Biom. J. 33 (1991) 3, 339–345 Akademie Verlag

On Fisher's Method of Combining p-Values

R. C. Elston

Department of Biometry and Genetics,
Louisiana State University Medical Center

Summary

The problem of combining p-values from independent experiments is discussed. It is shown that Fisher's solution to the problem can be derived from a "weight-free" method that has been suggested for the purpose of ranking vector observations (Biometrics 19: 85–97, 1963). The method implies that the value $p = 0.37$ is a critical one: p-values below 0.37 suggest that the null hypothesis is more likely to be false, whereas p-values above 0.37 suggest that it is more likely to be true.

Key words: Chi-square; Ranking vectors; Significance level; Weight-free ranking.

1. Introduction

The p-value, or observed significance level attained in a particular experiment (sometimes also called the posterior significance level), can be defined in words as follows: it is the probability, under the null hypothesis, of obtaining the observed result or any more extreme result. By a "more extreme" result we mean any outcome that would alert the experimenter, even more than did the observed result, to the possibility that the null hypothesis is false.

Now suppose two experiments are conducted independently to test the same null hypothesis. Experiment 1 leads to results 1 with a corresponding p-value p_1, and experiment 2 leads to results 2 with a corresponding p-value p_2. Thus, using the definition of the p-value and noting that the two experiments are independent, we can write

(1) P (results 1 or any more extreme result) $= p_1$
P (results 2 or any more extreme result) $= p_2$

and P (results 1 or any more extreme result *and* results 2 or any more extreme result) $= p_1 p_2$.

Is $p_1 p_2$ therefore the p-value attained as a result of both experiments taken together? It is easy to see that this cannot be so by considering what it would imply if n independent experiments were carried out, and each resulted in a p-value of, say, 0.9. The analogous combined p-value would then be $(0.9)^n$, and we

would arrive at the absurd conclusion that any null hypothesis could be made as significant as we please merely by testing it on the basis of a large enough number of experiments!

Alternatively, let us think of p as the probability of a Type I error. Then the overall probability of a Type I error for both experiments is the probability of such an error in either of the two experiments, and we are led to combining the p-values of the two experiments by taking

(2) $$\begin{aligned} &P \text{ (results 1 or any more extreme result } \textit{or} \\ &\quad \text{results 2 or any more extreme results)} \\ &= p_1 + p_2 - p_1 p_2 \\ &= 1 - (1 - p_1)\,(1 - p_2)\,. \end{aligned}$$

But this also implies an absurdity. Suppose we conduct n experiments and each results in a p-value of, say 0.05. If we take the combined p-value to be $1-(1-0.05)^n$, it is obvious that with this definition we can now make the p-values as close to 1 as we please, again merely by repeating the experiment enough times!

In this article I first introduce the method FISHER (1956) proposed for combining p-values. I then examine what is wrong with the above two paradoxical formulations and show that the problem of combining p-values can be considered identical to the problem of ranking vectors whose elements are in the interval [0, 1]. Finally, I recall a method of ranking vectors proposed over twenty-five years ago (ELSTON, 1963) and note that it leads directly to Fisher's method.

2. Fisher's Method for Combining p-Values

Fisher argued as follows. If the null hypothesis is true, p can be considered as a realization of a random variable P that is uniformly distributed on [0, 1], i.e. whose density function is

$$f_P(p) = \begin{cases} 1 & \text{if } \ 0 \leqq p \leqq 1 \\ 0 & \text{otherwise}\,. \end{cases}$$

Let $Y = -2 \ln P$, so that under the null hypothesis the density function of Y is given by

$$f_Y(y) = f_P(p) \left| \frac{dp}{dy} \right| = \frac{1}{2} e^{-\frac{y}{2}}\,, \quad 0 \leqq y \leqq \infty\,,$$

which is the density function of a chi-square distribution with 2 d.f.

Now if we have two independent statistics, each distributed as chi-square with 2 d.f., we know that their sum is distributed as chi-square with 4 d.f. Fisher therefore proposed comparing $-2 \ln p_1 - 2 \ln p_2$ to the chi square distribution with 4 d.f. Remembering that small values of p correspond to large values of y, the combined p-value is thus

(3) $$P\,(Y_4 \geqq -2 \ln p_1 - 2 \ln p_2)\,,$$

where Y_4 denotes a chi-square random variable with 4 d.f., i.e.

$$f_{Y4}(y) = \frac{1}{4} y e^{-\frac{y}{2}}, \quad 0 \leqq y \leqq \infty .$$

Thus, letting $p_1 p_2 = c$, so that $-2 \ln p_1 - 2 \ln p_2 = -2 \ln c$, the combined p-value is

$$(4) \qquad \int_{-2 \ln c}^{\infty} \frac{1}{4} y e^{-\frac{y}{2}} dy$$

$$= \frac{1}{4} \left[-2y e^{-\frac{y}{2}} - 4 e^{-\frac{y}{2}} \right]_{-2 \ln c}^{\infty} = -c \ln c + c .$$

In general, if we have n independent experiments yielding the p-values p_1, p_2, ..., p_n, then we consider $-2 \ln c = -2 \sum_{i=1}^{n} \ln p_i$ as a realization of a chi-square random variable with $2n$ d.f., so that the combined p-variable is

$$\int_{-2 \ln c}^{\infty} \frac{1}{(n-1)!\, 2^n} y^{n-1} e^{-\frac{y}{2}} dy .$$

Before continuing, it is interesting to note a particular implication of this general result. Let us ask: what value of p, if it were replicated a large number of times, would lead to a combined p-value of 0.5 by this criterion? For large n, the median of a chi-square random variable with $2n$ d.f. is approximately equal to its mean, or $2n$. We therefore answer the question by setting

$$-2n \ln p = 2n$$

and solving for p. The result is $p = e^{-1} \doteq 0.37$. This suggests that $p < 0.37$ replicated many times will imply that the null hypothesis is false, whereas $p > 0.37$ replicated many times will imply that it is not false. Should we therefore view 0.37 as some critical value, below which the null hypothesis is more likely and above which the alternative hypothesis is more likely? The answer to this question will be discussed in the conclusion.

3. Reformulation of the Problem

Consider again the case of two independent experiments. The resulting p_1 and p_2 can be considered, under the null hypothesis, as a realization of the vector random variable (P_1, P_2) that is uniformly distributed on the unit square, as illustrated in Figure 1. Probability (1) corresponds to the doubly hatched rectangle in the lower left-band corner of this figure, while probability (2) corresponds to the whole hatched part of the figure. We know that probability (1) is too small and probability (2) is too large, so the combined probability we are seeking must lie somewhere between these two extremes. It should correspond to the doubly

hatched area together with some part of the singly hatched areas, as illustrated in Figure 2. The question is: how do we define the curve in Figure 2 such that the combined p-value is equal to the area beneath it?

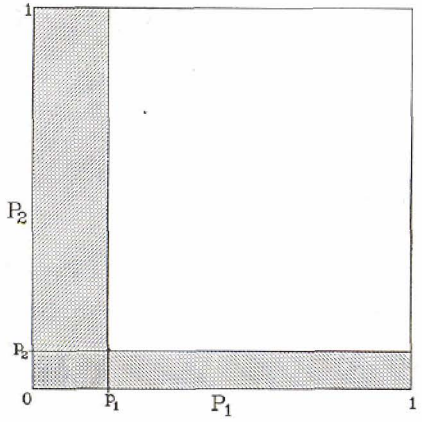

Fig. 1. The pair of p-values p_1, p_2 considered as a point on the unit square. The area of the doubly hatched rectangle is p_1p_2, the whole hatched area is $p_1+p_2-p_1p_2$.

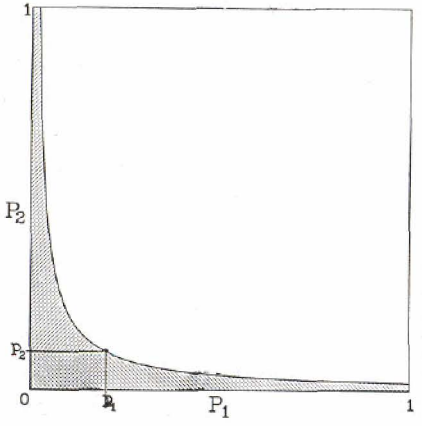

Fig. 2. Hypothetical area that corresponds to the combined p-value. The doubly hatched area is the same as in Figure 1; the singly hatched areas are parts of those in Figure 1. The curve defining the boundary of the hatched areas is everywhere convex decreasing.

Recalling the definition of the p-value as given in the opening sentence of this article, we see that the combined p-value is neither (1) nor (2) but rather

$$P \text{ (results 1 and results 2 or any more extreme result)}.$$

Now if we measure "results 1" by the statistic p_1 and "results 2" by the statistic p_2, we see immediately that we are trying to find

$$P \text{ (obtaining } p_1 \text{ and } p_2 \text{ or any more extreme pair of } p\text{-values)}.$$

In other words, we want all the points (P_1, P_2) that are more extreme than (p_1, p_2) to lie below the curve in Figure 2, and all those less extreme than (p_1, p_2) to be above it. The area below the curve will then be the desired p-value. This leads us to consider the problem of ranking points in the plane, or, more generally, of ranking n-dimensional vectors.

4. Ranking Vectors and Its Relation to Combining p-Values

A practical problem can be briefly described as follows (ELSTON, 1963). A poultry breeder wishes to select female breeding stock for broiler production. On each member of the flock the following measures are available: x_1, the number of eggs laid per year, and x_2, the weight in pounds at ten weeks of age. The problem is to rank all the birds "placing equal emphasis on each trait", so that a predetermined fraction can be selected to be the mothers of the next generation. Clearly it makes no sense to rank the birds on the basis of x_1+x_2, because on the scales of measurements used this would give virtually no weight to x_2. If x_1 were w times as variable as x_2, one could consider ranking the birds on the basis of x_1+x_2w, and then this would define what was meant by "placing equal emphasis on each trait". Instead, we can develop a weight-free index based on the following considerations.

If a bird lays no eggs at all, i.e. $x_1=0$, we want to be certain of not selecting it, however large x_2 might be. Similarly there must be some lower bound, k, for 10-week weight; and if a particular bird has $x_2=k$, we want to be certain of not selecting it, however large x_1 might be. Conversely, if a bird has the largest values possible for both x_1 and x_2, we want to be sure to select it. We therefore want an index function, I, of x_1 and x_2 that satisfies the following three conditions:

(i) I must take on its smallest value at $x_1=0$, whatever the value of x_2;
(ii) I must take on its smallest value at $x_2=k$, whatever the value of x_1;
(iii) I must take on its largest value when both x_1 and x_2 are largest.

In addition, if t is any given threshold value of I above which a bird will be selected, we want the equation $I=t$ to define a curve in the x_1, x_2 plane that is everywhere convex decreasing (analogous to the curve of Figure 2, in the P_1, P_2 plane). This requirement, which intuitively seems reasonable, can be more rigorously justified by reference to what economists call the law of substitution (SAMUELSON, 1955).

The simplest mathematical function that satisfies the above conditions is $I=x_1\,(x_2-k)$. More generally, if we have n traits $x_1, x_2, \ldots, x_n$, the analogous index is $I=\prod_{i=1}^{n}(x_i-k_i)$, where k_i is the lower bound of x_i. This index is weight-free in the sense that, provided the traits are each measured on a scale on which the smallest possible value is 0, the ranking of the individuals is invariant under magnification of any of the scales used. To see this, let $x_i'=x_i-k_i$ and suppose we give weight w_i to the i-th trait. The weighted index is then $I=\prod_{i=1}^{n} w_i x_i' = \left[\prod_{i=1}^{n} w_i\right] \times \times \left[\prod_{i=1}^{n} x_i'\right]$, and it is clear that the first of these two factors is the same for all individuals; it will therefore have no effect on how two individuals with different x's are ranked relative to each other. Let us now use this index to rank the points (P_1, P_2) in Figure 2. The lower bounds of both P_1 and P_2 are 0, and so the index

in this case is $I=(P_1-k_1)\,(P_2-k_2)=P_1P_2$. The curved line in Figure 2 is thus the locus of points given by $P_1P_2=p_1p_2=c$, and the area below the line is

$$P\,(P_1P_2 \leqq c)=\int\limits_{\substack{0\\P_1P_2\leqq c}}^{1}\int\limits_{0}^{1} 1\,dP_1dP_2\,.$$

Reference to Figure 3 shows that this integral can be split into parts and written as

$$\begin{aligned}
&2c-c^2+\int\limits_{c}^{1}\int\limits_{c}^{c/P_2} 1\,dP_1dP_2\\
&=2c-c^2+\int\limits_{c}^{1}\left(\frac{c}{P_2}-c\right)dP_2\\
&=2c-c^2+\,[c\ln P_2-cP_2]_c^1=c-c\ln c\,,
\end{aligned}$$

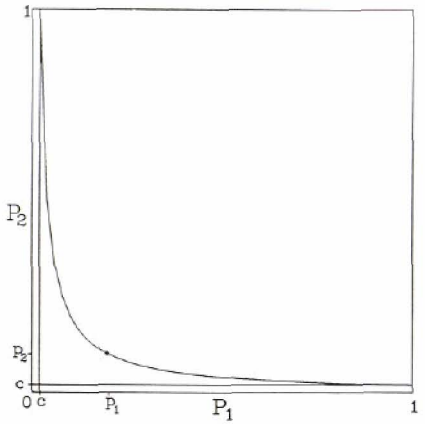

Fig. 3. The area on the unit square that lies beneath the curve $P_1P_2=p_1p_2$ can be split into parts: the square c^2, the two rectangles $c\,(1-c)$, and the rest.

which is the same as (4).

The relation between combining p-values and ranking vectors is seen immediately when we note that the weight-free index can be equivalently expressed on a log scale, i.e. $\ln I=\sum\limits_{i=1}^{n}\ln x_i'$, which in this case is $\sum\limits_{i=1}^{n}\ln P_i$. Thus

$$\begin{aligned}
P\,(P_1P_2\leqq c)&=P\,(\ln P_1+\ln P_2\leqq \ln c)\\
&=P\,(-2\ln P_1-2\ln P_2\geqq -2\ln c)\,,
\end{aligned}$$

which is the same probability as (3). In general, we have combined

$$\begin{aligned}
p&=P && \text{(obtaining } p_1, p_2, \ldots, p_n \text{ or any more extreme}\\
&&& \text{set of } n \text{ } p\text{-values)}\\
&=P\,(P_1P_2\ldots P_n\leqq c), && \text{where the } P_i \text{ are independently uniformly dis-}\\
&&& \text{tributed on } [0,\,1]\,,\\
&=P\left(-2\sum_{i=1}^{n}\ln P_i\geqq -2\ln c\right),
\end{aligned}$$

where $-2\sum_{i=1}^{n} \ln P_i$ is distributed as χ^2 with $2n$ d.f. We thus see that Fisher's method for combining p-values corresponds to this method of ranking vectors.

5. Conclusion

It has been shown that Fisher's method of combining p-values corresponds to a weight-free method of ranking vectors of p-values, i.e., vectors whose elements lie in the unit interval. In particular, the method of ranking has the property that all vectors in which one or more p-values equal zero would be ranked (equally) lowest, and the vector in which all p-values equal unity would be ranked highest. Although there are many other ways in which vectors of p-values could reasonably be ranked, and these would lead to different combined p-values, the particular method discussed here both has intuitive appeal and leads to a mathematically tractable result. If one accepts the rationale underlying this weight-free method of ranking, then the value $p = 0.37$ is a critical one: p-values below 0.37 suggest that the null hypothesis is more likely to be false, whereas p-values above 0.37 suggest that it is more likely to be true. This result is even more intriguing when we note that, in terms of Bahadur relative efficiency, Fisher's method has been shown to be asymptotically optimal among essentially all methods of combining independent tests (LITTELL and FOLKS, 1973).

References

ELSTON, R. C., 1963: A weight-free index for the purpose of ranking or selection with respect to several traits at a time. Biometrics **19**, 85–97.

FISHER, R. A., 1950: *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd, 11th edition, pp. 99–101.

LITTELL, R. C. and J. L. FOLKS, 1973: Asymptotic optimality of Fisher's method of combining independent tests II. J. Amer. Statist. Assoc. **68**, 193–194.

SAMUELSON, P. A., 1955: *Economics: An Introductory Analysis.* New York: McGraw-Hill, 3rd edition, p. 433.

Received Febr. 1990
Revised April 1990

ROBERT C. ELSTON
Department of Biometry and Genetics
LSU Medical Center
1901 Perdido Street
New Orleans, LA 70112