Reproductive BioMedicine Online (2016) 32, 348–349

www.sciencedirect.com
www.rbmonline.com

COMMENTARY

The *P*-value and the problem of multiple testing

The purpose of this editorial is to draw attention to the correct and rigorous interpretation of that widely used statistic the *P*-value. In particular we shall concentrate on the situation when, in a published paper, one or more significant treatment effects or differences are discovered amongst a (possibly) long list of potential effects. In what is usually called the "classical mode of statistical inference", the *P*-value occupies a crucial position.

The rationale in this classical approach to inference is to assume, initially, that the effect of interest is not present, and then to calculate the probability of such an extreme event occurring under those circumstances. Thus if an investigation involves estimating a certain treatment difference, or "effect", the *P*-value represents the probability of such an extreme difference as that observed occurring by chance, that is in the absence of any genuine and systematic effect, the deviations observed being simply due to random variation. The *P*-value is often also called the Type 1 error, since it represents the probability of wrongly concluding that the effect of interest is present. The absence of any systematic effect is usually referred to as the "Null Hypothesis".

A *P*-value of 5% (0.05) is often adopted as a critical value in making the inferences. Thus if the *P*-value associated with an effect is calculated as 0.05 the inference is that such a large effect would only occur by chance with a probability of 0.05, or once in 20. Generally this would be taken as evidence of the presence of the effect in question. It is important to realise however that under those circumstances there is a 1 in 20 chance of being wrong in this conclusion.

Unfortunately the *P*-value is a statistic that is widely misused and mis-interpreted in a great deal of experimental work. Very often a very large number of tests are cited in a publication but very little attention is paid to the probability of spurious significant results. After all, if 50 independent tests are carried out in a single investigation, one would expect more than 2 significant results at the 5% level simply by chance and random variation. Clearly for large investigations some allowances are needed to deal with this troublesome feature in order to guarantee reliable inferences. A strict adherence to a 5% or perhaps an even smaller significance probability will reduce the frequency of wasteful attempts at interpreting spurious effects. This property would also of course reduce the need to explain away the absence of significant results in other comparable research efforts. It must be acknowledged however that there is at present no universally accepted method of dealing with multiple testing.

One of the earliest attempts to deal with this problem was by Bonferroni (1936), who recommended that the nominal *P*-value, often set at 0.05, should be scaled down by the number of tests envisaged. Thus if there are 10 independent tests in an investigation, the nominal significant *P* value for each test should be reduced from 0.05 to 0.005, (i.e. 0.05/10) in order, to guarantee that the probability of one or more "spurious" significant effects is 0.05. This statistic is usually labelled the FWER (Family Wise Error Rate).

The rationale for this very simple solution to the problem proceeds as follows. If n is the number of independent tests envisaged in an investigation, and we wish to retain an "experiment-wise" probability of a Type 1 error of say 0.05, then the *P*-value applied to each individual test α_t is the solution of the equation:

$$1-(1-\alpha_t)^n = 0.05$$

Thus,

$$\alpha_t = 1-(1-0.05)^{1/n}$$

More generally, if α_e is the chosen "experiment-wise" Type 1 Error and α_t is the corresponding individual test *P*-value then

$$\alpha_t = 1-(1-\alpha_e)^{1/n} \approx \alpha_e/n$$

The exact evaluation of α_t would be rather troublesome, but the approximation listed will be satisfactory for most purposes.

In fact the Bonferroni approach is rather conservative, and numerous alternative procedures have been suggested over the years, important early contributions having been made

http://dx.doi.org/10.1016/j.rbmo.2016.02.008

Table 1 The nominal *P*-values for the 15 hypotheses in increasing numerical order.

n(H)	*i*	P_i	Bon.	Ben.
15	1	0.0001	0.0033	0.0033
6	2	0.0004	0.0033	0.0067
8	3	0.0019	0.0033	0.0100
5	4	0.0095	0.0033	0.0133
2	5	0.0201	0.0033	0.0167
13	6	0.0278	0.0033	0.0200
7	7	0.0298	0.0033	0.0233
4	8	0.0344	0.0033	0.0267
10	9	0.0459	0.0033	0.0300
12	10	0.3240	0.0033	0.0333
9	11	0.4262	0.0033	0.0367
3	12	0.5719	0.0033	0.0400
13	13	0.6528	0.0033	0.0433
14	14	0.7590	0.0033	0.0467
1	15	1.0000	0.0033	0.0500

n(H) = the identification number of the hypothesis; P_i = the nominal *P*-value; Bon. = the adjusted *P*-values according to the Bonferroni (1936) approach; Ben. = the adjusted *P*-values according to the approach recommended by Benjamini and Hochberg (1995).

by Scheffe (1953), Duncan (1955), and Tukey (1949). A more recent, and influential contribution to the topic is by Benjamini and Hochberg (1995), who advocate a sequential strategy that involves comparing the statistic P_i with $0.05i/n$, where n = the number of tests (or hypotheses), starting with $i = n$ and proceeding backwards until $P_i < 0.05i/n$.

We may now illustrate the ideas outlined above by using a list of *P*-values, which may be assumed to arise from an investigation involving 15 independent tests, listed in **Table 1** in ascending order. In fact this was the dataset used by Benjamini and Hochberg to describe their procedure.

A straightforward adoption of the conservative Bonferroni approach would lead us to reject all hypotheses when the observed *P*-value was less than 0.0033 (=0.05/15). Thus hypotheses 15, 6, and 8 would be rejected.

The sequential strategy advocated by Benjamini and Hochberg involves comparing the statistic P_i with $0.05i/15$, starting with $i = 15$, and proceeding, backwards, until $P_i < 0.05i/15$. It will be seen that the first occasion when the criterion is satisfied is when $P_i = 0.0095$. We conclude therefore that the first four hypotheses, that is hypotheses 15, 6, 8, and 5 are rejected, and that the corresponding effects are genuine.

Consider now a second numerical example, derived from the data displayed in Adler et al. (2015; table 2). All we require to illustrate these ideas are the 17 *P*-values, in ascending order, which have been regenerated from the information contained in the paper, and are shown below.

0.007, 0.020, 0.067, 0.095, 0.139, 0.168, 0.208, 0.284, 0.294, 0.384, 0.412, 0.418, 0.465, 0.653, 0.704, 0.881, 0.936

We note immediately that the smallest *P*-value (0.007) among the list of 17 values does not constitute statistical significance after the adoption of the Bonferroni or the Benjamini/Hochberg strategy. We would conclude therefore that evidence of systematic differences among the corresponding list of effects is not strong.

The paper by Benjamini and Hochberg demonstrates that their method has highly desirable properties regarding the False Discovery rate (FDR) and the Power of the tests.

In conclusion, I fear that the integrity of the significance test has in recent years been eroded by poor application. The eagerness of research workers to derive a positive result from their investigations has led to a rather lax attitude to statistical rigour. For example, this has led some workers to adopt a significant *P*-value of 0.10, rather than the more familiar 0.05. Also the rash interpretation of a small number of significant results abstracted from a large number of tests, without adopting any of the safeguards of multiple testing may well lead to unreliable inferences.

In view of these findings, it is clear that some form of adjustment for multiple testing is highly desirable. Further the recent procedure by Benjamini and Hochberg would appear to have highly desirable properties, and should be applied if at all possible. Otherwise the early Bonferroni approach, although being very conservative, has the appeal of being a very cautious procedure, and remarkably easy to apply.

What needs to be avoided of course is a universal application of the unmodified (perhaps 5%) test for all effects to be investigated. I would urge readers to pay attention to the recommendations in this editorial, as it will surely lead to more reliable conclusions in an important research area. Finally we should note that the Benjamini and Hochberg procedure has now been adopted as a requirement in many research areas. See for example Thissen et al. (2002).

References

Adler, A., Lee, H.-L., McCulloh, D.H., Ampeloquio, E., Clarke-Williams, M., Wertz, B.H., Grifo, J., 2015. Blastocyst culture selects for euploid embryos: comparison of blastomere and trophectoderm biopsies. Reprod. Biomed. Online 28, 485–491.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate; a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. (B) 57, 289–300.

Bonferroni, C.E., 1936. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni R Istit. Super. Sci. Econ. Commer. Fir. 8, 3–62.

Duncan, D.B., 1955. Multiple range and multiple F tests. Biometrics 11, 142.

Scheffe, H., 1953. A method for judging all contrasts in the analysis of variance. Biometrika 40, 87–110.

Thissen, D., Steinberg, L., Kuang, D., 2002. Quick and easy implementation of the Benjamini–Hochberg procedure for controlling the false positive rate in multiple comparisons. J. Educ. Behav. Stat. 27, 77–83.

Tukey, J., 1949. Comparing individual means in the analysis of variance. Biometrics 5, 99–114.

Eurof Walters
E-mail address: office@rbmonline.com