GFisher, which covers many classical statistics. Systematic simulations show that new p-value calculation methods based on moment-ratio matching and joint distribution surrogating are more accurate under the multivariate Gaussian and more robust under the generalized linear model and the multivariate t distribution. Relevant computation has been implemented in the R package GFisher, which is available in the Comprehensive R Archive Network.

The poolr package (Cinar and Viechtbauer [54]) provides an implementation of a variety of methods for combining p-values, including the inverse chi-square method (Liu [55]), a binomial test (Wilkinson [18]) and a Bonferroni/Holm method [56], which is an alternative to Simes' test [19]. Using an empirically derived null distribution based on pseudo-replicates that mimics a proper permutation test, an adjustment to account for dependence among the tests from which the p-values have been derived is made assuming multivariate normality among the test statistics. The poolr package has been compared with several other packages that can be used to combine p-values. Dewey's [57] metap v1.9 package provides an implementation of a wide variety of methods for combining independent p-values described in Becker [58].

Liu and Xie [59] suggested a statistic defined as a weighted sum of the Cauchy transformation of individual p-values, implying that the tail of the null distribution can be well approximated by a Cauchy distribution under arbitrary dependency structures. The p-value calculation of the test is accurate and as simple as the classical z-test or t-test, making it well suited for analyzing massive data. On the other hand, Ham and Park [60] showed that the Cauchy combination test provides the best combined p-value in the sense that it had the best performance among the examined methods while controlling type I error rates.

As the independence assumption is clearly a strong limitation when it comes to combining p-values, in 1975, Brown [61] discussed a method for combining non-independent tests of significance. The combination of p-values in correlated setups, for instance, in genome research requiring the analysis of Big Data, is currently a very active field of research, cf. Makambi [46], Hou [45], Yang [62], and Chuang and Shih [44]. In 2002, Kost and McDermott [63] derived an approximation to the null distribution of Fisher's statistic for combining p-values when the underlying test statistics are jointly distributed as a multivariate t with a common denominator.

As already mentioned, Fisher's statistic is the most used for combining p-values and generalizing it for dependence contexts has also been a constantly revisited research topic (see, for instance, Yang [47], Dai et al. [51] or Li et al. [64]). Chen [65] investigated new Gamma-based combination of p-values, based on the test statistic

$$T_{G(\alpha,1/\delta)}(P_1,\ldots,P_n) = \sum_{k=1}^{n} F^{-1}_{G(\alpha,1/\delta)}(1-P_k),$$

where $F^{-1}_{G(\alpha,1/\delta)}$ denotes the inverse of the Gamma cumulative distribution function with shape parameter α and scale parameter $1/\delta$, and showed that in many situations it provides an asymptotically Uniformly Most Powerful test.

Wilson [66] recommends the use of the harmonic mean p-value, i.e.,

$$T_{\mathcal{H}_n}(P_1, \dots, P_n) = \frac{n}{\sum_{k=1}^{n} 1/P_k},$$

for combining dependent p-values, since it controls the overall type I error, i.e., the probability of falsely rejecting the overall null hypothesis H_0^* in favor of at least one alternative hypothesis H_{Ak}. It is a complementary method to Fisher's method by averaging only valid p-values when these are mutually exclusive but not necessarily independent. The sampling distribution of $T_{\mathcal{H}_n}(P_1, \dots, P_n)$ is known to be in the domain of attraction of the heavy-tailed Landau skewed additive (1,1)-stable law, is robust to positive dependency between p-values and also to the distribution of the weights w used in its computation. Furthermore, it is insensitive to the number of tests and is mainly influenced by the smallest p-values.

Chien [67] compared the performances of Wilson's [66] harmonic mean method and of Kost and McDermott's [63] method to the performance of an empirical method based on the gamma distribution for combining dependent p-values from multiple hypothesis testing, which robustly controls the type I error and keeps a good power rate.

Based on recent developments in robust risk aggregation techniques, Vovk and Wang [68] by combining a number of p-values without making any assumption about their dependence structure, extended those results to generalized means, and showed that n p-values can be combined by scaling up their harmonic mean by a factor of $\ln n$.

E-values, defined as expectations, in contrast to p-values, defined as probabilities, are nonnegative random variables whose expected values under the null hypothesis are bounded by 1 (Shafer et al. [69]), as in Bayes factors and likelihood ratios in the case of a simple null hypothesis (Grünwald et al. [70]; Shafer et al. [69]). The combination of e-values via e-merging functions is a more recent and active field of research (cf. Grünwald et al. [70], Shafer [71], Vovk et al. [72,73], and Vuursteen et al. [74]). For instance, the product of independent e-values is clearly an e-value. However, so far, little is known about the power of these combination procedures, although this is now the main focus of research in this field.

6. Conclusions

The meta-analysis of p-values poses some challenges, especially in today's world in which academic and scientific achievements are largely measured (and funded) by the number of papers published, thus putting much pressure on researchers. For this reason, possibly some—but almost certainly a very few—of the P_k's, $k = 1, \dots, n$, to be used in a statistic $T(P_1, \dots, P_n)$ are fake p-values (minimum of Two P), when in an honest world, they should all be genuine p-values (not Two P). Therefore, it is a good idea to perform a comparison between the conclusions drawn from different combined tests, assuming that among the observed p_k's there are $n_f = 0, 1, \dots, j \ll n$ fake p-values, to ensure a more informed decision on the overall hypothesis.

The tables with quantiles of the most used methods for combining p-values that take into consideration the existence of a small number of fake p-values in a sample, obtained by the authors and provided in Brilhante et al. [10], can be a useful tool to assess the reliability of the conclusions drawn from meta-analyses of p-values in the event of their unknown presence.

Author Contributions: Conceptualization, M.F.B., M.I.G., S.M., D.P. and R.S.; Funding acquisition, M.I.G.; Investigation, M.F.B., M.I.G., S.M., D.P. and R.S.; Methodology, M.F.B., M.I.G., S.M., D.P. and R.S.; Project administration, M.F.B., D.P. and R.S.; Resources, M.F.B., M.I.G. and R.S.; Software, M.F.B. and R.S.; Supervision, M.F.B., M.I.G. and D.P.; Writing—original draft, M.F.B., M.I.G., S.M., D.P. and R.S.; Writing—review and editing, M.F.B., M.I.G., S.M., D.P. and R.S. All authors have read and agreed to the published version of the manuscript.

Funding: Research partially financed by national funds through FCT—Fundação para a Ciência e a Tecnologia, Portugal, under the project UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1900**, *50*, 157–175. [CrossRef]
2. Arbuthnot, J. An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philos. Trans. R. Soc. Lond.* **1710**, *27*, 186–190. [CrossRef]
3. Fisher, R.A. On the interpretation of χ^2 from contingency tables, and the calculation of *P. J. R. Stat. Soc.* **1922**, *85*, 87–94. [CrossRef]
4. Fisher, R.A. *Statistical Methods for Research Workers*; Oliver and Boyd: Edinburgh, UK, 1925.
5. Fisher, R.A. The arrangement of field experiments. *J. Minist. Agric.* **1926**, *33*, 503–515.
6. Greenwald, A.G.; Gonzalez, R.; Harris, R.J.; Guthrie, D. Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology* **1996**, *33*, 175–183. [CrossRef]
7. Colquhoun, D. The reproducibility of research and the misinterpretation of p-values. *R. Soc. Open Sci.* **2017**, *4*, 171085. [CrossRef]
8. Tippett, L.H.C. *The Methods of Statistics*; Williams & Norgate: London, UK, 1931. [CrossRef]
9. Fisher, R.A. *Statistical Methods for Research Workers*, 4th ed.; Oliver and Boyd: London, UK, 1932.
10. Brilhante, M.F.; Gomes, M.I.; Mendonça, S.; Pestana, D.; Santos, R. Meta-analysis of genuine and fake p-values. *Preprints* **2024**. [CrossRef]
11. Wasserstein, R.L.; Lazar, N.A. The Asa statement on p-values: Context process, and purpose. *Am. Stat.* **2016**, *70*, 129–133. [CrossRef]
12. Wasserstein, R.L.; Schirm, A.L.; Lazar, N.A. Moving to a world beyond "$p < 0.05$". *Am. Stat.* **2019**, *73*, 129–133. [CrossRef]
13. Jin, Z.C.; Zhou, X.H.; He, J. Statistical methods for dealing with publication bias in meta-analysis. *Stat. Med.* **2015**, *34*, 343–360. [CrossRef]
14. Lin, L.; Chu, H. Quantifying publication bias in meta-analysis. *Biometrics* **2018**, *74*, 785–794. [CrossRef] [PubMed]
15. Givens, G.H.; Smith, D.D.; Tweedie, R.L. Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Stat. Sci.* **1997**, *12*, 221–250. [CrossRef]
16. Fisher, R.A. Has Mendel's work been rediscovered? *Ann. Sci.* **1936**, *1*, 115–137. [CrossRef]
17. Deng, L.Y.; George, E.O. Some characterizations of the uniform distribution with applications to random number generation. *Ann. Inst. Stat. Math.* **1992**, *44*, 379–385. [CrossRef]
18. Wilkinson, B. A statistical consideration in psychological research. *Psychol. Bull.* **1951**, *48*, 156–158. [CrossRef] [PubMed]
19. Simes, R.J. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **1986**, *73*, 751–754. [CrossRef]
20. Edgington, E.S. An additive method for combining probability values from independent experiments. *J. Psychol.* **1972**, *80*, 351–363. [CrossRef]
21. Pearson, K. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika* **1933**, *25*, 379–410. [CrossRef]
22. Brilhante, M.F.; Gomes, M.I.; Mendonça, S.; Pestana, D.; Pestana, P. Generalized Beta models and population growth, so many routes to chaos. *Fractal Fract.* **2023**, *7*, 194. [CrossRef]
23. Pearson, K. On a new method of determining "goodness of fit". *Biometrika* **1934**, *26*, 425–442. [CrossRef]
24. Owen, A.B. Karl Pearson's meta-analysis revisited. *Ann. Stat.* **2009**, *37*, 3867–3892. [CrossRef] [PubMed]
25. Stouffer, S.A.; Schuman, E.A.; DeVinney, L.C.; Star, S.; Williams, R.M. *The American Soldier: Adjustment during Army Life*; Princeton University Press: Princeton, NJ, USA, 1949; Volume I. [CrossRef]
26. Mudholkar, G.S.; George, E.O. The logit method for combining probabilities. In *Symposium on Optimizing Methods in Statistics*; Rustagi, J., Ed.; Academic Press: New York, NY, USA, 1979; pp. 345–366.
27. Birnbaum, A. Combining independent tests of significance. *J. Am. Stat. Assoc.* **1954**, *49*, 559–574. [CrossRef]
28. Mosteller, F.; Bush, R. Selected quantitative techniques. In *Handbook of Social Psychology: Theory and Methods*; Lidsey, G., Ed.; Addison-Wesley: Cambridge, MA, USA, 1954.
29. Littell, R.C.; Folks, L.J. Asymptotic optimality of Fisher's method of combining independent tests, I. *J. Am. Stat. Assoc.* **1971**, *66*, 802–806. [CrossRef]
30. Littell, R.C.; Folks, L.J. Asymptotic optimality of Fisher's method of combining independent tests, II. *J. Am. Stat. Assoc.* **1973**, *68*, 193–194. [CrossRef]
31. Loughin, T.M. A systematic comparison of methods for combining p-values from independent tests. *Comput. Stat. Data Anal.* **2004**, *47*, 467–485. [CrossRef]
32. Hartung, J.; Knapp, G.; Sinha, B.K. *Statistical Meta-Analysis with Applications*; Wiley: Hoboken, NJ, USA, 2008. [CrossRef]

33. Kulinskaya, E.; Morgenthaler, S.; Staudte, R.G. *Meta Analysis. A Guide to Calibrating and Combining Statistical Evidence*; Wiley: Chichester, UK, 2008. [CrossRef]
34. Tsui, K.; Weerahandi, S. Generalized p-values in significance testing of hypothesis in the presence of nuisance parameters. *J. Am. Stat. Assoc.* **1989**, *84*, 602–607. [CrossRef]
35. Weerahandi, S. *Exact Statistical Methods for Data Analysis*; Springer: New York, NY, USA, 1995. [CrossRef]
36. Hung, H.; O'Neill, R.; Bauer, P.; Kohn, K. The behavior of the p-value when the alternative is true. *Biometrics* **1997**, *53*, 11–22. [CrossRef] [PubMed]
37. Brilhante, M.F. Generalized p-values and random p-values when the alternative to uniformity is a mixture of a Beta(1,2) and uniform. In *Recent Developments in Modeling and Applications in Statistics*; Oliveira, P., Temido, M., Henriques, C., Vichi, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 159–167. [CrossRef]
38. Pires, A.M.; Branco, J.A. A statistical model to explain the Mendel-Fisher controversy. *Stat. Sci.* **2010**, *25*, 545–565. [CrossRef]
39. Franklin, A.; Edwards, A.W.; Fairbanks, D.J.; Hartl, D.L. *Ending the Mendel-Fisher Controversy*; University of Pittsburgh Press: Pittsburgh, PA, USA, 2008. [CrossRef]
40. Gomes, M.I.; Pestana, D.; Sequeira, F.; Mendonça, S.; Velosa, S. Uniformity of offsprings from uniform and non-uniform parents. In Proceedings of the ITI 2009, 31st International Conference on Information Technology Interfaces, Cavtat/Dubrovnik, Croatia, 22–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 243–248.
41. Brilhante, M.; Pestana, D.; Sequeira, F. Combining p-values and random p-values. In Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, Cavtat/Dubrovnik, Croatia, 21–24 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 515–520.
42. Good, I.J. On the weighted combination of significance tests. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1955**, *17*, 264–265. [CrossRef]
43. Bhoj, D.S. On the distribution of the weighted combination of independent probabilities. *Stat. Probab. Lett.* **1992**, *15*, 37–40. [CrossRef]
44. Chuang, L.L.; Shih, Y.S. Approximated distributions of the weighted sum of correlated chi-squared random variables. *J. Stat. Plan. Inference* **2012**, *142*, 457–472. [CrossRef]
45. Hou, C.D. A simple approximation for the distribution of the weighted combination of non-independent or independent probabilities. *Stat. Probab. Lett.* **2005**, *73*, 179–187. [CrossRef]
46. Makambi, K.H. Weighted inverse chi-square method for correlated significance tests. *J. Appl. Stat.* **2003**, *30*, 225–234. [CrossRef]
47. Yang, T.S. A New Weighted Combination Procedure. Master's Thesis, Fu Jen Catholic University, Taipei, Taiwan, 2012.
48. Alves, G.; Yu, Y.K. Combining independent weighted P-values: Achieving computational stability by a systematic expansion with controllable accuracy. *PLoS ONE* **2011**, *6*, e22647. [CrossRef]
49. Lancaster, H. The combination of probabilities: An application of orthonormal functions. *Aust. J. Stat.* **1961**, *3*, 20–33. [CrossRef]
50. Chen, Z. Is the weighted z-test the best method for combining probabilities from independent tests? *J. Evol. Biol.* **2011**, *24*, 926–930. [CrossRef]
51. Dai, H.; Leeder, J.S.; Cui, Y. A modified generalized Fisher method for combining probabilities from dependent tests. *Front. Genet.* **2014**, *5*, 32. [CrossRef]
52. Hou, C.D.; Yang, T.S. Distribution of weighted Lancaster's statistic for combining independent or dependent P-values, with applications to human genetic studies. *Commun. Stat. Theory Methods* **2023**, *52*, 7442–7454. [CrossRef]
53. Zhang, H.; Wu, Z. The generalized Fisher's combination and accurate p-value calculation under dependence. *Biometrics* **2022**, *79*, 1159–1172. [CrossRef] [PubMed]
54. Cinar, O.; Viechtbauer, W. The poolr package for combining independent and dependent p values. *J. Stat. Softw.* **2022**, *101*, 1–42. [CrossRef]
55. Liu, J.Z.; Mcrae, A.F.; Nyholt, D.R.; Medland, S.E.; Wray, N.R.; Brown, K.M.; AMFS Investigators; Hayward, N.K.; Montgomery, G.W.; Vissher, P.M.; et al. A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **2010**, *87*, 139–145. [CrossRef]
56. Holm, S. A simple sequentially multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
57. Dewey, M. *metap: Meta-Analysis of Significance Values*, R Package Version 1.9. RDocumentation. Available online: https://www.rdocumentation.org/packages/metap/versions/1.9 (accessed on 2 September 2024).
58. Becker, B.J. Combining significance levels. In *The Handbook of Research Synthesis*; Cooper, H., Hedges, L.V., Eds.; Russell Sage Foundation: New York, NY, USA, 1994; pp. 215–230.
59. Liu, Y.; Xie, J. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* **2020**, *115*, 393–402. [CrossRef]
60. Ham, H.; Park, T. Combining p-values from various statistical methods for microbiome data. *Front. Microbiol.* **2022**, *13*, 990870. [CrossRef]
61. Brown, M.B. A method for combining non-independent, one-sided tests of significance. *Biometrics* **1975**, *3*, 987–992. [CrossRef]
62. Yang, J.J. Distribution of Fisher's combination statistic when the tests are dependent. *J. Stat. Comput. Simul.* **2010**, *80*, 1–12. [CrossRef]
63. Kost, J.T.; McDermott, M.P. Combining dependent p-values. *Stat. Probab. Lett.* **2002**, *60*, 183–190. [CrossRef]
64. Li, Q.; Hu, J.; Ding, J.; Zheng, G. Fisher's method of combining dependent statistics using generalizations of the gamma distribution with applications to genetic pleiotropic associations. *Biostatistics* **2014**, *15*, 284–295. [CrossRef]

65. Chen, Z. Optimal tests for combining p-values. *Appl. Sci.* **2022**, *12*, 322. [CrossRef]
66. Wilson, D.J. The harmonic mean p-value for combining dependent tests. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 1195–1200. [CrossRef] [PubMed]
67. Chien, L.C. Combining dependent p-values by gamma distributions. *Stat. Appl. Genet. Mol. Biol.* **2020**, *19*, 20190057. [CrossRef] [PubMed]
68. Vovk, V.; Wang, R. Combining p-values via averaging. *Biometrika* **2020**, *107*, 791–808. [CrossRef]
69. Shafer, G.; Shen, A.; Vereshchagin, N.; Vovk, V. Test martingales, Bayes factors and p-values. *Stat. Sci.* **2011**, *26*, 84–101. [CrossRef]
70. Grünwald, P.; De Heide, R.; Koolen, W.M. Safe testing. Information Theory and Applications Workshop (ITA). *J. R. Stat. Soc. Ser. B* **2020**, 1–54. [CrossRef]
71. Shafer, G. Testing by betting: A strategy for statistical and scientific communication. *J. R. Stat. Soc. Ser. A (Stat. Soc.)* **2021**, *184*, 407–431. [CrossRef]
72. Vovk, V.; Wang, R. E-values: Calibration, combination and applications. *Ann. Stat.* **2021**, *49*, 1736–1754. [CrossRef]
73. Vovk, V.; Wang, B.; Wang, R. Admissible ways of merging p-values under arbitrary dependence. *Ann. Stat.* **2022**, *50*, 351–375. [CrossRef]
74. Vuursteen, L.; Szabó, B.; van der Vaart, A.; van Zanten, H. Optimal testing using combined test statistics across independent studies. In Proceedings of the Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.