International Journal of Statistics and Probability; Vol. 11, No. 6; November 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Combining Correlated P-values From Primary Data Analyses

Jai Won Choi[1], Balgobin Nandram[2], & Boseung Choi[3]

[1] Statistical consultant, Meho Inc., U.S.A.

[2] Professor, Worcester Polytechnic Institute, U.S.A.

[3] Professor, Korea University, Sejong Campus, Korea

Correspondence: Jai Won Choi, 9504 Mary Knoll Drive, Rockville, MD, U.S.A.

Received: July 31, 2022 Accepted: October 3, 2022 Online Published: October 20, 2022

doi:10.5539/ijsp.v11n6p12 URL: https://doi.org/10.5539/ijsp.v11n6p12

Abstract

Research results on the same subject, extracted from scientific papers or clinical trials, are combined to determine a consensus. We are primarily concerned with combining p-values from experiments that may be correlated. We have two methods, a non-Bayesian method and a Bayesian method. We use a model to combine these results and assume the combined results follow a certain distribution, for example, chi-square or normal. The distribution requires independent and identically distributed (iid) random variables. When the data are correlated or non-iid, we cannot assume such distribution. In order to do so, the combined results from the model need to be adjusted, and the adjustment is done "indirectly" through two test statistics. Specifically, one test statistic (TS^{**}) is obtained for the non-iid data and the other is the test statistic (TS) is obtained for iid data. We use the ratio between the two test statistics to adjust the model test statistic (TS^{**}) for its non-iid violation. The adjusted TS^{**} is named as "effective test statistics" (ETS), which is then used for statistical inferences with the assumed distribution. As it is difficult to estimate the correlation, to provide a more coherent method for combining p-values, we also introduce a novel Bayesian method for both iid data and non-iid data. The examples are used to illustrate the non-Bayesian method and additional examples are given to illustrate the Bayesian method.

Keywords: assumed distribution, Correction ratio, Correlation, Model assumptions, P-values, Effective test statistic, Statistical inference

1. Introduction

Researchers use a model to combine the results, p-values or Z-scores, from sample surveys or clinical trials for the same subject or purpose. We consider these results are iid random variables and assume a certain distribution, for example normal, for statistical inference. Such a distribution requires iid-random variables.

However, these variables are more likely correlated as they are from the similar sample surveys or clinical trials for a specific topic or purpose. For example, poll results of presidential election or clinical trial results of one medication executed from different locations, or from the repeated trials at a same place (see Example 1). These results are often reported as p-values. We do not consider the previous procedures in obtaining p-values, and the k p-values are really the random variables. However, we are attacking a problem that is, indeed, very difficult because no aspect of the correlation is known, and moreover, there is a single sample of p-values, thereby making it impossible to find Pearson correlation.

The resulting p-values are non-iid random variables (see Example 1 and Appendix B). We present a method to show how an assumed distribution, which requires iid-random variables, can be applied to non-iid variables. To do so, non-iid variables need to be adjusted indirectly through its test statistics (TS^{**}). This adjustment is done by comparing two test statistics, one from the non-iid model and other from the iid model. The test statistic (TS^{**}) comes from a model with non-iid data, given null hypothesis, sample size and test level. Similarly, the other test statistic, (TS), comes from an assumed distribution with iid-random variables. We define correction factor as the ratio of TS^{**} to TS. Finally, we can get effective test statistic (ETS) of TS^{**} divided by the correction factor and this ETS is used to make statistical inference with the assumed distribution.

We use one of the two methods to combine the non-iid results or p^* values, Non-Bayesian or Bayesian. We show two methods for non-Bayesian in Section (3.1) show how to obtain ETS of correlated data (Choi and McHugh,1989), and in Section (3.2) show how to obtain ETS for TS^{**} of non-iid data, that involve not only correlation but also other non-iid-conditions, if any. Then, we use ETS with the assumed distribution.

The case of iid random variables to obtain TS

TS is based on a test statistic. It is the standard test statistic with which two other test statistics, TS^* or TS^{**}, are compared to measure the size of its deviation from TS, where TS^* is from a distribution of correlated variables and TS^{**} is from a distribution of non-iid variables. Below, we show how TS is obtained.

Suppose, p = $(p_1, \ldots, p_n)$, $0 \leq p_i \leq 1$, i=1,…, n, are iid random variables with a known distribution function h(p|θ). One can make statistical inferences on p. Let the global null hypothesis H_o: $\theta_1 = \ldots. = \theta_n = \theta$ against alternative hypothesis $H_1: \theta_i \geq \theta$ for some i = 1, …, n. The hypothesis H_0 is reasonable as all the tests are done for a same purpose. We assume that h(p|θ) is a monotone function, and therefore it is optimal for combining p-values (Birnbaum, 1954).

We define test statistic (TS) as

$$TS_\alpha \text{or } t_\alpha = \text{T}(h(p|\theta), \alpha, n),$$

where the rejection test level α is obtained as

$$\alpha = 1 - \int_{-\infty}^{t_\alpha} \text{h(p}|\theta) \text{ dp.}$$

TS does not involve in hypothesis testing and it is based on the assumed distribution function $h(p|\theta)$ of iid p-values for given α and n. For example, h(p|θ) is Normal(μ, σ^2), or x_{2n}^2 chi-square 2n degrees of freedom. When we use $h(p|\theta)$ as base distribution of TS, we do not need actual p values, but the $h(p|\theta)$ implies p as iid random variables. For example, we only need sample size n and test level α to have table value of TS for x_{2n}^2, chi-square 2n degrees of freedom. The test level α is pre-selected by researcher. This TS is used only to compare to study test statistic, TS^* or TS^{**}, to measure its deviation from TS, and they involve in testing a null hypothesis at the same sample sized n and test level α of TS.

Above TS, based on $h(p|\theta)$ of iid-random variables p, is its own ETS. TS is compared to two study test statistics, TS^* based on correlated data and TS^{**} based on non-iid variables. We ignore the pre-procedures to obtain these data, and consider these data are the variables of our interest.

This paper has five more sections. In Section 2, we review pertinent literature. In Section 3, we present the non-Bayesian method. In Section 4, we show examples to illustrate the non-Bayesian method. In Section 5, we present Bayesian method to find the posterior mean of the combined p-value and some additional examples are presented. Section 6 includes a brief conclusion.

2. Pertinent Literature

Yoon et al (2021) used Meta analyses to increase statistical power by combining statistics (e.g., effect sizes, z- scores, or p-values) from multiple studies when they share the same null hypothesis under the assumption that all the data in each study have an association with a given phenotype. However, specific experimental conditions in each study can result in independent statistics that are derived from a null distribution. They showed the power of Meta analysis rapidly decreases as they were combined, Fisher's Method (Fisher, 1932), Weighted Fisher's method (wFisher), and Ordered p-values (ordMeta) increased power. The last two methods (i.e., wFisher and ordMeta), outperformed existing Meta-analysis when only a small number of studies n=2 is combined. The weighted Fisher's method (wFisher) assigned non-integer weights to each p-value, that are proportional to sample sizes. The wFisher and ordMeta are more robust than the test statistic of the Meta method.

Vovk and Wang (2020) got the average of k p-values $p_1,\ldots,p_k$ to obtain one combined value without any parametric or distribution assumption. They reviewed previous results of arithmetic mean (AM $\bar{p}$) by multiplying 2 as $2\bar{p}$ and geometric mean (GM) replacing 2 by e (=2.718). They extended the recent risk aggregation technique to harmonic mean (HM) by multiplying log K for $K \geq 2$, scaling up by a factor of log k, where k is number of p-values. They also explore several other weighted averages of p-values. Note that the inequality of HM$\leq$ GM $\leq$AM, related to scaling factors, which is proved using Jensen's inequality (Casella and Berger, 2002).

Vovk and Wang (2020) showed several models to combine $p_1,\ldots,p_k$ into a single p-value. assuming, $p_1,\ldots,p_k$ are independent random variables. The simplest way to combine them is the Bonferroni method,

$$\text{F}(p_1,\ldots,p_k) = K \min(p_1,\ldots,p_k),$$

when $\text{F}(p_1,\ldots,p_k)$ exceed 1, it can be replaced by 1. Other method, used to smooth out overestimation of above-mentioned method, is a general average:

$$M_{\emptyset,k}(p_1,\ldots,p_k) = \varphi\left[\frac{\emptyset(p_1\}+,\ldots,+\emptyset(p_K\}}{K}\right],$$

where $\emptyset(0,1) \to (-\infty, \infty)$ is a continuous strictly monotonic function and $\varphi[(0,1)] \to (0,1)$ is its inverse. For example, AM corresponds to the identity function $\emptyset(p)$= p, GM corresponds to $\emptyset(p)$= log p, and HM corresponds to $\emptyset(p)$= 1/p. They present more extensions of this basic idea.

Loughin (2004) compared several methods, when only p-values are available, in combining p-values from independent tests under combined hypothesis heuristically through simulation. They are minimum value (Tippett, 1931), Chi-square combining model (Fisher, 1932), scaled normal (Liptak, 1958), maximum value (Wilkinson, 1951), combinatoric uniform (Edington, 1972) and approximately scaled logistic (Rastogi, 1979).

Fisher's Model (FM) (1932) is $g(\boldsymbol{p}^*|\boldsymbol{\theta}) = -2\sum_{i=1}^{n}\log(p_i^*) = -2\log(p_1^* \dots p_n^*) = log\frac{1}{\{p_1^* \dots p_n^*\}^2}$. to combine $p_1^* \dots p_n^*$.

FM assumes the null hypothesis distribution follows x_{2n}^2, chi-square with 2n degrees of freedom for n independent random variables. This is not true when p^* are correlated. Other problem of FM arises when combining a large number of p^*-values. When n $\to \infty$, FM value$\to \infty$, i.e., combined value of even non-significant p-values becomes significant for a large n (Choi and Nandram, 2021).

Hess and Iyer (2007) used Fisher's Score combining p-values to detect differential genes array using Affymetrix expression arrays. Others (Tippett,1931, and Wilkinson,1951, George,1977, Stouffer,1949) suggest non-parametric methods to combine p-values.

Most methods, presented above, assumed independent p-values and did not address correlation or non-iid problems for statistical inference. Our research addresses a solution for this problem. However, this is a difficult problem because one cannot estimate the correlation in a straightforward manner, and this is an innovation in this paper as well. In a recent paper, Heard and Rubin-Delanchy (2018) showed how to choose between different methods to combine p-values. They also discussed the likelihood ratio for combining p-values and the weighted average of the logarithms of the p-values. However, there was no discussion about correlated p-values nor any discussion of the Bayesian approach, presumably there is none.

There is virtually no Bayesian attempt on the specific problem we are considering in this

paper. Specifically, we are combining a number of p-values, which may be dependent because the experiments are done under the same protocol, and similar procedures may be followed at the different experimental sites or laboratories. However, there is a sparse literature on the study of Bayesian p-values, not the combination of p-values. See Casella and Berger (1987) and the discussions that followed on reconciling Bayesian and frequentist evidence on the one-sided testing problem.

3. Non-Bayesian Method

Test statistics, TS^* for correlated variables and TS^{**} from non-iid variables, are compared by the standard rule, TS, for iid variables to see the size of their deviations from TS. We introduce these two test statistics, TS^* in (3.1) and TS^{**} in (3.2). We also present the correction factors, C^* and C^{**}, for TS^* and TS^{**} and its estimations. We also present Table 1 to illustrate practical application to clinical data.

In the introduction, we discussed the base test statistic TS for $h(p|\theta)$ with iid random variables p = $(p_1, \dots, p_n)$.as a standard rule to which TS^* or TS^{**} are measured.

In 3.1, the TS^* of g($p^*|\theta$) for correlated variables $p^* = (p_1^*, \dots, p_n^*)$ for given sample size and test level is compared to the base test statistic TS of $h(p|\theta)$ to find its difference, which is expressed as ratio, C^*= TS^* /TS. We call C^* correction factor (CF) as it corrects the impact of correlation on TS^*.

In 3.2, TS is now compared to TS^{**} for non-iid $p^{**} = (p_1^{**}, \dots, p_n^{**})$, which may carry not only correlation but also all other non-iid violations, if any. The difference between these two test statistics expressed as the ratio C^{**}= TS^{**} /TS. Here C^{**} corrects the impacts not only correlation but all other violations of iid condition.

In 3.3, we show how to estimate C^{**}. Three candidates are presented.

In 3.4, we illustrate TS, TS^{**}, and C^{**} in Table 1, using Fisher's Model F for TS^{**} and chi-square distribution C for TS. Table 1 is continuously used in the next Section 4. It shows for Fisher's Model users how to use the table values of C^{**} for possible violations of correlation or non-iid problem.

3.1 Correlated Random Variables, Model 1

Previously we introduced the base test statistic TS = T($h(p|\theta)$,α,n) for a known distribution h(p|θ) of iid random variables p = $(p_1, \dots, p_n)$, $0 \le p_i \le 1$, i=1,…, n, for given test level α and sample size n.

Now we consider. We can obtain the test statistic (TS^*) for the combining model g($p^*|\theta$) of these correlated variables, $p^* = (p_1^*, \dots, p_n^*), 0 \le p_i^* \le 1$, for a given hypothesis H_O^*, test level α^*, correlation ρ and sample size n. We can assume g($p^*|\theta$) is its pseudo distribution and write

$$TS^* = \mathrm{T}(\mathrm{g}(p^*|\theta),\ H_O^*, \alpha^*, \rho, n).$$

Choi and McHugh (1989) discussed how to reduce the TS^* for the correlated variables in Chi-square testing. The g($p^*|\theta$) is erroneously assumed to follow h(p|θ), chi-square distribution x_{2n}^2. When the test statistic (TS) for distribution h(p|θ) is compaired to TS^* of the actual model g($p^*|\theta$), the test statistic, TS^* is largely inflated because of the correlation. Hence TS^* is reduced, dividing it by the correction factor C^*= [1 + ρ(n-1)] , ρ is the positive correlation among p^*-values, n is the sample size.

Choi and McHugh (1989) showed how to obtain the effective test statistic (ETS) of test statistic TS^* with this correction factor, C^*,

$$\mathrm{ETS}=\frac{TS^*}{C^*},$$

on $1 \le C^* < \infty$. It implies that the correlation of the variables $p^* = (p_1^*, \dots, p_n^*)$, is indirectly adjusted by the correction factor C^*. After such correction, we can now make statistical inference on the effective test statistic ETS with assumed distribution h(p|θ), for example chi-square distribution.

We can also achieve the same goal through effective sample size n_e of n, $n_e = \frac{n}{C}$ to obtain ETS (Choi,1980). For example, for binomial variables, x_i, i=1,…,n, that are correlated, its normal approximation of test statistic TS^* under the null hypothesis $H_0: p = 0$, is given as N(1,0)= $\frac{n\hat{p}}{\sqrt{p(1-p)}}$. We can use the reduced sample size $n_e = n/C$, to obtain effective test statistic, ETS =$\frac{n_e \hat{p}}{\sqrt{p(1-p)}}$.

3.2 Non-iid Random Variables, Model 2

In this section, we try to find the differences between the test statistic TS^{**} and basic test statistic TS, TS=T($h(p|\theta)$,α, n), and TS^{**}= T(g($p^{**}|\theta$), $H_O^{**}, \alpha^{**}, \rho, n^{**}$). Two types of differences can be considered: One is the correlation ρ in the variables $p^{**} = (p_1^{**}, \dots, p_{n^{**}}^{**})$, and other includes all other known or unknown differences such as $h(p|\theta) \neq \mathrm{g}(p^{**}|\theta)$, $p^{**} \neq$ p, null hypothesis H_O^{**}, $\alpha \neq \alpha^{**}$, n $\neq n^{**}$.

The model $g(p^{**}|\theta)$ in TS^{**} is used to combine the non-iid variables $p^{**} = (p_1^{**}, \dots, p_{n^{**}}^{**})$. The distribution $h(p|\theta)$ in TS is based on iid variables p= $(p_1, \dots, p_n)$. Users of the model $g(p^{**}|\theta)$ assume that $g(p^{**}|\theta)$ follows the distribution $h(p|\theta)$ as if p^{**} = p. It is a wrong assumption if $p^{**} \neq$ p. The aim of this section is to correct the wrong assumption indirectly by adjusting the test statistic, TS^{**}, while TS of assumed distribution h(p|θ) remains the same.

We have shown when TS is compared against TS^* for correlated variables $p^* = (p_1^*, \dots, p_n^*)$ in 3.1. Here in 3.2, we compare TS to TS^{**} for variables $p^{**} = (p_1^{**}, \dots, p_{n^{**}}^{**})$, which is not only correlated but also violated non-iid and other conditions, if any.

The total difference between the two test statistics, TS^{**} and TS, is defined as the ratio of these two test statistics:

$$C^{**}=\frac{TS^{**}}{\mathrm{TS}}, 0 < C^{**} < \infty.$$

Note that $TS^{**} \ge TS$ (Appendix A) when $1 < C^{**} < \infty$, and $TS^{**} < TS$ when $0 < C^{**} < 1$. The turning point greater than 1 or less than 1 depends on the size of p-values and the number n of the p-values as well as on the different changing speed, increasing or decreasing, of the TS^{**} and TS (see Table 1). We can ignore $TS^{**} < TS$ when $0 < C^* < 1$, since we assume only positive correlation of $p_1^{**}, \dots, p_{n^*}^{**}$ or consider only $TS^{**} \ge TS$ to correct positive correlation and other violation of TS^{**}.

To correct the impacts of non-iid and other violations, if any, we adjust TS^{**} by C^{**}as

$$ETS^{**} = \frac{TS^{**}}{C^{**}}, 0 < C^* < \infty.$$

Note $ETS^{**} > TS^{**} > TS$ on the interval $1 \leq C^{**} < \infty$ (Appendix A). The ETS** is the effective test statistic of the test statistic (TS**) on the interval,$1 < \boldsymbol{C}^{*} < \infty$. Here, the non-iid violation of the variables $p_1^{**}, \dots, p_{n^*}^{**}$, is indirectly corrected through C^{**}.

Lemma

The difference between the two test statistics, TS^{**} and TS can be expressed as the ratio , C^{**}=TS^{**}/ TS, $0 < C^{**} < \infty$, the correction factor, C^{**}, indirectly correct the correlation and other iid violations of TS^{**}. The effective test statistic is ETS^{**}=TS^{**} /C^{**}, on $1 < C^{**} < \infty$. Then, the effective test statistic ETS^{**} of test statistic, TS^{**}, is used for statistical inference with the originally assumed distribution $h(p|\theta)$.

Proof is outlined in Appendix A

3.3 Estimation of Correction Factor C^{**}

The correction factor C^{**} indirectly measures all violations including non-iid condition of p^{**}. In actual situation, it is difficult to obtain exact TS^{**} and hence C^{**}. To estimate $C^{**} = \frac{TS^{**}}{\text{TS}}, 1 < C^{*} < \infty$, we compare TS = T($h(p|\theta)$,$\alpha, n$) of assumed iid random variables p = $(p_1, \dots, p_n)$ to TS^{**}= T(g($p^{**}|\theta$), $H_0^{**}, \alpha^{**}, \rho, n^{**}$), of non-iid variables $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$. While the TS remains the same for given $h(p|\theta)$, α, n, the TS^{**} can be estimated by how we use $(p_1^{**}, \dots, p_{n^*}^{**})$ in the combing model g($p^{**}|\theta$). Below shows three ways of different use of these variables.

The three candidates are (1) is to use the minimum value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, expressed as C_{Min}^{**}, (2) uses the maximum value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, expressed as C_{Max}^{**}, (3) is the sum of individual values of TS^{**}, expressed as C_{Mix}^{**}, each term of $TS^{**}$is divided or individually weighted by all member weights (Example 1). All member weight is used because the weight of one member is one: when sample size is one (i.e., n=1), it is independent automatically regardless of the size of p-values, i.e., T($h(p|\theta), \alpha, n = 1$)= T(g($p^{**}|\theta$), $H_0^{**}, \alpha^{**}, \rho = 0, n^{**} = 1$), $h(p|\theta)$ = g($p^{**}|\theta$) for given $\alpha = \alpha^{**} = p = p^{**}$, ignoring the null hypothesis, H_0^{**}.as assumed distribution $h(p|\theta)$ is not involved in any null hypothesis. This is the only time the assumption is correct, or $h(p|\theta)$ = g($p^{**}|\theta$) (see First row, Table 1, Example 1).

Three possible correction factors are C_{Min}^{**}, C_{Max}^{**},and C_{Mix}^{**} (Appendix C). The choice depends on researcher's need. Thus, three different effective test statistics, ETS^{**}=TS^{**} /C^{**}, can be obtained when TS^{**} reduced by respective new correction factor:

$$ETS_{min}^{*} = \frac{TS^{**}}{C_{Min}^{*}},$$

$$ETS_{max}^{*} = \frac{TS^{**}}{C_{Max}^{*}},$$

and

$$ETS_{mix}^{*} = \frac{TS_1^{**}}{C_{Mix,1}^{*}} + \dots + \frac{TS_n^{**}}{C_{Mix,n}^{*}},$$

where $ETS_{max}^{**} \leq ETS_{mix}^{**} \leq ETS_{min}^{**}$, because $C_{Min}^{**} < C_{Mix,}^{**} < C_{Max,}^{**}$ (see Example 1). We may have the extreme cases of ETS_{max}^{**} and ETS_{min}^{**} when n^*-values of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$ are widely spread out, and the minimum or maximum value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$ is comparatively very small or large, far away from the mean. In this situation, one may avoid the use of the two extreme cases and prefer to use middle value ETS_{mix}^{*} for the statistical inference in combining the value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$. Note the weights $C_{Mix,1}^{**}, \dots, C_{Mix,n}^{**}$ are each term weights for each TS_i^{**} of all member p_n^{**} (see Example 1, n=5 fifth row, for all 5 members, under each column of p-values).

3.4 Table 1, Numerical Example of Correction Factors C^{**}

The Table 1 below shows the numerical calculation to construct the test statistic TS (C), $TS^{**}(F)$, correction factors C^{**}, using chi-square value (C) for TS and Fisher's Model (F) for TS^{**} and the clinical trial data for $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$ (Example 1, Section 4).

One reason of presenting Table 1 here is to remind the users of Fisher's Model (FM) to be more careful if the data are correlated or non-iid variables. Often we find that, especially in medical journals, many people are still using FM without proper consideration of the problem as if data are iid random variables, Table 1 can be used to correct non-iid problems of their data when they use FM in combining p-values. Another reason to have Table 1 here is to help

understanding the text of next Section 4.

Table 1 shows the three numbers, FM (F), Chi-square model (C), and correction factor (C^{**}), by the p-vales on the columns, i.e., p= 0.01,0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9, and the 15 numbers on the rows, i.e., n =1, 2, 3, 4, 5, 6, 7, 8, 9,10,11,12,13,14,15, each n-number means the same n p-values.(See Appendix A for the reason why we use the same p for n times). Recall that

$F = TS^{**}$ = T(g($p^{**}|\theta$), $H_O^{**}, \alpha^{**}, \rho, n^{**}$), test statistic for Fisher's Model g($p^{**}|\theta$), for given p^{**}, $H_O^{**}, \alpha^{**}, \rho, n^{**}$,

C = TS = T($h(p|\theta)$,α, n) of assumed base distribution $h(p|\theta)$ given α, n,

C^{**}= $\frac{F}{C} = \frac{TS^{**}}{\mathrm{TS}}, 1$

$0 < C^{**} < \infty$, in the Table 1, is the correction factor expressed as ratio of F

and C to compare them on the equal bases, (i.e., n=n**, and $\alpha=\alpha^{**} = p_i^{**}, \mathrm{i} = 1, \dots, n^{**}$),

except correlation ρ and the forms of models g(.) and h(.), on the interval,$1 < C^{**} < \infty$,

this condition implies that C^{**} shows only impacts of correlation and model difference.

Note that here we use the five same values of p to induce the maximum correlation to F in C^{**}=$\frac{F}{c}$, while C remains the same, hence giving larger C^{**}, which, in turn, provides conservative or smaller ETS^{**}= $\frac{F}{C^{**}}$. Thus, users of C^{**}, in Table 1 will have conservative effective test statistic, ETS^{**}, when F is corrected by C^{**}.

To illustrate for the calculation of F, C, and C^{**}in Table 1, we take one cell for n=5, the fifth row and p=0.05 on the third column, Fisher's Model (FM), F = -2log 0.05 0.05 0.05 0.05 0.05)=29.96, using the same values five time for n=5 for the reason given above. The basic distribution, Chi-square value (C), C =18.31, for x_{2n}^2, 2n=10 degrees of freedom at $\alpha=\alpha^{**} = p_i^{**} = p^{**}$ = 0.05, from the table. The result is C^{**}= $\frac{F}{C}$= $\frac{29.96}{18.31}$.=1.64 as shown in the 5th row, n=5, and third column p=0.05 in Table 1. Other cells in Table1 follow the same steps to obtain F, C, and C^{**}.

Note we set the sample size n=n**=5, test level $\alpha=\alpha^{**} = p_i^{**} = 0.05$, to compare C and F on the equal bases except the correlation and the forms of two models, g(.) and h(.), i.e., g(.) ≠ h(.). Thus, the C^{**} shows the impacts of correlation and the wrong assumption of the model F in comparison to C.

We call C**=$\frac{F}{C} = \frac{TS^{**}}{TS}, 1 < C^{**} < \infty$, correction factor as they are indirectly used to correct or reduce TS^{**} for the violation of iid conditions and model assumption, for the data p_5^{**} = (05,.08, 0,09 0.10, 0.20), (see Appendix B). Effective Test Statistic (ETS^{**}=$\frac{TS^{**}}{C^{**}}$) is finally used for statistical inference. Note $ETS^{**} > TS, 1 < C^{**} < \infty$. (Appendix A).

Table 1. shows Fisher's Model F=TS^{**} and Chi-square Table value C= TS, and Correction Factors C**= F/C by the size of the nine p's, p=0.01, …, 0.9 on the columns, and the 15 numbers n=1,…,15 for the same n p-values on the rows

n of p	α =p→	0.01	0.02	0.05	0.1	0.2	0.3	0.5	0.7	0.9
n=1	F	9.21	7.82	5.99	4.61	3.22	2.41	1.39	0.71	0.21
	C	9.21	7.82	5.99	4.61	3.22	2.41	1.39	0.71	0.21
	C**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
n=2	F	18.42	15.65	11.98	9.21	6.44	4.82	2.77	1.43	0.42
	C	13.28	11.67	9.49	7.78	5.99	4.88	3.36	2.19	1.06
	C**	**1.39**	**1.34**	**1.26**	**1.18**	**1.07**	**0.99**	**0.83**	**0.65**	**0.40**
n=3	F	27.63	23.47	17.97	13.82	9.66	7.22	4.16	2.14	0.63
	C	16.81	15.03	12.59	10.64	8.56	7.23	5.35	3.83	2.20
	C**	**1.64**	**1.56**	**1.43**	**1.30**	**1.13**	**1.00**	**0.78**	**0.56**	**0.29**
n=4	F	36.84	31.30	23.97	18.42	12.88	9.63	5.55	0.56	0.84
	C	20.09	18.17	15.51	13.36	11.03	9.52	7.34	5.53	3.49
	C**	**1.83**	**1.72**	**1.55**	**1.38**	**1.17**	**1.01**	**0.76**	**5.53**	**0.24**
n=5	F	46.05	39.12	29.96	23.03	16.09	12.04	6.93	3.57	1.05
	C	23.21	21.16	18.31	15.99	13.44	11.78	9.34	7.27	4.87
	C**	**1.98**	**1.85**	**1.64**	**1.44**	**1.20**	**1.02**	**0.74**	**0.49**	**0.22**
n=6	F	55.26	46.94	35.95	27.63	19.31	14.45	8.32	4.28	1.26
	C	26.22	24.05	21.03	18.55	15.81	14.01	11.34	9.03	6.30
	C**	**2.11**	**1.95**	**1.71**	**1.49**	**1.22**	**1.03**	**0.73**	**0.47**	**0.20**
n=7	F	64.47	54.77	41.94	32.24	22.53	16.86	9.70	4.99	1.48
	C	29.14	26.87	23.68	21.06	18.15	16.22	13.34	10.82	7.79
	C**	**2.21**	**2.04**	**1.77**	**1.53**	**1.24**	**1.04**	**0.73**	**0.46**	**0.19**
n=8	F	73.68	62.59	47.93	36.84	25.75	19.26	11.09	5.71	1.69
	C	32 00	29.63	26.3	23.54	20.47	18.42	15.34	12.62	9.31
	C**	**2.30**	**2.11**	**1.82**	**1.56**	**1.26**	**1.05**	**0.72**	**0.45**	**0.18**
n=9	F	82.89	70.42	53.92	41.45	28.97	21.67	12.48	6.42	1.90
	C	34.81	32.35	28.87	25.99	22.76	20.60	17.34	14.44	10.86
	C**	**2.38**	**2.18**	**1.87**	**1.59**	**1.27**	**1.05**	**0.72**	**0.44**	**0.17**
n=10	F	92.10	78.24	59.91	46.05	32.19	24.08	13.86	7.13	2.11
	C	37.57	35.02	31.41	28.41	25.04	22.77	19.34	16.27	12.44
	C**	**2.45**	**2.23**	**1.91**	**1.62**	**1.29**	**1.06**	**0.72**	**0.44**	**0.17**
n=11	F	101.3	86.06	65.91	50.66	35.41	26.49	15.25	0.44	2.32
	C	40.29	37.66	33.92	30.81	27.3	24.94	21.34	18.1	14.04
	C**	**2.51**	**2.29**	**1.94**	**1.64**	**1.30**	**1.06**	**0.71**	**0.43**	**0.17**
n=12	F	110.5	93.89	71.9	55.26	38.63	28.90	16.64	8.56	2.53
	C	42.98	40.27	36.42	33.20	29.55	27.10	23.34	19.94	15.66
	C**	**2.57**	**2.33**	**1.97**	**1.66**	**1.31**	**1.07**	**0.71**	**0.43**	**0.16**
n=13	F	119.7	101.7	77.89	59.87	41.85	31.30	18.02	9.27	2.74
	C	45.64	42.86	38.89	35.56	31.79	29.25	25.34	21.79	17.29
	C**	**2.62**	**2.37**	**2.00**	**1.68**	**1.32**	**1.07**	**0.71**	**0.43**	**0.16**
n=14	F	128.9	109.5	83.88	64.47	45.06	33.71	19.41	9.99	2.95
	C	48.28	45.42	41.34	37.92	34.03	31.39	27.34	23.65	18.94
	C**	**2.67**	**2.41**	**2.03**	**1.70**	**1.32**	**1.07**	**0.71**	**0.42**	**0.16**
n=15	F	138.2	117.4	89.87	69.08	48.28	36.12	20.79	10.7	3.16
	C	50.89	47.96	43.77	40.26	36.25	33.53	29.34	25.51	20.6
	C**	**2.71**	**2.45**	**2.05**	**1.72**	**1.33**	**1.08**	**0.71**	**0.42**	**0.15**

Note in Table 1, C** = F/C is increasing from1.39 to 2.71 when n=2 increases to n=15 on the first column of p=0.01. It means that F is increasing faster than C as the number n of same p-values is increasing. This trend is reversed in the seventh column of p=0.5, C** is decreasing from 0.83 to 0.71 when n=2 increases to n=15. i.e., F decreasing faster than

C.

Similar trend exists on the rows, for the second-row n=2, C** is decreasing from 1.39 to 0.40 when p=0.01 increases to p=0.9. The change point C** greater than 1 to less than 1 is p=0.5, it is true for all the 15 rows.

Note that we ignore when $C^{**} = \frac{TS^{**}}{TS}, 0 < C^{**} < 1$, it happens data are negatively correlated. or

$TS^{**} < TS$ which happens when C^{**} does not reduce the impacts of non-iid inflation on TS^{**}.

4. Examples

Two examples are presented. (1) Effective Test Statistics ETS* of the Fisher's Model (FM) to combine p^*-values from clinical trial data at Minneapolis Veterans Administration (VA) Hospital. (2) Random group method for a large sample of n variables (Choi and Nandram, 2021). Using random grouping, we divide a large sample into k manageable random groups and obtain one p value from each group. Then the k p-values are combined, using FM.

4.1 Example 1. Fisher's Model (Fisher, 1932) to Combine Clinical Trial Results

All Parkinson patients, visiting the Neurology Department of Minneapolis VA hospital, are the population during the study period in 1970 (Choi, 1970). In our example, a sample of 36 patients is randomly selected from all the visitors. The 36 patients randomly ordered and took either Symmetrel, a candidate for Parkinson medication, or placebo, for 20 weeks crossover design, starting by coin toss, one week medication and one week placebo double blindly.

After each week, they took 5 tests: walking, tremor, stiffness, arm movement, and eye movement, to measure the impacts of medication or placebo. These tests are equally weighted assuming no residual effects, and calibrated from one to ten, one for no effect and 10 for the best result. The differences of on and off weeks are measured. Each patient provides 10 differences during 20 trial weeks and obtain one mean difference for each patient.

Again, find one mean differences from 36 patients for each of 5 tests, providing one mean difference from each of 5 tests. Using student-t test for the mean differences under the null hypothesis of no difference, we have 5 p-values from 5 tests, n=5, combined with Fisher's Model (FM), assuming they are iid random variables and follow Chi-square 10 degrees of freedom, x^2_{10}.

We have five p values of t-test under the null hypothesis of no mean differences. Once we have p-values, we ignore the previous procedures to obtain them and they are the random variables of our interest and may have their own distribution. The five p values are p^*_5 = (05,.08, 0,09 0.10, 0.20).

Fisher's model (FM) combines these 5 p-values.

FM = -2 log (0.05 x 0.08 x 0,09 x 0.10 x 0.2)

= $-$ 2(log 0.05 + log 0.08 + log 0.09 + log 0.10 + log 0.20)

=$-2(-2.9957$ - 2.5257 - 2.4080 - 2.3026 - 1.6094)

= 23.6828.

When we compare FM=23.6828 to the assumed Chi-square 10 degrees of freedom at α = 0.01 = 23.209, FM is significant as 23.6828 $>$ 23.209 at α = 0.01 of x^2_{10}.

However, the clinical trial data p^*_5 = (05,.08, 0,09 0.10, 0.20) are correlated (see Appendix B) or non-iid random variables, and thus, we cannot assume FM is distributed as chi-square 10 degrees of freedom. Therefore, FM = 23.6828 should be reduced for the violations of iid condition of p^*_5.

Most data are correlated in the real world as there is hardly any independent data.

But statisticians, in general, blindly assume their data are iid random variables. Thus, it is necessary to check out the independence and other characteristics of their data beforehand.

The three candidates, C^{**}_{Min}, C^{**}_{Max}, and $*\ C^{*}_{Mix}$ of Correction Factor are introduced in 3.3. They are used to reduce FM for iid violations.

(1) $\mathrm{C}^{**}_{\min} = \frac{\boldsymbol{F(min)}}{\boldsymbol{C(min)}} = \frac{-2\log(0.05\ 0.05\ 0.05\ 0.05\ 0.05),\alpha=0.05}{x^2_{10,}(p=\alpha=0.05,n=5)} = 1.64$**,** using minimum($p^*_5$)=0.05.

(2) $C^{**}_{max} = \frac{F(max)}{C(max)} = \frac{-2\log(0.2\ 0.2\ 0.2\ 0.2\ 0.2),,\alpha=0.2}{x^2_{10,}(p=\alpha=0.2,n=5)} = 1.29$, using maximum($p^*_5$)=0.2.

Since individual weights are C^{**}=1 for n=1 (see first row, Table 1), we use an alternative weight.

(3) $C^{**}_{Mix,5} = \frac{F_i}{C_i}$ =1.64, 1.52, 1.46, 1.44. 1.20 (see, C^{**} in Table 1**,** row 5 for n=5 and corresponding columns of p= 0.05, 0.08, 0,09, 0.10, 0.20.

TS^{**} for Fisher's Model result (F) is adjusted by this correction factor (C^{**}) to obtain the effective test statistics (ETS^{**}) as shown below.

First, we find the minimum value of p_5^{**} = (05,.08, 0,09 0.10, 0.20, which is 0.05, and use 0.05 five times to find FM (F) as explained the reason why we use the same number 0.05 five times. Then adjust FM by $C^{**}_{\min}$=1.64 (Table 1, row n=5 and column p=0.05). We have

FM(min=0.05) = -2 log(0.05 0.05 0.05 0.05 0.05) = 2(5 x 2.99573) =29.9573,

$$\text{ETS**(min=0.05)} = \frac{TS^{*}(0.05)}{C^{**}_{\min=0.05}} = \frac{29.9573}{1.64} = \mathbf{18.2667.}$$

Second, similarly, the maximum value 0.2 is used five times in FM, and FM(max=0.2)

is adjusted by $C^{**}_{max=0.2}$=1.29 (Table 1, row n=5 and column p=0.2). We have

FM(max=0.2)= -2 log(0.2 0.2 0.2 0.2 0.2)= 2(5x 1.60944)=16.0944.

$$\text{ETS**(max=0.2)} = \frac{TS^{*}(0.2)}{C^{**}_{max=0.2}} = \frac{16.0944}{1.29} = \mathbf{12.4763.}$$

Third, we obtain FM(mix) of individual value adjusted by individual combined weights C_i^{**} = 1.4, 1.52, 1.46,1.44,1.20 (Table 1, row 5, n =5 and columns corresponding to 0.05, 0.08, 0.09, 0.1, 0.2). The main reason why we use individual combined weights is, when n=1, individual weights C^{**}=1 regardless of p-vales. One sample is always independent so both FM and assumed chi-square distribution remain the same for given test level when sample size is one (see Table 1, row 1, C^{**}=1 for all p=values). We have

FM(mix) = -2 {log 0.05 + log 0.08 + log 0.09 + log 0.1 +log 0.2}, and each term is divided by the corresponding individual combined weight for the given reason. Hence, we have been

$$\text{ETS**(mix)} = \frac{-2\{\log 0.05\}}{1.64} + \frac{-2\{\log 0.08\}}{1.52} + \frac{-2\{\log 0.09\}}{1.46} + \frac{-2\{\log 0.1\}}{1.44} + \frac{-2\{\log 0.2\}}{1.20}$$

$$= \frac{2x2.99573}{1.64} + \frac{2x2.5257}{1.52} + \frac{2x2.408}{1.46} + \frac{2x2.3026}{1.44} + \frac{2x1.6094}{1.20}$$

$$= \frac{5.99146}{1.64} + \frac{5.0514}{1.52} + \frac{4.816}{1.46} + \frac{4.6052}{1.44} + \frac{3.2188}{1.20}$$

=3.4521+3.3233 +3.2986 +3.1981 +2.6823 =**15.9544.**

Results show that

ETS**(max=0.2)= **12.4763** < ETS**(mix) =**15.9544** < ETS** (min=0.05)= **18.2667.**

ETS**(min=0.05)=18.2667 is significant at α =0.05 of $x^2_{10,}$ (=18.307), but other two, ETS**(max=0.2)= **12.4763** and ETS**(mix) =**15.9544** are not significant.

In the beginning of this example, Fisher's Model gives FM=23.6826, without correction, which is significant at α = 0.01 of x^2_{10} (23.209). This FM is very much inflated when compared to above corrected results. Only one not-corrected value 29.9537 of FM(min=0.05) is bigger than the not-corrected FT=23.6826.

When the Maximum, here 0.2 or Minimum, 0.05, of p-values are too far away from the mean or relatively too small or too big, one may prefer the mixed value, ETS**(mix)=15.9544, for statistical inference, which is not significant at α = 0.01 of x^2_{10} (23.209), even at α =0.05 of $x^2_{10,}$ (=18.307).

4.2 Example 2. P-values from Random Groups of a Large Sample

When the existing methods, for example normal test or student t-test, are used for statistical inference, we encounter the large sample problems (Choi and Nandram, 2021). The reason is such test is the function of its variance, which in turn, function of sample size. The variance becomes too small when the sample size is large or too large when sample size is

too small. We consider the case of too large sample size, and test statistic becomes significant for the sample size over certain level (Choi and Nandram, 2021).

4.2.1 The Large Sample Problem

We indicate the large sample problem and show a solution using Random Group Method (Choi and Nandram, 2021). A concrete example is as follows. Let $x_1, x_2, \ldots, x_n$ be the realization of iid random variables $X_1, X_2, \ldots, X_n$, distributed as N(μ, σ^2), where σ^2 is known and inference is required about $\boldsymbol{\mu}$. We test the null hypothesis H_0: $\mu = \mu_0$ against alternative H_1: $\mu < \mu_0$. Let $\bar{\boldsymbol{x}}_{\boldsymbol{0}}$ be observed value of the sample mean, $\bar{x}$. Then the p-value of the test is

$$\mathrm{P}(\bar{\boldsymbol{x}} \leq \bar{\boldsymbol{x}}_{\boldsymbol{0}} \mid \boldsymbol{H_0})$$

$$= \mathrm{P}\left(\frac{\bar{x} - \boldsymbol{\mu_0}}{\boldsymbol{\sigma}/\sqrt{\boldsymbol{n}}} \leq \frac{\bar{\boldsymbol{x}}_{\boldsymbol{0}} - \boldsymbol{\mu_0}}{\boldsymbol{\sigma}/\sqrt{\boldsymbol{n}}} \mid \boldsymbol{H_0}\right)$$

$$= \boldsymbol{\Phi}\{\sqrt{\boldsymbol{n}}\ (\bar{\boldsymbol{x}}_{\boldsymbol{0}} - \boldsymbol{\mu_0})\ \}$$

Here $\boldsymbol{\Phi}(.)$ is the cdf of standard normal random variable. Therefore, if n is very large and $\bar{\boldsymbol{x}}_{\boldsymbol{0}} \leq \boldsymbol{\mu_0}$, p-value $\approx \boldsymbol{0}$ which shows large sample problem (Choi and Nandram, 2021). We use the following steps to solve this problem.

Step one

We divide a large sample of size n into a number of random groups so that each can be tested by the usual method. Let $\boldsymbol{x_n} = x_1, x_2, \ldots, x_n$ be a large sample of size n from N(μ, σ^2). When n is a large number, we cannot do the usual test. We want to divide the sample into h smaller samples of size m, 1< m < n, using Random Group Method. The smaller samples enable us to perform a traditional test (e.g., Normal test, t-test) for testing a hypothesis, H_0: $\mu = \mu_0$. Choi and Nandram (2021) showed how to divide the large sample into h smaller samples. Each sample provides one test statistic

$$t_i = \mathrm{T}(f(\mathrm{p}|\ \mu,\ \sigma^2),\ m_i,\ H_0, \alpha),\ m_i = \mathrm{m}, \quad \mathrm{i} = 1, \ldots, \mathrm{h}.$$

and the h test statistics provide h test scores $p_1, \ldots, p_h$ at the test level $\alpha_i = \alpha$, i=1, …,h.

Step two

When h p-values are iid variables, we can use Fisher's Model is assumed to be chi-square 2h degrees of freedom. We assume random groups are independent, we may assume h p-values are also independent, p = $p_1, \ldots, p_h$ are distributed as chi-square distribution, f(p|θ). We can make statistical inference with chi-square test result. However, If the p-values are correlated, we can use the correction factor in Table 1, to correct such impacts on Fisher's Model value.

Numerical example

A student presented data analysis of three sets of data; each includes 1500 persons' dental records. All the three t-tests of hypothesis H_0: $\mu = \mu_0$ were significant due to large sample size. Suggestion was to randomly divide 1.500 into 50 groups of 30 persons. If out of 50 t-tests, 45 tests (90%) of the 50 tests were significant at p=0.05, then it is also 90% significant for the 1500 persons' data at the same level at p=0.05 (Choi and Nandram, 2021). Similarly, it can be done for the remaining two groups.

5. Bayesian Model for Combining P-values

The Bayesian paradigm has the advantage of coherence, but the calculation of p-values is incoherent within the Bayesian paradigm because the computation of a tail area of a posterior distribution is not coherent. This is why Bayesians have hardly worked on this problem; see Casella and Berger (1987) and the discussions that followed. The combined p-value is an appropriate posterior mean, μ, say. However, note that μ is a parameter in the Bayesian paradigm, and it is a random variable.

It is not simple to include a correlation among the p-values since the sample of p-values is small. For the non-Bayesian method, we have constructed a correlation based on a distance measure (see Appendix B); otherwise, it is impossible to estimate this correlation. Here we will separate the data into groups to get an intra-cluster correlation.

The problem of combining a number of p-values, from the studies on the same subject, is one of data integration, which is currently a hot topic, see, for example, Nandram et al (2021) for model-based methods using both non-Bayesian and Bayesian approaches.

5.1 The Case of Independence

Suppose that we have the results of p-values $\hat{p}_1, \ldots, \hat{p}_n$ from n data sets, and these values are independent. We can also use an appropriate prior to reflect previous procedures to obtain p-values.

Let iid $\hat{p}_1,\ldots, \hat{p}_n \sim$Beta$\{\mu\frac{1-z}{z}, (1-\mu)\frac{1-z}{z})\}$ and E$(\hat{p}_i) = \mu$, $0 \leq \mu, z \leq 1$.

This is a useful reparameterization of the parameters of the Beta distribution in which both (μ, z) lie in (0,1), which leads to easy computation. See Nandram (2016) where this reparameterization was first introduced. A priori, we assume that

$$\mu, z \sim \text{U}(0,1),$$

essentially a non-informative prior.

We want to make inference about μ, combined p values. Letting $\hat{p}_a = \prod_{i=1}^{n} \hat{p}_i$, and $\hat{p}_b = \prod_{i=1}^{n}(1-\hat{p}_i)$, the posterior density of (μ, z) is

$$\pi(\mu,z|\hat{p}) \sim \left[\frac{\Gamma(\frac{1-z}{z})}{\Gamma\mu(\frac{1-z}{z})\Gamma(1-\mu)(\frac{1-z}{z})}\right]^n \hat{p}_a^{\mu\frac{1-z}{z}-1} \hat{p}_b^{(1-\mu)\frac{1-z}{z}-1}, 0 \leq \mu, z \leq 1.$$

For the samples from the posterior density, one can also use the Gibbs sampler (Casella and George,1992) to obtain μ and z for given p-values; but we use a random sampler that does not need any convergence monitoring.

The posterior summaries we use are the posterior mean (PM), posterior standard distribution (PSD), posterior coefficient of variation (PCV) and 95% highest density interval (HPDI).

Consider Example 1 on combining the five p-values, .05, .08, ,09, .10, .20. Applying our method based on the Beta model to these p-values, we computed the combined p-value, μ, and the posterior summaries are PM=.121, PSD=.032, PCV=.266, HPDI=(.069, .191). Therefore, the null hypothesis is not significant at the 5% significant level and perhaps not even at 10% significant level.

Table 2 has results of a small simulation study, which is used to provide many different examples. We generated n p-values, n=10,…,100, and we compare the combined p-value, the posterior mean of μ; we also look at z. Again, we show posterior summaries in Table 2 of the two variables, μ *and* Z, by sample size on the columns, and posterior mean (PM), posterior standard deviations (PSD), coefficient of variations (PCV) and 95% HPDIs of μ and z on the rows. Again, not that μ-values represent the posterior mean of the p-values, which range $0.05529 < \mu < 0.09157$. Note that the PSDs are decreasing as the sample size n increases. This also gives smaller PCVs and narrower 95% HPDIs e.g., at n=2 the 95% HPDI for μ is (.02945, .16355).

Table 2. Posterior summaries of μ *and* z including intervals

Sample size n	**PM**	**PSD**	**PCV**	**95% Lower bound**	**95% Upper bound**
n=10 μ	**0.09157**	**0.03414**	**0.37282**	**0.03945**	**0.16355**
z	0.09908	0.05641	0.56934	0.02105	0.20441
n=20 μ	**0.06136**	**0.01395**	**0.22729**	**0.04007**	**0.09056**
z	0.05462	0.02196	0.40201	0.02156	0.09700
n=30 μ	**0.05716**	**0.01028**	**0.17992**	**0.04096**	**0.07916**
z	0.04721	0.01501	0.31800	0.02101	0.07358
n=40 μ	**0.05810**	**0.00821**	**0.14122**	**0.04099**	**0.07149**
z	0.03979	0.01092	0.27439	0.02117	0.06064
n=50 μ	**0.05596**	**0.00675**	**0.12061**	**0.04206**	**0.06934**
z	0.03771	0.00901	0.23902	0.02110	0.05349
n=60 μ	**0.05545**	**0.00640**	**0.11540**	**0.04149**	**0.06795**
z	0.03787	0.00818	0.21608	0.02107	0.05085
n=70 μ	**0.05975**	**0.00616**	**0.10310**	**0.05117**	**0.07057**
z	0.04001	0.00760	0.19006	0.03101	0.06021
n=80 μ	**0.05529**	**0.00616**	**0.11149**	**0.04092**	**0.06617**
z	0.04357	0.00808	0.18538	0.03098	0.05878
n=90 μ	**0.05751**	**0.00571**	**0.09927**	**0.04879**	**0.07038**
z	0.04436	0.00798	0.17976	0.03099	0.05867
n=100 μ	**0.05859**	**0.00573**	**0.09778**	**0.05099**	**0.07015**
Z	0.04580	0.00778	0.16985	0.03101	0.05922

We may be able to include all information of first stage as prior replacing $\mu, z \sim$U(0,1). This

Will be done in a future study. We can use independent Beta distributions with specified parameters, and this will depend on the amount of information available.

To motivate the case, where we include an intra-class correlation, we provide another Bayesian analogue of Fisher's model of combining p-values. Let $p_{i,}$ i= 1,…,n, denote the n p-values, and let $q_i = \log\{p_i/(1-p_i)\}$, independent, then a simple model is

$$q_i \mid \mu, \sigma^2 \sim \text{Normal}(\theta, \sigma^2)$$

$$\pi(\theta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

This is a standard non-informative prior (a version of Jeffrey's objective prior), but as always leading to proper posterior distribution for (θ, σ^2).

Here the combined p-value is $\emptyset = e^{\theta}/(1+e^{\theta})$. The posterior density of θ is a Student's t density, and inference about $\emptyset$ is obtained by sampling the Student's t density and computing $\emptyset$. For the example on the five p-values, for inference about $\emptyset$, we have posterior summaries, which are PM=0.099, PSD=0.033, PCV=0.334, HPDI=(0.044, 0.162). Again, the test is not significant at the 1 % significant level.

5.2 Including Correlation

We add an intra-cluster correlation as follows**.** We find all l = n(n-1)/2 distinct pairs of $q_i, \ldots, q_n$, and we form a Bayesian one-way random effect model, each cluster having just two values. Let y_{i1}, y_{i2}, i= 1,…,l, denote the distinct pairs which form the clusters. Then we assume the model,

$$y_{i1}, y_{i2} \mid \mu_i, \sigma^2 \widetilde{ind} N(\{\mu_i, (1- \rho)\sigma^2\}$$
$$\mu_i \mid \theta, \sigma^2, \rho \widetilde{ind} \text{N}(\theta, \rho\sigma^2), \quad \text{i= 1,…,}l,$$

$$\pi(\theta, \sigma^2, \rho) \propto \frac{1}{\sigma^2}.$$

It is important to note that $\text{cor}(y_{i1}, y_{i2} \mid \vartheta, \sigma^2, \rho) = \rho$ in (0,1). We have actually used the traditional non-informative prior for $\pi(\theta, \sigma^2, \rho)$; this prior causes no impropriety issues (see Nandram, Toto and Choi, 2011) for proofs.

Also, note that we are actually assuming a composite likelihood because the pairs are not independent (i.e., each pair has one common unit), for example, see Varin, Reid and Firth (2011) for a discussion of composite likelihood. Again, the combined p-value is $\emptyset = e^{\theta}/(1+e^{\theta})$. This is the same as for the case when no correlation is assumed.

Using Bayes' Theorem, the joint posterior density is

$$\pi(\boldsymbol{\mu}, \theta, \sigma^2, \rho \mid \mathbf{q}) =$$
$$\pi_1(\boldsymbol{\mu} \mid \theta, \sigma^2, \rho \mid \mathbf{q})\ \pi_2(\theta \mid \sigma^2, \rho \mid \mathbf{q})\ \pi_3(\sigma^2 \mid \rho \mid \mathbf{q})\ \pi_3(\sigma^2 \mid \rho \mid \mathbf{q})\ \pi_4(\rho \mid \mathbf{q}).$$

Here, $\pi_1(\boldsymbol{\mu} \mid \theta, \sigma^2, \rho \mid \mathbf{q})$, $\pi_2(\theta \mid \sigma^2, \rho \mid \mathbf{q})$, and $\pi_3(\sigma^2 \mid \rho \mid \mathbf{q})$, have simple forms, and $\pi_4(\rho \mid \mathbf{q})$ has nonstandard form but it can be sampled using a grid method (e.g., Nandram, Toto and Choi, 2011). It is also true that the joint posterior density is proper, provided $l \geq 2$, see Nandram, Toto, and Choi (2011). Therefore, it is easy to sample the posterior density of θ and so $\emptyset$. To make inference about $\emptyset$, we draw 10,000 samples of the posterior density of $\emptyset$. No monitoring is required because a Markov chain Monte Carlo sampler is not used.

As summaries of the posterior density of $\emptyset$, we have PM=0.078, PSD=0.017, PCV=0.217, and the 95% HPDI= (0.048, 0.112). Therefore, the combined test is not significant at 5% significant level. Note that when we assume no correlation, PM=0.099 a bit larger, and the HPDI= (0.044, 0.162) a bit wider. The posterior summaries of ρ are PM=0.147, PSD=0.125, PCV=0.851, 95% HPDI=(0.001, 0.603); so, there is a small correlation.

As another example, when we increased the number of p-values to 10 (i.e., duplicate the five p-values to get 05, .08, ,09, .10, .20, 05, .08, ,09, .10, .20); there is an increase in precision but the results remain essentially the same. The posterior summaries of ρ are PM= 0.147, PSD= 0.125, PCV= 0.851, 95% HPDI= (0.001, 0.393); so that there is a small correlation, not much of a difference

6. Conclusion

We have used a model combine test scores on the same topic. Here, we assume a distribution for the data model. We compare the two test statistics, one from assumed distribution h(.) of iid-data and other from pseudo-distribution g(.) of

non-iid data. We define the differences between them as the ratio of the two. As the actual data may include impacts of not only correlation but also other difference of iid and non-iid conditions. We describe how to reduce the test statistics of non-iid data to make statistical inferences with the assumed distribution of iid variables.

We have considered two-stage procedure. The first stage is sampling and pre-processing to obtain the p-values. The second stage is the analysis of the first stage results.

Suppose that h independent samples. $y_1, \ldots, y_{n_i}$, i=1,…,h, are randomly taken from the population for an investigation on a same subject and suppose the sample follows true distribution f(y|θ). Each sample provides one test result from significant testing at a critical level α under a null hypothesis, providing test statistics.

$$t_i = \mathrm{T}(f(y_i|\theta), H_0, \alpha, n_i), \ \alpha_i=\alpha, \ \mathrm{i}=1,\ldots, \mathrm{h},$$

These test statistics provide h p*-values,

$$\alpha = 1 - \int_{-\infty}^{t_i} f(y_i|\theta) \ \mathrm{d}y_i, \ \mathrm{i} = 1,\ldots,\mathrm{h}.$$

Some assume the two stages are connected and the second stage is a continuation of the first. If the information such as sample design, sample, $f(y_i|\theta), H_0, \alpha,$ and sample size n_i are available, we can use this information in the second stage to combine the p_i^*-values to increase efficiency. Yoon et al.(2021) incorporate sample size n_i to combine p*-values. If one wants to include other information in Bayesian modeling, it is possible to use them as prior information.

The validity check of these estimations can be added in the future extension using the variance or coefficient of variation, and 95% confidence interval of each estimation through simulation.

It will be useful to carry out further study of the combination of correlated p-values in the Bayesian paradigm. For one thing, it will allow us to incorporate further information that can improve posterior inference. When available, information such as sample size and site covariates can be included in the combination of correlated p-values.

References

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association, 82*(397), 106-111.

Casella, G., & Berger, R. L. (2002). Statistical Inference, Second Edition. Duxbury, Pacific Grove, Ca.

Casella, G., & Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*(3), 167-174.

Choi, J. (1970). Effectiveness of Symmetrel for Parkinson patients. Technical Report, Parkinson Laboratory, Neurology Department, VA Hospital, Minneapolis.

Choi, J. W. (1980). Ph. D. Thesis, University of Minnesota.

Choi, J. W., & McHugh, R. (1989). An adjustment factor for goodness and independent test for correlated and weighted observations. *Biometrics, 43*, 976-996.

Choi, J., & Nandram, B. (2021). Large sample problems. *International Journal of Statistics and Probability, 10*(2), 81-89.

Fisher, R. A. (1932). Statistical Methods for Research Workers (4th Edition), London: Oliver and Boyd.

George, E. O. (1977). *COMBINING INDEPENDENT ONE-SIDED AND TWO-SIDED STATISTICAL TESTS--SOME THEORY AND APPLICATIONS*. University of Rochester.

Hartung, J., & Knapp, G. (2005). Models for combining results of different experiments: retrospective and prospective. *American Journal of Mathematics and Management Sciences, 25*, 149-188.

Hartung, J., Bockenhoff, A., & Knapp, G. (2003). Generalized Cochran-Wald statistics in combination of experiments. *Journal of Statistical Planning and Inference, 113*, 215-237**.**

Heard, N. A., & Rubin-Delanchy, P. (2018). Choosing between methods of combining-values. *Biometrika, 105*(1), 239-246.

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Boston Academic Press.

Held, L., Pawel, S., & Schwab, S. (2020). Replication power and regression to the mean. *Significance, 17*(6), 10-11.

Hess, A., & Iyer, H. (2007). Fisher's combined p-value for detecting differentially expressed genes using Affymetrix expression arrays. *Bmc Genomics*, *8*(1), 1-13.

Higgins, J. P., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in medicine*, *23*(11), 1663-1682.

Iyer, H. K., Wang, C. J., & Mathew, T. (2004). Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, *99*(468), 1060-1071.

Li, Z., & Begg, C. B. (1994). Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *Journal of the American Statistical Association*, *89*, 1523-1527.

Lipt åk, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, *3*, 171-197.

Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Computational statistics & data analysis*, *47*(3), 467-485.

Matthews, R. (2021). The p-value statement, five years on. *Significance*, *18*(2), 16-19.

Nandram, B. (2016). Bayesian predictive inference of a proportion under a two-fold small area model. *Journal of Official Statistics*, *32*(1), 187-208.

Nandram, B., Choi, J. W., & Liu, Y. (2021). Integration of nonprobability and probability samples via survey weights. *International Journal of Statistics and Probability*, *10*(6), 4-17**.**

Nandram, B., Toto, M. C., & Choi, J. W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation*, *81*, 1593-1608.

Rustagi (Ed.) (1979). Symposium on Optimizing Methods in Statistics, Academic Press, New York, 1979. 345-366.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. A., Jr. (1949). The American Soldier, Volume I. Adjustment during Army Life. Princeton, N.J.: Princeton University Press.

Tippett, L. H. C. (1931). The Methods of Statistics. London: Williams and Norgate Ltd.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*(1), 5-12.

Vovk, V., & Wang, R. (2020). Combining p-values via averaging. *Biometrika*, *107*(4), 791-808.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, *48*, 156-158.

Yoon, S., Baik, B., Park, T., & Nam, D. (2021). Powerful p-value combination methods to detect incomplete association. *Scientific reports*, *11*(1), 1-11.

Appendix A, outline for the proof of Lemma

Correlation, Model 1

Consider the correlated random variables $p^* = (p_1^*, \dots, p_n^*)$. Choi and McHugh (1989) show how to adjust the TS_α^* based on correlated variables in Chi-square testing. Test Statistic (TS_α^*) for correlated data p^* is largely inflated and corrected by the correction factor C= [1 + ρ(n-1)] , ρ is the correlation among n p^*-values. $1 < C < \infty$.

$ETS_\alpha^* = \frac{TS_\alpha^*}{C}$. ETS_α^* can also be obtained by effective sample n_e of n, $n_e = \frac{n}{C}$. (Choi, 1980).

Non-iid case, correlation and other non-iid violations, Model 2

Here, we try to find the non-iid problem of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, indirectly through its test statistics TS^{**}, which is compared to test statistic TS of iid variables. The total difference between the two test statistics, TS^{**} and TS, can be expressed as the ratio of these two, $C^{**} = \frac{TS^{**}}{\mathrm{TS}}$, is used to get effective test statistics (ETS), which is used for statistical inference with $h(p|\theta)$.

$C^{**} = \frac{TS^{**}}{\mathrm{TS}} = \frac{\mathrm{T}*(\mathrm{g}(p^{**}|\theta), H_O^{**}, \alpha^{**}, \rho, n^{**}),}{\mathrm{T}(h(p|\theta), \alpha, n)}$.

The ratio, $C^{**} = \frac{TS^{**}}{\mathrm{TS}}$, $0 < C^{**} < \infty$ We consider C^{**} only on $1 \le C^{**} < \infty$, for positive correlation or TS^{**}> TS.

We do not consider or ignore TS^{**} < TS on $0 < C^{**} < 1$, for it does not reduce inflated TS^{**} for the impacts of non-iid violation (see Proof below). It happens also for negative correlation in C= [1 + ρ(n-1)] (see Method 1).
To prove TS^{**}> TS, consider two disjoint intervals, $(0 < C^{**} < \infty) = \{(0 < C^{**} < 1)\ U\ (1 \leq C^{**} < \infty)\}$.

Let the effective test statistic be $ETS^{**} = \frac{TS^{**}}{C^{**}}$, and correction factor be $C^{**} = \frac{TS^{**}}{\mathrm{TS}}$.

It is easy to see that $ETS^{**} < TS^{**}$ from $ETS^{**} = \frac{TS^{**}}{C^{**}}$, TS^{**} < TS from $C^{**} = \frac{TS^{**}}{\mathrm{TS}}$, on the interval $(0 < C^{**} < 1)$.

Similarly, $ETS^{**} \geq TS^{**}$ and TS^{**} > TS, on the other interval $(1 \leq C^{**} < \infty)$.

The difference between TS^{**} and TS, $C^{**} = \frac{TS^{**}}{\mathrm{TS}} = \frac{\mathrm{T}(g(p^{**}|\theta), H_O^{**}, \alpha^{**}, \rho, n^{**}),}{\mathrm{T}(h(p|\theta), \alpha, n)}$, C^{**} is less than 1 or greater than 1 depending also on n, p = α, and the increasing or decreasing speed of TS^{**} and TS (see Table 1).
If all the above conditions of TS^{**} and TS are same except ρ of $p_i^{**}s$, ignoring H_0^*, and $\alpha = \alpha^{**} = p^{**}$, and n = n^{**}, the proof depends only on correlation ρ : $0 \leq \rho(p_i^{**}, p_{i'}^{**}) \leq 1$, i $\neq i'$, for i, i' = 1, …, n. Model 1 can be used in this case.

(1) If $\rho = 0, C^{**} = \frac{TS_{\alpha^{**}}^{**}}{TS_\alpha} = \frac{\mathrm{T}(g(p^{**}|\theta),,\alpha^{**},\rho,n^{**})}{\mathrm{T}(h(p|\theta),\alpha,n)} = 1$,. It is also true $C^* = 1$ when n=1. The sample size one is always independent, ρ = 0 and $\mathrm{T}(h(p|\theta),,\alpha,n=1) = (g(p^{**}|\theta),,\alpha^{**},\rho=0,\ n^{**}=1)$ for g(.) =h(.) and $\alpha = \alpha^{**} = p^{**}$=p. This is the only time that FM for g(.) assumed correctly to be distributed as c hi-square C for h(.)

(2) If $0 < \rho \leq 1$ and $2 \leq n$, the correction factor C*=$1 + \rho(n-1), 1 < C^* < \infty$ (Choi and McHugh 19 89) and, if $\alpha = \alpha^{**}$= p = $p_i^{**}, i = 1 \ldots, n^{**}, and\ n^{**} = n$, the effective test statistic $ETS_{\alpha^{**}}^*$= $\frac{TS_{\alpha^{**}}^{**}}{C^*} = \frac{TS_{\alpha^{**}}^{**}}{1+\rho(n-1)}$, C^* reduces the correlation impact of $TS_{\alpha^{**}}^{**}$.

For example: If the correlation among the 5 p-values of data 0.05, 0.08, 0.09, 0.10, 0.20, is ρ=0.42 (Appendix B). The correction factor $C = 1 + \rho(n-1) = 1 + 0.42(5-1) = 2.68$ and the Fisher's Model Test Statistic FM= $TS_{\alpha^{**}}^{**}$=23.68 is reduced as, $ETS_{\alpha^{**}}^{**}$= $\frac{23.68}{2.68}$=8.8361, this effective Test Statistic not significant at α =0.01 of $x_{10,}^2$ (=23.209).

If $\rho = 1$, for n= 5, C= $[1 + \rho(n-1)] = 1 + 1.0(5-1) = 5.00$, which is the largest correction value for any given n, and it ,in turn, gives the smallest $ETS_{\alpha^{**}}^*$ = $\frac{23.68}{5}$ = 4,74.

(3) We can also use the effective sample size n_e^*, $n_e^* = \frac{n^*}{C^*}, 1 \leq C^* < \infty$ to obtain ETS^* (Choi, 1980) .

(4) The turning point also depends on the increasing or decreasing speed of $TS_{\alpha^{**}}^{**}$ and TS_α, $TS_{\alpha^{**}}^{**} < TS_\alpha$ when $0 < C^{**} < 1$ and $TS_{\alpha^{**}}^{**} > TS_\alpha$ when $1 < C^{**} < \infty$. We can ignore the case $TS_{\alpha^{**}}^{**} < TS_\alpha\ on\ 0 < C^{**} < 1$, as it happens for negative correlation of p^{**} variables. The change point from less than 1 to more than 1 also depends on the sample size n^{**} and size of p^{**}, for example, Table 1 shows the turning point is $at\ p^{**}$= 0.5 in the column and for all n on the rows,

Appendix B, the correlation of one sample

For one group of data including n variables $p_1, \dots, p_n$, currently there is no formula available to calculate ρ between the variables. We define $\rho_{(p_i p_j)} = \frac{\mathbf{1}}{|\boldsymbol{p_n - p_1}|} \frac{\sum_{i>j}^{n} |\boldsymbol{p_i - p_j}|}{\boldsymbol{n(n-1)/2}}$ for the continuous variables, $p_1, \dots, p_n$.

For example, p = (05,.08, 0,09, 0.10, 0.20),

$$\rho_{(p_i p_j)} = \frac{(0.03+ 9.04+ 0.05+ 0.15)+(0.01+ 0.02+ 0.12)+(0.01+ 0.11)+0.1}{|0.2-0.05)/|(5x4)/2} = \frac{0.27+0.15+0.12+0.1}{0.15x\ 10} = \frac{0.64}{1.5} = 0.4207$$

Appendix C, the three candidates of correction factor

TS = $T(h(p|\theta),\alpha,n)$ of iid random variables p = $(p_1, \dots, p_n)$ remain the same for given test level α and sample size n, while TS**= T(g($p^{**}|\theta$), $H_O^{**}, \alpha^{**}, \rho, n^{**}$) on the non-iid variables $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$

(1) C_{Min}^{**} uses the minimum value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, all n** valuers are the same $p_{min.}^{**}$= $Min(p^{**})$ $= p_{min,i}^{**}$, i=1,…, n^{**}. to obtain the test statistic (TS**). The same minimum values are used to induc e the maximum correlation and in turn conservative TS**. (see Example 1 and Table 1)

$$C_{Min}^{**} = \frac{TS_{\boldsymbol{\alpha^*,min}}^{**}}{TS_\alpha} = \frac{T(g(p_{min,}^{**}, \dots, p_{min,}^{**}|\theta),\ H_O^{**}, \alpha^{**}, \rho, n^{**})}{T(h(p|\theta), \alpha, n)}.$$

(2) C_{Max}^{**} uses the maximum value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, similarly all n** valuers are $p_{max.}^{**}$= $Max(p^{**}) = p_{max,i}^{**}$, i=1,…, n^{**}.

$$C_{Max}^{**} = \frac{TS_{\alpha,max}^{**}}{TS_\alpha}, = \frac{T(g(p_{max.}^{**}, \dots, p_{max.}^{**}|\theta),\ H_O^{**}, \alpha^{**}, \rho, n^{**})}{T(h(p|\theta), \alpha, n\ )}.$$

(3) $C_{Mix}^{**} = C_{Mix,1}^{**} +, \dots, + C_{Mix,n^{**}}^{**}$,

where $C_{Mix,i}^{**} = \frac{TS_{\boldsymbol{i}}^{**}}{TS_n} = \frac{(T(g(p_i^{**}|\theta), H_O^*, \alpha^{**}, \rho, i))}{T_i(h(p_i|\theta), \alpha, n^{**})}$. $i = 1, \dots, n^{**}$.