Gamma Approximation of Stratified Truncated Exact test (GASTE-test) & Application

Alexandre WENDLING[1] and Clovis GALIEZ[1]

[1]**Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France**

Corresponding author:
Alexandre WENDLING[1]

Email address: alexandre.wendling@univ-grenoble-alpes.fr

ABSTRACT

The analysis of categorical data, particularly the study of associations between binary outcomes and binary features, is crucial across various scientific disciplines, such as assessing the impact of vaccination on health outcomes. Traditional 2x2 contingency tables are commonly used to summarize binary counts; however, these analyses can be confounded by external factors like age or gender, necessitating stratification to create sub-tables. Stratified analysis is prevalent in medical, epidemiological, and social research, as well as in meta-analyses. Current methodologies for testing associations across strata struggle with small sample sizes and heterogeneity of the effect among strata. To cope with these limitations, exact tests can be used, but at a very high computational cost, preventing their use in most situations. Here, we propose the Gamma Approximation of Stratified Truncated Exact (GASTE) test as a robust alternative. The core of this paper presents a method for approximating the exact statistic of combination of p-values with discrete support, leveraging the gamma distribution to approximate the distribution of the test statistic under stratification. We show that this approximation maintains a high test power while keeping a low level of type I error. The GASTE method provides fast and accurate p-value calculations even in the presence of homogeneous and heterogeneous effects between strata, and is robust in scenarios with varying levels of significance. Our findings demonstrate that the GASTE test outperforms traditional methods, offering more sensitive and reliable detections. This advancement not only enhances the robustness of stratified analyses, but thanks to its fast computation also broadens the applicability of exact tests in various research fields. Firstly, we illustrate our method through the ecological application that motivated its development, consisting of the study of Alpine plant associations. Secondly, we apply our method to a well-known case study of stratified binary data, concerning admissions to the University of California at Berkeley in 1973. Overall, the GASTE method is a powerful and flexible tool for researchers dealing with stratified binary data, offering substantial improvements over traditional methods such as the CMH (Cochran-Mantel-Haenszel) test. An open-source python package is provided at https://github.com/AlexandreWen/gaste.

INTRODUCTION

Categorical data analysis is ubiquitous across various scientific fields, particularly in examining binary outcomes in relation to binary features, such as the influence of a vaccination on sickness of individuals. The objective often entails testing whether the probability of outcomes (sick or healthy) differs between two groups (vaccinated or not). To investigate the association between these features and outcomes, 2x2 contingency table is frequently employed to summarize binary counts within a sample. However, observations may be influenced by external factors. To address this, stratification is utilized to create sub-tables based on these confounding variables, such as age or gender. This approach of multilevel analysis is frequently employed in medical studies, epidemiology, social research, and meta-analyses (Agresti, 2012; Smaardijk et al., 2020; Reilly, 2023).

To test the overall association between binary features and outcomes across several strata, one way is to calculate a combined effect given by the weighted average of some summary statistic across all studies. Two models can be considered for assigning weights (Borenstein

et al., 2010), a fixed effect model assumes that all strata are estimating the same underlying effect size (the strength of association), and a random effect model accounts for variability both within and between studies. The latter assumes that the true effect size varies among studies, and the observed effect sizes reflect this distribution. Four methods are widely used, three fixed-effect methods, namely Mantel-Haenszel (Mantel and Haenszel, 1959), Peto (Yusuf et al., 1985) and inverse variance; and one random-effects method, DerSimonian and Laird inverse variance (DerSimonian and Laird, 1986). These methods employ asymptotic tests hampered by small counts and do not cope with heterogeneity (uneven strength of association between feature and outcome across strata) even for random effect models due to averaging over strata. The chapter 10 of Cochran handbook gives useful insight and clear summary on the topic (Deeks JJ, 2023). The indispensable stratification carried out in studies to manage confounding variables increases the number of combinations of subgroup categories and creates problems known as *combination fatigue*. These include concerns about small sample size, reduced power, spurious results and complexity of statistical design and analysis (Evans, 2024). However, as we have mentioned, despite their widespread use, the aforementioned methods may lack robustness and power in these configurations. This is particularly true for the Mantel-Haenszel test (Mantel and Haenszel, 1959) also called Cochran-Mantel-Haenszel (CMH) test (Cochran, 1954), despite its very popular use, which is sometimes misapplied (Berger et al., 2006).

Leveraging many strata without being constrained by rare events (low cell frequency) or heterogeneity of association, calls for exact testing. In essence, an exact test aims at determining the probability of observing a specific stratified set of tables under a null model stating that the outcomes (e.g. sick or healthy) are indeed independent of the features (vaccinated or not). Given the number of samples in each table, there exists a finite number of possibilities for different stratified tables, allowing for the calculation of the exact probability of the observed data. No assumption of homogeneity or sufficient counts are henceforth needed. While Jung (2014) proposed an exact test based on this principle, its computational complexity grows exponentially with the number of strata and the number of samples in each stratum, rendering the test intractable for many scenarios.

In this paper, we first reformulate the exact test proposed by Jung (2014) as a combination of independent one-tailed Fisher exact tests for each stratum's p-value. To address the complexity issue, drawing inspiration from the combination of dependent p-values (Brown, 1975; Kost and McDermott, 2002; Yang, 2010; Zheng et al., 2012; Li et al., 2014), we propose an approximation of this p-value combination using a gamma distribution. We infer the gamma distribution parameters by matching its moments with the analytical moments of the exact distribution. In contrast to asymptotic approximations, directly approximating the exact test allows handling cases of rare events, particularly in the case of a high number of strata leading to a high complexity of exact calculation. We call our test method GASTE-test for Gamma Approximation of Stratified Truncated Exact test. Moreover, this approximation allows us to reduce the computational complexity by more than three orders of magnitude and makes the exact test, by its approximation, more practical for real-world applications. We demonstrate by simulation that this approximation remains accurate and preserves high test power, 16% more powerful than CMH test on average when the effect is applied to all strata and 58% more powerful when the effect is applied to only half the strata, while controlling the type I error level.

We also illustrate the efficacy of our method through real-world applications, comparing it to other statistical tests for assessing the overall significance between outcomes and features in stratified 2x2 contingency tables. The initial application of our test was in the field of ecology, which spurred its development. We had multiple observation zones in the French Alps, and within each zone, points were recorded for the presence or absence of various plant species. For a given pair of plants, we constructed stratified tables by environment to study plant associations by counting coexistence. This application was particularly interesting because the number of survey points was relatively low (≤ 300) and not controlled, and within these observation zones, species could be either rare or very abundant. GASTE-test method stands out from the rest by detecting twice as many associations as the CMH test, and handles the heterogeneity of effects between strata, which is crucial in this context. The second application is a well-known case study of stratified binary data. It puts at stake whether the admissions to the University of

California at Berkeley in 1973 are gender biased. When aggregated, the data appears to show a selection bias favoring men. However, when stratified by department, the usual tests give information of heterogeneity of association between strata and surprisingly no overall association between gender and department. GASTE-test method gives a different result.

METHODS

Stratified 2x2 contingency table

When dealing with observations of individuals with binary features (e.g. vaccinated or not), associated with binary outcome (e.g. sick or healthy), it is convenient to represent the data as a 2x2 contingency table. Observations may be influenced by some external factors, to control these confounding variables, stratification can be used to create sub-tables based on these factors. Assuming S strata $s = (1, \ldots, S)$, for a stratum s, the Table 1 contains the following data: k_s number of positive outcomes with the feature, K_s total number of positive outcomes, n_s total number of features and N_s the sample size.

	Features		
Outcomes	Yes	No	Total
Positive	k_s	$K_s - k_s$	K_s
Negative	$n_s - k_s$	$N_s - n_s - K_s + k_s$	$N_s - K_s$
Total	n_s	$N_s - n_s$	N_s

Table 1. 2x2 contingency table for stratum s

Fisher exact test

For a stratum s, the ratio $\theta_s = \frac{k_s/(K_s-k_s)}{(n_s-k_s)/(N_s-n_s-K_s+k_s)}$, called odd-ratio, is used to test the association between features and outcomes. When the outcome is independent of the features, the ratio θ_s is close to 1. Indeed, if the outcomes are not affected by the presence or absence of features, i.e. a positive outcome appear randomly with feature or not, the ratio of positive outcomes in the case of feature to those in the case of a non-feature must be similar to the ratio of negative outcomes between feature and non-features.

So in this case, with fixed marginals n_s, K_s and fixed sample size N_s (conditional test), the random variable X_s representing the number of individuals of both positive feature and outcome follows a hypergeometric distribution (HG). Formally, the probability of observing k_s positive outcomes with the feature is given by $P(X_s = k_s | N_s, n_s, K_s) = \frac{\binom{K_s}{k_s}\binom{N_s-K_s}{n_s-k_s}}{\binom{N_s}{n_s}}$, that is also the exact probability to observe a specific table with fixed marginals. The support of the distribution is discrete and finite and is given by $\text{supp}(X_s) = [\![\max(0, n_s + K_s - N_s), \min(K_s, n_s)]\!]$. To test the association between features and outcomes, two cases are taken into account : the case where k_s are observed less than expected ($E[X_s]$) (resp. more than expected), it is an under-association (resp. over-association) of positive outcomes with the feature.

Therefore, two tests can be constructed, respectively,

$$H^-_{0_s} : \theta_s \geq 1 \text{ against } H^-_{1_s} : \theta_s < 1$$

and

$$H^+_{0_s} : \theta_s \leq 1 \text{ against } H^+_{1_s} : \theta_s > 1$$

where the p-value of the test is call respectively $p^-_s := P(X_s \leq k_s)$ and $p^+_s := P(X_s \geq k_s)$, that are respectively the left and right tail of the distribution X_s. The non-strict inequality is very important in order to take into account the different table possibilities up to the one observed. It is worth to be noted that $p^+_s \neq 1 - p^-_s$, the fact of not being under-associated does not necessarily imply the fact of being over-associated. Finally, the one-tailed Fisher's exact test (Fisher et al., 1936) has just been recalled here, also known as the hypergeometric test. The

exact probability to observe one single table has been described, now the way to compute the exact probability to observe a set of table will be seen.

Stratified Truncated Fisher exact test

To assess the overall association between features and outcomes across different strata, as the Fisher exact test describe the exact probability to observe one table, we can describe the exact probability of observing a set of tables. In this approach, the p-values of each stratum are combined assuming independence between strata, achieved by multiplying the p-values. The product may not be very pleasant to handle, which is why it is preferable to use its monotonic logarithmic transformation.

$$T_1 = \prod_{s=1}^{S} P_s \quad \text{and} \quad Y_1 = -2\log(T_1) = -2\sum_{s=1}^{S}\log(P_s)$$

where P_s is the random variable associated to the p-value of each independent test to be combined. Typically, this method is employed when combining p-values from tests with continuous distributions. In this case, so when the p-values follow a uniform distribution, the combination is the classic case of combination by Fisher's method (Fisher et al., 1936), where the combination Y_1 follows a chi-squared distribution χ^2 with $2S$ degrees of freedom.

Additionally, inspired by Zaykin et al. (2002) and Zhang et al. (2018), to increase the statistical power of the test a truncated version of the Fisher's method can be used to combine only the p-values that contribute a significant amount of information to the overall signal, taking into account that some p-values are discarded. To preserve a certain softness and re-spread the retained p-values of the combination between 0 and 1, only the amount of signal below the truncation threshold τ is taken into account. The weighted truncated combination is given by :

$$T_\tau = \prod_{s=1}^{S}\left(\frac{P_s}{\tau}\right)^{\mathbb{I}(P_s\leq\tau)} \quad \text{and} \quad Y_\tau = -2\log(T_\tau) = -2\sum_{s=1}^{S}(\log(P_s)-\log(\tau))\,\mathbb{I}(P_s\leq\tau)$$

where $\mathbb{I}(\cdot)$ is the indicator function. Always in the case where the p-values are uniformly distributed, Zhang et al. (2018) gives the expression of the survival function of Y_τ for any $\tau \in [0,1]$.

When the p-values P_s come from the Fisher exact test of each table. So the P_s do not therefore follow a uniform distribution but a discrete distribution resulting from the hypergeometric test, which may be quite different from the distributions in the continuous case, also call sub-uniform distribution. The core of this paper presents a method for combining sub-uniform p-values.

Given the combination Y_τ has a discrete distribution, the finite set of values taken by Y_τ can be explored to calculate the combined p-value exactly through an analytic solution. As before, there are two cases for the two one-sided p-values of each test. To study the overall under-association (resp. over-association) of positive outcomes with the feature, the p-values from each under-association (resp. over-association) test are combined. Combining the p-values involves studying the intersection of all the null sub-hypotheses.

For the overall under-association, the hypotheses are :

$$H_0^- = \cap H_{0_s}^- : \text{"}\forall s, \theta_s \geq 1\text{" against } H_1^- : \text{"}\theta_s < 1 \text{ for at least one stratum } s\text{"}$$

Similarly, for the overall over-association :

$$H_0^+ = \cap H_{0_s}^+ : \text{"}\forall s, \theta_s \leq 1\text{" against } H_1^+ : \text{"}\theta_s > 1 \text{ for at least one stratum } s\text{"}$$

Further details on the exact calculation of the combined p value as well as the consideration of numerical problems and cases of the same combination value for different realizations are provided in Supplementary Materials 1.1 and Figure S1.

As mentioned by Jung (2014); Chiba (2017), the exact approach is computationally heavy. Indeed, the set of all the possible values taken by Y_τ has to be iterated through to take into account only realizations that are greater than the observed one. The size of this set is equal to the product

of the size of the supports of each hypergeometric distribution. The complexity of the exact calculation therefore increases exponentially with the number of strata $O\left(\prod_{s=1}^{S} |\text{supp}(X_s)|\right)$. In some applications, the order of magnitude is $10^6 - 10^7$, and it is perfectly possible to explore all the possible combinations of the different strata and therefore obtain an exact calculation of the overall significance across the strata. Our first contribution is a tool for computing this exact calculation. However, in other contexts, while maintaining the discrete nature of the p-values to be combined, the set of possible values may be excessively large, and it is impossible to calculate an exact value. In this case, instead of using an asymptotic test, which is a family of tests widely used in the literature, it is preferable to try to approximate the statistic of the exact test.

Gamma approximation

Given the potentially high computational cost of exact computation for the p-value of the stratified Fisher combination in discrete cases, we propose an approximation of the exact distribution followed by the statistic Y_τ. As mentioned in the previous section, the approximation of Y_1 by a χ^2_{2S} as in the case of uniform distribution on [0,1] of p-value is well known as the Fisher's method. However, through simulation, we observe that the distribution resulting from a combination of p-values from Fisher's exact test differs slightly from a χ^2 distribution in terms of scale and amplitude. Since the Gamma distribution Γ is a generalization of the χ^2 distribution ($\chi^2_{2n} = \Gamma(n, 1/2)$), we approximate the discrete combination by finding appropriate parameters α and β for the $\Gamma(\alpha, \beta)$ distribution. This type of method has already been used in the literature to combine continuous dependent p-values (Brown, 1975; Kost and McDermott, 2002; Li et al., 2014; Poole et al., 2016; Yang et al., 2016; Zhang and Wu, 2023), but to our knowledge, has never been used for combinations of independent discrete tests. The approximation is therefore as follows:

$$Y_1 = -2\sum_{s=1}^{S} \log(P_s) \approx Z = \Gamma(\alpha, \beta)$$

The inference of α and β is done by matching the moments of Y_1 and Z by analytically calculating the moments of the exact law of combination. Different moments can be chosen to infer α and β, but higher moments do not systematically involve a better approximation. In the Supplementary Materials 1.6 with the Figure **??**, we show, using simulations with random margins ($\sim U[1, N_s]$), that the gamma approximation asymptotically converges to the chi-squared distribution when the sample size N_s increases.

For the case where $\tau < 1$, due to our truncation, the approximation of the truncated law of combination Y_τ is in no longer a Γ distribution but a mixture that we describe in Supplementary Materials. More details on the choice of moments are given in Discussion section and details of the computation are also given in the Supplementary Materials 1.2.

The complexity of the gamma approximation is much lower than the exact calculation, since the exact moments of each stratum statistic $log(P_s)$ can be analytically calculated, and linearly combined to get the moments of the global statistic Y_τ (see Supplementary Materials 1.2 for further details). The complexity is therefore linear in the number of strata $O\left(\sum_{s=1}^{S} |\text{supp}(X_s)|\right)$.

Comparison to other statistical tests

Other statistical tests for assessing the overall significance between outcomes and features in stratified 2x2 contingency tables are documented in the literature. These tests vary in sensitivity and the conditions under which they are applied. We will compare our GASTE-test method to the following popular alternatives:

- Bonferroni's procedure (Bonferroni, 1936) controls for the combination $H_0 = \cap_{s=1}^{S} H_{0_s}$. At a threshold α, H_0 is rejected if $\min_{s\in(1,\ldots,S)}(p_s) \leq \frac{\alpha}{S}$. While this procedure tends to be conservative, it exhibits power in the presence of a small number of strata (down to a single one) with strong effect.

- A test widely used in meta-analyses and often considered as a reference method for stratified 2x2 contingency tables is the Cohran-Mantel-Haenszel test (CMH test) (Cochran, 1954;

Mantel and Haenszel, 1959). Intermediate tests are not performed on each stratum; rather, data from all tables are used to construct a single test statistic based on the hypergeometric distribution. The test statistic is given by:

$$\chi^2_{CMH} = \frac{\left(\sum_{s=1}^{S} k_s - E[X_s]\right)^2}{\sum_{s=1}^{S} Var[X_s]} = \frac{\left(\sum_{s=1}^{S} (k_s - \frac{n_s K_s}{N_s})\right)^2}{\sum_{s=1}^{S} \frac{n_s K_s (N_s - K_s)(N_s - n_s)}{N_s^2 (N_s - 1)}}$$

While this statistic rigorously follows a discrete distribution based on combinations of hypergeometric distributions, it is considered to asymptotically follow a χ^2 distribution with 1 degree of freedom. However, the CMH test assumes homogeneity of the effect of the association between features and outcomes across strata. Furthermore, opposite effect in different strata may cancel out, potentially escaping detection. Therefore, the CMH test must be accompanied by the Breslow-Day test (Breslow et al., 1980) to verify this homogeneity assumption. Given its widespread use, comparing this method with our own, which does not rely on this homogeneity assumption, could be insightful.

- Classical Fisher's p-value combination (Fisher et al., 1936) assumes uniformly distributed continuous p-values for each stratum. It has long been known that the Fisher combination is therefore inappropriate and can lead to false results if applied in discrete cases (Kincaid, 1962; Mielke Jr et al., 2004). Applied to hypergeometric tests, one thus ignores the discrete nature of the p-values. The χ^2_{2S} distribution followed by the combined statistic Y_τ in case of uniform p-values, will be compared to the exact and approximate distribution by GASTE to assess its deviation.

RESULTS

Importance of exact calculation

In this section, we compare under different scenarios the exact distribution (analytically computed) of the statistic Y_τ to both the gamma approximation developed in this paper and the chi-square distribution used in the Fisher's combination method. To this end, five strata are simulated, resulting in a total number of combinations across five different tables on the order of magnitude of 10^7, with an average support size of 30 for each table. Specifically focusing on cases where positive outcomes equate the number of features (equal marginals), we vary the number of samples N_s across four different contexts, such that the support size represents 50%, 20%, 10%, and 5% of the total, i.e. $n_s = K_s = 0.5N_s$ for the first case. This deliberate selection places us in asymmetric scenarios where study factors (positive outcomes and features) are not rare in the first two cases and can be considered as rare in the latter two. Subsequently, for each of these four cases, five tables are simulated with an underlying hypergeometric distribution to conform to the null hypothesis context. Initially, we examine the non-truncated case ($\tau = 1$) for this stratified table, evaluating the exact distribution (Y_1), the uniform p-value approximation (χ^2_{10}), and the gamma distribution ($\Gamma(\alpha, \beta)$) as a suitable approximation. As depicted in Figure 1, due to the asymptotic nature of the chi-squared distribution, the rarer the features relative to the number of samples, the greater the difference between the exact value and the uniform p-value approximation (chi-squared). The Fisher chi-square combination appears to be a conservative test, and the gamma approximation fits the exact distribution very well in all cases. Indeed, the order of magnitude of the approximation error of the gamma distribution is a few units of percent, whereas the error of the chi-squared distribution is of the order of hundreds of percent.

In this example, the number of possible combinations is exactly $24.3M$ and the calculation time for gamma approximation is $0.01s$ on a 13th Gen Intel® Core™ i7-13700H CPU compared with 55 seconds in parallel calculation on 20 CPUs of the same type. For general comparison of computation time, see Figure S4 of the computation times with respect to the number of strata given in appendix 1.4. For this simulation, the gamma approximation boots the computation time by more than 1000 times, while keeping the computation accurate (e.g. error $< 6\%$ against error $> 300\%$ for χ^2, at a significant level of 10^{-3} for the first two cases). More details of accuracy and computation time for the gamma approximation and χ^2 are given in Supplementary Materials 1.3 and Figure S3.

The gamma approximation emerges as a suitable balance between precision and computational efficiency. This shows the importance of performing exact tests in the case of sub-uniform p-values and that the gamma approximation seems to be a good general approximation.

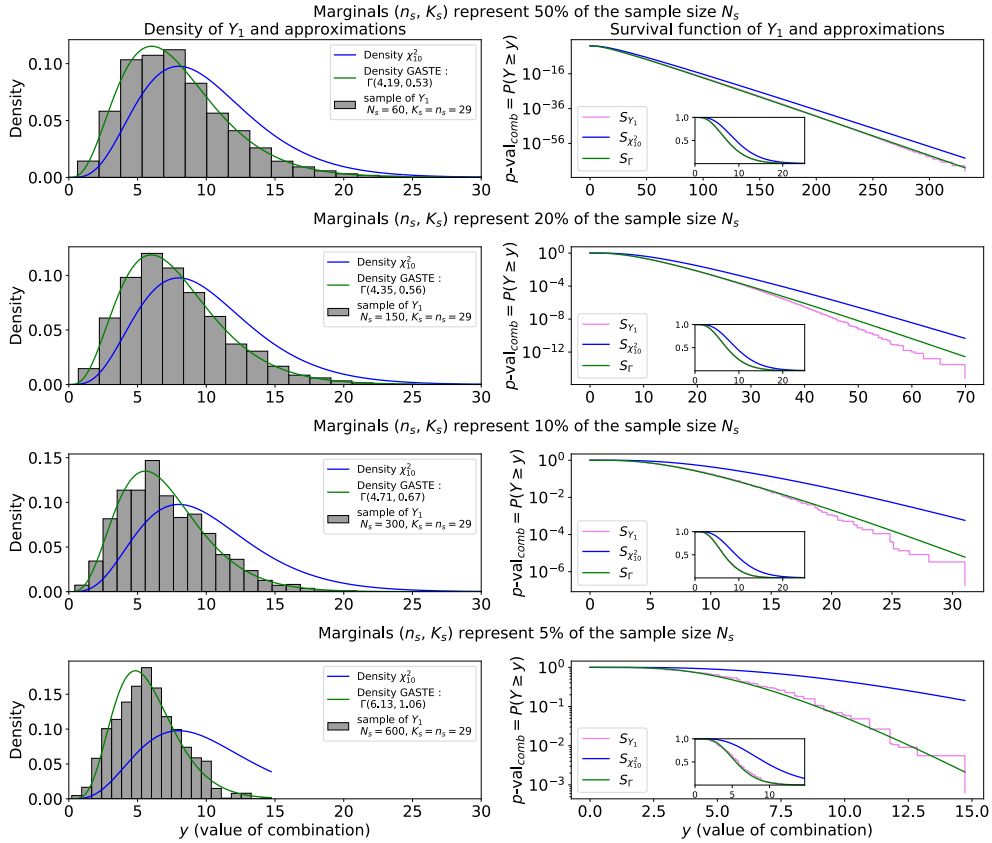

Figure 1. Comparison of stratified exact test, Y_1, its gamma approximation, Γ, and χ^2 distribution for the p-value combination of under-association. Study of four cases, depending on if features are rare or not, in the first case one in two samples has the features $N_s = 60$, $K_s = n_s = 29$, second case one in five $N_s = 150$, $K_s = n_s = 29$, third case one in ten $N_s = 300$, $K_s = n_s = 29$ and fourth case one in twenty $N_s = 600$, $K_s = n_s = 29$. On the right side the density function is plotted, the histogram is built from a sample of Y_1 given by the exact calculation. And on the left side the survival function of the distributions, i.e. the combined p-values for each possible Y_1-combination value.

Power of the test on simulations

Having shown that it is essential to be able to approximate the stratified exact test in order to reduce computing time considerably, it is now possible to run simulations to compare it with other tests. A second study was therefore carried out to compare the power of different tests and their behavior in homogeneous or non-homogeneous cases. We will compare our gamma approximation method of the stratified truncated exact test (GASTE) for three truncations $\tau = 1$ (no truncation), $\tau = 0.2$, $\tau = 0.05$ with the Bonferroni test, the CMH test and the classic Fisher combination following a χ^2 degree of freedom $2S$.

For the parameters of the simulation, to avoid having identical marginals in each stratum and create small variation on the input data, we draw the marginals according to a Poisson distribution with parameter $\mu = 30$ and then consider them as fixed to remain in the case of our conditional tests. Four strata are chosen arbitrarily, each with a sample size of $N_s = 150$, to be in the conditions of the second scenario described in the first simulation (previous section). This scenario involves a relatively minor rare factor compared with the sample size ($\approx 20\%$), so which

could reflect a realistic data case, and the choice is not biased towards favoring the GASTE method over the other method, as we do not opt for conditions where the method performs best (e.g., low marginals and large sample size).

We create 500 different sets of 4 tuples (N_s, n_s, K_s) to define the sample size and the marginals of 500 stratified contingency tables of 4 strata. For these 500 sets of stratified tables parameters, we apply an association effect to each stratum to reject at a certain level the null hypotheses $H^-_{0_s}$ and $H^+_{0_s}$. When simulating positive effects (resp. negative), the probability, under the null hypothesis, of drawing an individual belonging to the class of positive outcomes is multiplied (resp. divided) by the chosen effect size. Based on this new probability, the number of positive outcomes with features (k_s) is drawn. This generates an over-association (resp. under-association) of positive outcomes with the feature. Subsequently, the various global significance tests are applied to compare their power. Where the effect size is positive (resp. negative), the global over-association (resp. under-association) is studied. For a chosen effect size, and for each of the 500 simulated marginal tuples, we simulate 50 contingency tables (such that only k_s is re-drawn). For one effect size, we therefore run 25,000 stratified table simulations (50×500) and as many global significance tests. Figure 2 illustrates the results of this study.

In cases of homogeneity, our gamma approximation method without truncation demonstrates comparable, or slightly superior power compared to the CMH test (16% more powerful on average). Unsurprisingly, the Bonferroni test is the most conservative and χ^2 test is less powerful than GASTE. In instances of heterogeneity, where the CMH test cannot be applied due to strata heterogeneity for somes effects size, our GASTE method emerges as the most powerful (54% more powerful on average on effect size outside non-homogeneity). Given that the marginals n_s and K_s are less than half the sample size N_s, for the same absolute effect size, we observe weaker statistical power for under-association compared to over-association, owing to the natural rarity of positive outcomes with the feature. GASTE proves particularly potent in scenarios involving multiple small effects spread across strata. In instances of compensation, truncated GASTE methods exhibit greater power compared to the non-truncated approach. This is because in cases of compensation, certain strata exhibit significant effects in opposite directions, which truncation helps to mitigate by removing opposing effects while retaining significant ones. Furthermore, the use of the CMH test in compensation scenarios is erroneous, in our case due to the asymmetric margins relative to the sample size, the effects of compensation are not balanced. This falsely inflates the importance of under-association due to the inherently more powerful observation of over-association. This underscores the importance of separately studying over-association and under-association in non-homogeneous strata, as well as the necessity of verifying assumptions before applying the CMH test to avoid misleading conclusions. Finally, in the case of null effect sizes, all tests remain well below or just at the significant level alpha of 0.05. However, for a significant level alpha of 10^{-4}, the CMH test demonstrates over-sensitivity, leading to greater significance than observed, even under null conditions. This highlights an additional concern with the CMH test, as lower significant levels correspond to higher false positive rates, posing challenges for its reliability, particularly if the p-value of the CMH test is used in a multiple testing context.

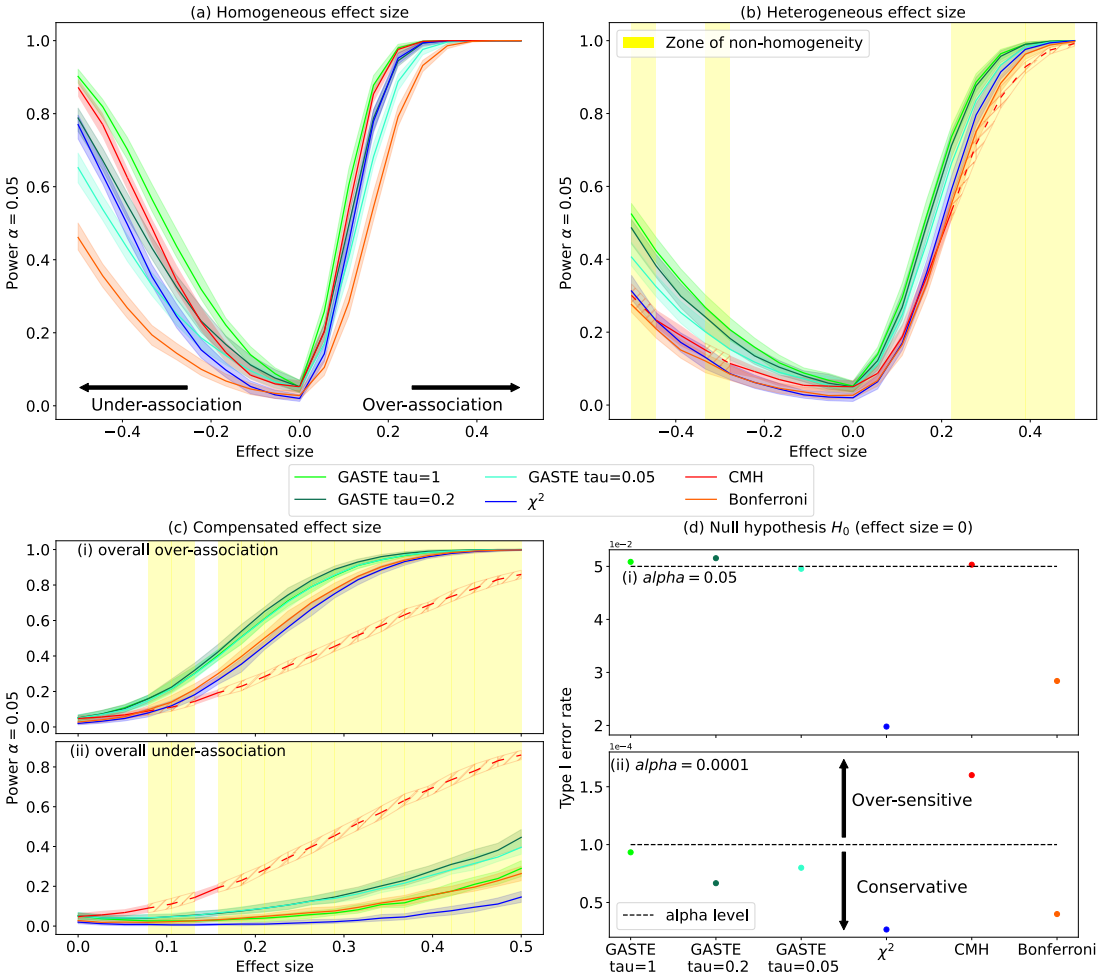

Figure 2. Study of 3 different effect size configurations, the first configuration, sub-figure (a), corresponds to a homogeneous case where the effect size is applied to the all 4 strata in the same way. The second configuration, sub-figure (b), corresponds to a case of heterogeneity where the effect size is only applied to two strata and the other two are in null condition. The third configuration, sub-figure (c), corresponds to a case of compensation where the effect size is applied positively to two strata and negatively to the other. For these configurations, the areas where the significance of the strata is not homogeneous using the Breslow-Day test are illustrated by a yellow background. In this zone, the CMH test appears as a hash because it should only be applied if the homogeneity test is satisfied. For sub-figure (d), the effect size is zero, i.e. the simulation is under the null hypothesis. The three configurations are concatenated to obtain 75,000 simulations of the overall significance test under null hypothesis (effect size = 0), and the type I error is studied for significant levels alpha, 0.05 and 10^{-4}.

Application to real datasets

To further illustrate our method, we will present two application examples.

Application I, Detection of association between plants species

We begin by outlining the application that motivated the development of our method. Our dataset comprises information on the presence or absence of plants (1048 different species) across various observation zones (177 zones) in the French Alps (ORCHAMP `https://orchamp.osug.fr/home`), each considered independent. The goal is to ascertain whether the observed coexistence between two species differs significantly from what would be expected under a null model knowing that they are seen in the same observation zone (co-occurring in the same environment). To achieve this, 2x2 contingency tables are employed, where rows denote the presence or absence of one species and columns denote the presence or absence of the other. One-sided Fisher's exact test is used to evaluate independence between the two species in a given environment. Subsequently, our GASTE method is employed to derive global insights across all environments, with each environment treated as an independent stratum.

Each observation zone consists of approximately 300 survey points, with some species being highly prevalent (existing more than 250 times out of 300) while others are rare (existing only a few times). For every pair of plants in the dataset, the exact combined p-value is computed if the support of Y_τ is less than five million combinations; otherwise, a gamma approximation with the fourth-order moment is used. To reduce the number of plant pairs to study, a filter is applied to the data by excluding pairs where the smallest achievable exact level of significance is over 0.2. Indeed, by example, it is not feasible to study the under-association of two rare species, as the probability of not observing any point with both plants is already high by nature. At the end, 63749 pairs of plants are studied in over-association and 10513 in under-association. Multiple testing correction is applied when examining all plant pairs of interest, with a Bonferroni correction utilized to identify significant positive and negative plant pair associations. Moreover, a truncation threshold of $\tau = 0.2$ is applied to increase the statistical power of the test, to cope with potential heterogeneity of association signal between plant among environments.

Similar to the approach illustrated in Figure 2, we compare the detection of coexistence using the various tests mentioned above on real data of ORCHAMP. This comparison is visualized through a Venn diagram (Figure 3) and under the null model (Table 2). The null model was obtained by randomly distributing the species 400 times over the survey points while maintaining the abundance of each species in each observation zone.

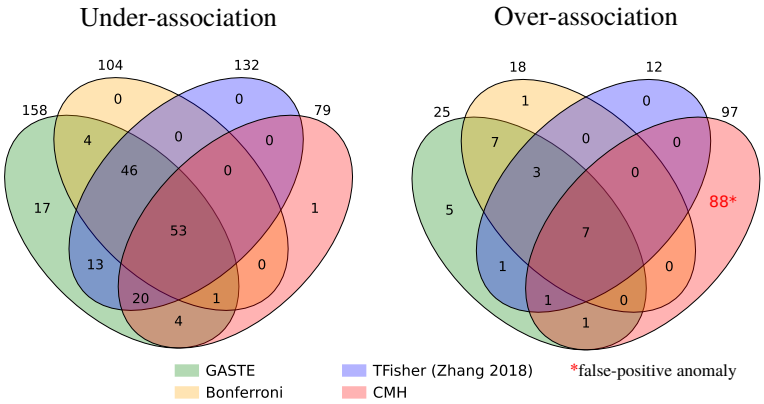

Figure 3. Venn diagram of the detection of association between two plants species in the ORCHAMP data. The first diagram represents the detection of under-association, the second diagram represents the detection of over-association. The different tests are represented by the different circles, the Bonferroni test, the CMH test, the TFisher test and our GASTE method. Note that the over-associations detected by the CMH method are mostly spurious detections, as witnessed by our FWER assessment.

Figure 3 illustrates that the GASTE method detects the most pairs ($\approx 20\%$ more than TFisher (uniformity assumption) and twice as many as CMH) while maintaining the probability

		GASTE	Bonferroni	TFisher	CMH
$\alpha = 0.2$	under-assoc	0.0475	0.0125	0.0125	0.015
	over-assoc	0.05	0.015	0.015	1
$\alpha = 0.05$	under-assoc	0.015	0.005	0.0025	0.005
	over-assoc	0.015	0.005	0.005	1
$\alpha = 0.01$	under-assoc	0.005	0.0	0.0	0.0
	over-assoc	0.005	0.0025	0.0025	1

Table 2. Family-wise error rate (FWER), under the null hypothesis, for the different association detection methods on ORCHAMP data.

of making at least one type I error (FWER) under the α significant level (Table 2). More precisely, our GASTE method detects 158 (resp. 25) pairs of plants significantly under-associated (resp. over-associated) across the different zone of observation, means co-occurring in the same zone (environment) but significantly not coexisting (resp. coexisting) in the same survey point. These two sets overlap on 6 plant pairs, meaning that the two plants are significantly under- and over-associated at the same time. This does not lead to an erroneous result, but shows that the association between the two plants depends on the environment, as a harsh environment leads plants to evolve in the same restricted area, and inversely, a rich environment leads plants to expand and conquer an area, producing competition between plants. A more precise example is given in Supplementary Materials 1.5.1 with the Figure S5 on the association between *Agrostis schraderiana* and *Rumex arifolius* species, concluding that the pair brings a significantly overall signal of under-association in an open environment like grassland and brings a significantly overall signal of over-association in a closed environment like forest.

One unusual observation in the result of null model is the family-wise error rate of the CMH test. Firstly, this anomaly is asymmetrical; it occurs in the case of over-association but not under-association. Secondly, the error rate is significantly higher than the α significant level for all tested levels, as it is always equal to 1. This means that during each randomization, more than one pairs of plants is detected in over-association under the null hypothesis. These detections correspond to cases where the marginals are very small, just a few counts, which explains why they are found only in over-association. Our pre-selection of pairs prevents them from being studied in under-association, but it is possible to evaluate the over-association of two rare species. However, in the case of the CMH test, these pairs do not comply with Cochran's rule, which states that the counts in the contingency tables must each be greater than 5. Consequently, the detection of over-association by the CMH method exceeds that of the other methods, with 88 pairs detected. These detections are very likely false positives that do not adhere to this Cochran's rule, and this hypothesis would need to be verified for each pair. This highlights the problem of the over-sensitivity and generality of the CMH test, particularly in cases where the marginals are low.

In addition, excluding this anomaly, the FWERs are relatively lower than the α significance level due to the Bonferroni correction in the multiple tests. However, once again, our GASTE method has a better detection rate than the other methods, while the conservative nature of the Bonferroni and TFisher methods is salient.

Application II, Berkeley's admission (1973)

For this second application, we take a widely used example of stratified data which illustrates Simpson's paradox (Simpson, 1951) and highlights the value of stratifying data. The dataset known as UCBAdmissions lists admissions to the University of California, Berkeley, in 1973 according to gender and by department (from A to F). The study of these data was popularised by, at first glance, discrimination against women in admissions (Bickel et al., 1975). Indeed, looking at the overall admission rate, 44% of men were admitted (1198 out of 2691) compared with 30% of women (557 out of 1835). However, by stratifying the data by department, this trend tends to be reversed.

For each department, a 2x2 contingency table is created with men and women in the columns and admissions and refusals in the rows. The association studied is therefore men and admissions,

in order to detect whether there are more or fewer men admitted than in the case of a null model (independence between sex and admission). We recreated, Figure 4, a Forest plot summarizing the data from the tables stratified by department (odds ratio, confidence interval, CMH weights), and we added the information from our two global over- and under-association tests as well as the popular CMH test and the Breslow-Day homogeneity test. The homogeneity criterion is not fulfilled (Breslow-Day test p-value 0.002) and prevents the use of the CMH test. However, as was done in the original study without checking homogeneity, the overall association via CMH test is announced as non-significant (p-value 0.22). The non-homogeneity of the strata already distorts the result of the CMH test. Moreover, in department A there is a strong under-association of admissions by men compared with women (p-value Fisher's exact test 1.15e-5). This under-association in this stratum (department) drives the overall under-association test, making it also significant ($p\text{-value}_{combined}$ 0.0012). The result of our GASTE test is therefore that if men and women had been admitted to the departments in the same way, the probability of observing such results is one chance in a thousand. By artificially adding departments with results similar to departments B to F, for it to be likely to observe such a difference in department A at random ($p\text{-value}_{combined} > 0.05$), there would have to be 20 departments in total. These results were obtained using our GASTE approximation method, as the calculation cannot be made exactly (one thousand billion combinations).

In conclusion, while there appears to be an overall under-association of men compared to women in admissions, this does not necessarily indicate a gender bias in admissions. It is possible that women outperformed men in department A to such an extent that it created the perception of an overall under-association of men.

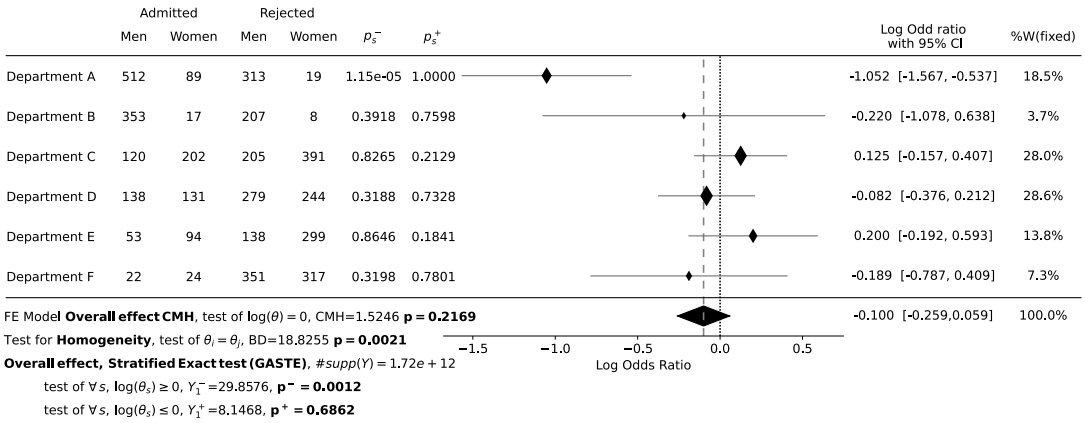

Figure 4. Berkeley's admission (1973) stratified analyse

DISCUSSION

We believe that our GASTE method offers substantial value in the statistical analysis of stratified binary data. As demonstrated in both our simulations and on real data applications, the GASTE method is a powerful test that operates without assuming homogeneous association between features and outcomes among strata and with no minimum data requirement in the contingency tables. This contrasts with the CMH test, which, despite its widespread use, suffers from both of these limitations. Applying statistical tests such as CMH without complying with their hypothesis (e.g. homogeneity, uniformity of the p-values for Fisher combination, sufficient counts for CMH) can lead to both false discoveries and undetected effects. Drawing general rules on approximate fulfillment of the underlying hypothesis is hard in general: as shown in the simulation (cf. Figure 1, bottom plot), when the marginals are low, the asymptotic test is far from being accurate even with large sample sizes (e.g. even with $N > 600$). To avoid these limitations, it is crucial to conduct exact tests whenever possible, or use approximate computations of the exact statistic when computationally intractable, offering a high accuracy at a low computational cost (Supplementary Materials 1.3 and Figure S3.). The GASTE method performs effectively

in scenarios with both homogeneous and heterogeneous effects between strata and in cases of both low and high significance. The major point of our method is to approximate the exact test rather than relying on asymptotic tests, enabling a wide variety of applications thanks to the computational efficiency. Our method can trigger detection as soon as odd ratio of a single stratum is sufficiently far from 1(i.e. the stratum has a sufficiently significant p-value) compared with the other strata, as well as when the signal is diffuse (i.e. in each stratum the p-value has a low level of significance, but this low significance is found across several strata), example Application I, Supplementary Materials 1.5.2, Figure S6. Computing the exact probability of observing a set of 2x2 contingency tables under the null model can be straightforwardly rephrased into a problem of combining the discrete p-values of each stratum. This formulation, in turn, led us to naturally propose an approximation of the exact test. Contrary to usual method, we think that, it is essential to consider under-association and over-association separately (not used two-sided p-value), as in cases of heterogeneity (cf Application I 1.5), both associations can be significant at the same time. Incorporating truncation into the p-value combination enhances statistical power in scenarios featuring few effects or contradictory effects between strata. However, when effects are homogeneous, truncation may not augment power and could potentially diminish it. Thus, truncation should be considered beforehand depending on the considered problem.

Moreover, we think that our method will be of great interest for many applications. In the case of meta-analysis using 2x2 tables, our method does not require selecting a set of studies complying with homogeneity criterion, thereby avoiding "cherry-picking" risk for meta-analysis (Yoneoka and Rieck, 2023). This mitigates both study selection bias (Felson, 1992; Eisend and Tarrahi, 2014), particularly in heterogeneous scenarios, and publication bias (Rothstein et al., 2005; Kicinski et al., 2015), as even studies with insignificant results provide information within a multiple-test or p-value combination framework. We also posit that our method holds promise in epidemiology, where sample sizes (i.e., marginals) are not always controllable and can turn out to be low. Even if our method performs well in general, there are still some pitfalls when using the gamma approximation. Indeed, inference of the parameters of the gamma distribution is based on the analytical matching of moments, and the choice of moments can yield different results in extreme cases. Selecting the appropriate moments is crucial for obtaining a good approximation. By default, we recommend choosing the highest moment that has been implemented, the fourth-order moment, as it better captures the tail distribution of the exact combination law Y_τ, especially when truncation is applied ($\tau \neq 1$). However, it's worth noting that in scenarios where the support of Y_τ is small compared to its variance, selecting a smaller moment may yield a more accurate p-value. When achieving a resolution for the parameter α is unattainable with a higher-order moment, then a lower-order moment is automatically chosen. These considerations are illustrated in Supplementary Materials 1.2.4 with the Figure S2. For further research, it would be interesting to develop a method to select the best moment for a given problem. Finally, GASTE method focuses on the combination of p-values stemming from the hypergeometric distribution, but the same idea both for computing the exact law and the gamma approximation can easily be transferable to other discrete laws with finite support.

CONCLUSIONS

We have reformulated the stratified exact test problem as a Fisher combination of one-sided sub-uniform p-values. We propose an exact algorithm as well as a gamma approximation of the exact combination statistic when exact calculation is prohibitively expensive. Furthermore, we have extended this test and its approximation to incorporate truncation, enhancing statistical power in scenarios with non-homogeneous and opposing effects between strata. The GASTE method offers a generalization suitable for both low and high significance cases in homogeneous and heterogeneous settings. Compared to the conventional CMH test, our method appears more suitable for studies involving stratified 2x2 tables.

DATA AND CODE AVAILABILITY

To use the method in practice, we have developed a Python package available on PyPI at `https://pypi.org/project/gaste-test/` (open-source code available at `https://github.com/AlexandreWen/gaste`) and the data and code used in this study are available at `https://gricad-gitlab.univ-grenoble-alpes.fr/wendlina/gaste_method`.

REFERENCES

Agresti, A. (2012). *Categorical data analysis*, volume 792. John Wiley & Sons.

Berger, V. W., Stefanescu, C., and Zhou, Y. Y. (2006). The analysis of stratified 2× 2 contingency tables. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 48(6):992–1007.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111.

Breslow, N. E., Day, N. E., and Heseltine, E. (1980). Statistical methods in cancer research. *International Agency for Research on Cancer Lyon*.

Brown, M. B. (1975). 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, pages 987–992.

Chiba, Y. (2017). Stratified exact tests for the weak causal null hypothesis in randomized trials with a binary outcome. *Biometrical Journal*, 59(5):986–997.

Cochran, W. G. (1954). Some methods for strengthening the common χ 2 tests. *Biometrics*, 10(4):417–451.

Deeks JJ, Higgins JPT, A. D. (2023). Undertaking meta-analysis. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, pages 249–325.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188.

Eisend, M. and Tarrahi, F. (2014). Meta-analysis selection bias in marketing research. *International Journal of Research in Marketing*, 31(3):317–326.

Evans, C. R. (2024). Overcoming combination fatigue: Addressing high-dimensional effect measure modification and interaction in clinical, biomedical, and epidemiologic research using multilevel analysis of individual heterogeneity and discriminatory accuracy (maihda). *Social Science & Medicine*, 340:116493.

Felson, D. T. (1992). Bias in meta-analytic research. *Journal of Clinical Epidemiology*, 45(8):885–892.

Fisher, R. A. et al. (1936). Statistical methods for research workers. *Statistical Methods for Research Workers.*, 6th Ed.

Jung, S.-H. (2014). Stratified fisher's exact test and its sample size calculation. *Biometrical Journal*, 56(1):129–140.

Kicinski, M., Springate, D. A., and Kontopantelis, E. (2015). Publication bias in meta-analyses from the cochrane database of systematic reviews. *Statistics in Medicine*, 34(20):2781–2793.

Kincaid, W. (1962). The combination of tests based on discrete distributions. *Journal of the American Statistical Association*, 57(297):10–19.

Kost, J. T. and McDermott, M. P. (2002). Combining dependent p-values. *Statistics & Probability Letters*, 60(2):183–190.

Li, Q., Hu, J., Ding, J., and Zheng, G. (2014). Fisher's method of combining dependent statistics using generalizations of the gamma distribution with applications to genetic pleiotropic associations. *Biostatistics*, 15(2):284–295.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4):719–748.

Mielke Jr, P. W., Johnston, J. E., and Berry, K. J. (2004). Combining probability values from

independent permutation tests: a discrete analog of fisher's classical method. *Psychological reports*, 95(2):449–458.

Poole, W., Gibbs, D. L., Shmulevich, I., Bernard, B., and Knijnenburg, T. A. (2016). Combining dependent p-values with an empirical adaptation of brown's method. *Bioinformatics*, 32(17):i430–i436.

Reilly, M. (2023). *Controlled Epidemiological Studies*. Chapman and Hall/CRC.

Rothstein, H. R., Sutton, A. J., and Borenstein, M. (2005). Publication bias in meta-analysis. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, pages 1–7.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

Smaardijk, V. R., Maas, A. H., Lodder, P., Kop, W. J., and Mommersteeg, P. M. (2020). Sex and gender-stratified risks of psychological factors for adverse clinical outcomes in patients with ischemic heart disease: A systematic review and meta-analysis. *International Journal of Cardiology*, 302:21–29.

Yang, J. J. (2010). Distribution of fisher's combination statistic when the tests are dependent. *Journal of Statistical Computation and Simulation*, 80(1):1–12.

Yang, J. J., Li, J., Williams, L. K., and Buu, A. (2016). An efficient genome-wide association test for multivariate phenotypes based on the fisher combination function. *BMC bioinformatics*, 17:1–11.

Yoneoka, D. and Rieck, B. (2023). A note on cherry-picking in meta-analyses. *Entropy*, 25(4):691.

Yusuf, S., Collins, R., Peto, R., Furberg, C., Stampfer, M., Goldhaber, S., and Hennekens, C. (1985). Intravenous and intracoronary fibrinolytic therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. *European Heart Journal*, 6(7):556–585.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002). Truncated product method for combining p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 22(2):170–185.

Zhang, H., Tong, T., Landers, J. E., and Wu, Z. (2018). Tfisher tests: Optimal and adaptive thresholding for combining p-values. *arXiv preprint arXiv:1801.04309*.

Zhang, H. and Wu, Z. (2023). The generalized fisher's combination and accurate p-value calculation under dependence. *Biometrics*, 79(2):1159–1172.

Zheng, G., Wu, C. O., Kwak, M., Jiang, W., Joo, J., and Lima, J. A. (2012). Joint analysis of binary and quantitative traits with data sharing and outcome-dependent sampling. *Genetic Epidemiology*, 36(3):263–273.