Tidsskriftet Tidsskrift for Den norske legeforening

Adjustment of p-values for multiple hypotheses

MEDICINE AND NUMBERS

STIAN LYDERSEN

stian.lydersen@ntnu.no
Stian Lydersen, PhD, professor of medical statistics at the Regional Centre for Child and Youth Mental Health and Child Welfare, Norwegian University of Science and Technology.

The author has completed the ICMJE form and declares no conflicts of interest

It is quite common to investigate multiple hypotheses in a single study, which increases the probability of Type I errors. This can be dealt with in various ways.

A researcher may have various reasons for testing multiple hypotheses in the same study, for example to investigate the effect on several outcome variables, compare more than two groups or undertake separate analyses for sub-groups.

Different adjustment methods

Consider a study where six hypothesis tests are performed. If all tests are made at a significance level of 5 %, each of them will have a 5 % probability of making a Type I error, that is, erroneously rejecting the null hypothesis (1). The probability of a Type I error in at least one of the hypothesis tests, also referred to as the *family-wise error rate* (FWER) (2), will then be substantially higher than 5 %, and at worst almost 30 %. Sometimes it is desirable to control this error rate to prevent it from exceeding a pre-defined threshold, for example a significance level of 5 %.

The simplest method is a so-called Bonferroni correction. This means multiplying the *p*-values by the number of hypotheses, in this case six, before comparing with the significance level. However, the Bonferroni correction is very conservative, which means that the statistical power, and thereby the probability of determining true hypotheses, will be greatly reduced. By using the Šidák correction, only a marginal improvement is achieved. Alternative methods, in order of increasing statistical power, are Holm's *step-down* correction, Hochberg's *step-up* correction and the Hommel correction (3). These methods are valid under general assumptions, and can be generally recommended.

In some situations, a large number of hypotheses are tested. For example, genetics studies may involve several hundred thousand hypotheses. In practice it will thus be impossible to control for the family-wise error rate. Instead, we have to content ourselves with controlling for the *false discovery rate* (FDR) (2). We allow for a certain proportion, normally 5 %, of the hypotheses that we mark out as true in one and the same study, to be false positives. When controlling for the family-wise error rate, on the other hand, we would not 'accept' even a single false-positive finding. The most common method for controlling for the false discovery rate is called the Benjamini-Hochberg correction (4). Controlling for the false discovery rate can also be relevant in trials, for example with as few as 8 to 16 hypothesis tests, although its benefits are greater for testing a large number of hypotheses (4).

Let us look at an example where we have six unadjusted *p*-values listed by size (Table 1). We can see how methods that make for higher statistical power typically give lower *p*-values. We see that the lowest adjusted *p*-value is the same as that obtained by the Bonferroni correction, irrespective of method. The final column with *p*-values adjusted with the Benjamini-Hochberg correction controls only for the false discovery rate. With only six hypothesis tests, another method would be used in practice.

Table 1

An example with six p-values, unadjusted and adjusted by different methods of correction.

Unadjusted *p*-value	Bonferroni	Šidák	Holm's *step-down*	Hochberg's *step-up*	Hommel	Benjamini-Hochberg
0.0003	0.0018	0.0018	0.0018	0.0018	0.0018	0.0018
0.009	0.054	0.053	0.045	0.042	0.028	0.021
0.013	0.078	0.076	0.052	0.042	0.039	0.021
0.014	0.084	0.081	0.052	0.042	0.042	0.021
0.04	0.24	0.22	0.08	0.08	0.06	0.048
0.06	0.36	0.31	0.08	0.08	0.06	0.06

Always adjust?

Do we always need to adjust for multiple hypotheses? This is a controversial question. The epidemiologist Kenneth Rothman argues against adjusting for multiplicity in some contexts (5). To put this into relief: imagine a researcher who studies the effect of a treatment on three outcome variables. Does he need to adjust for multiplicity if he splits the results into three different publications with only one hypothesis in each? Or should he perhaps adjust for all the hypotheses that he has tested during his career?

There are some alternatives to adjustment. In a study with several outcome variables it is normal to specify which is the primary one. Hypothesis tests are performed without adjusting, but in any findings 'less weight' is placed on secondary outcome variables. In other situations it may be relevant to choose a pragmatic solution, such as setting the significance level at 1 %, rather than 5 %. This will give some protection against false-positives, but usually without reducing statistical power as much as a formal adjustment would have done.

There is no general consensus regarding when, and if so, how, we should adjust for multiple hypotheses. However, the choice of procedure must be specified in advance in the protocol or analysis plan in order to avoid 'fishing' for significant findings.

LITERATURE

1. Lydersen S. Type I-feil og type II-feil. Tidsskr Nor Legeforen 2021; 141. doi: 10.4045/tidsskr.21.0013. [PubMed][CrossRef]

2. Lydersen S. Justering av p-verdier på norsk. Tidsskr Nor Legeforen 2021; 141. doi: 10.4045/tidsskr.21.0360. [CrossRef]

3. Dmitrienko A, D'Agostino R. Traditional multiplicity adjustment methods in clinical trials. Stat Med 2013; 32: 5172–218. [PubMed][CrossRef]

4. Benjamini Y, Hochberg Y. Controlling the false discovery rate – A practical and powerful approach to multiple testing. J R Stat Soc B 1995; 57: 289–300. [CrossRef]

5. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990; 1: 43–6. [PubMed][CrossRef]

Publisert: 28 September 2021. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.21.0357