International Journal of Business Analytics and Intelligence
7 (1) 2019, 01-02
http://publishingindia.com/ijbai/

Submitted: 12 January, 2019
Accepted: 12 February, 2019

Analytically Yours:

One Data, Many Tests

Arnab Kumar Laha*

In many areas of research in management, social science, medical science, genomics, business studies and psychology it is a common practice for researchers to test their theoretical understanding of a phenomenon through formulation of appropriate hypotheses which can be proved or disproved on the basis of data. These researchers argue that if the data provides support to the formulated hypotheses then it can be concluded that the theory based on which these hypotheses were derived is also empirically validated. Typically a piece of research may depend on testing more than one hypothesis and the empirical validation of the theory requires all these hypotheses being supported.

The subject of statistical test of a hypothesis has a long history going back to the work of John Arbuthnot more than three centuries ago. Formally Sir R. A. Fisher introduced what is now termed as the *P-value approach* to hypothesis testing in 1925. The main idea of this approach is as follows: Let H be a hypothesis that the researcher wants to test. Towards this goal she decides on a test statistic T and obtains its probability distribution assuming the hypothesis H to be true. Then she collects the required data and computes the value of T for this dataset. Suppose the obtained value of T is τ. The P-value of the test - which is the probability of obtaining a value of T more extreme than τ when H is true - is next obtained using the probability distribution of T computed earlier. If the P-value turns out to be "small" then it is concluded that the hypothesis H is false and is rejected.

J. Neyman and E.S. Pearson gave a different formulation of testing a hypothesis which is sometimes called the *fixed-α approach*. In this formulation the researcher has to specify two hypotheses H_0 and H_1 which are called the *null hypothesis* and *alternative hypothesis* respectively. H_0 here plays the same role as H in the Fisher's formulation with the additional understanding that if H_0 is rejected then H_1 is assumed to hold true. The researcher now decides on a suitable test statistic T and obtains its probability distribution assuming H_0 is true. She also decides on a value α and determines a "rejection region" (R) such that $P(T \in R|H_0 \text{ is true}) = \alpha$. She then collects the data D and computes the value of T say τ. If $t \in R$ then H_0 is rejected. α is called the level of significance of the test. Also, note that $P(\text{Reject } H_0|H_0 \text{ is true}) = \alpha$.

Now suppose there are m hypotheses H_i, $t = 1, \ldots, m$ that are to be tested on the same dataset. Of these m hypotheses suppose m_0 are actually true and the remaining $m_1 - m - m_0$ are actually false. Without loss of generality suppose that the hypotheses $H_1, \ldots, H_{m_0}$ are actually true. Each of these hypotheses are tested separately and the p-values obtained are $q_i < \alpha$. In the fixed-α approach any hypothesis for which $q_i < \alpha$ is rejected. The problem with this approach in the multiple testing context is that chance of rejecting at least one of the m_0 true hypotheses is larger than α. Hence we need to use a different strategy to ensure that the overall probability of rejecting a true hypothesis is not larger than α.

The following Table 1 summarizes the situation after the m-tests are carried out. Note that only N, D and m are observed.

Table 1

	H_0 retained	*H_0 rejected*	
H_0 True	True Negative (TN)	False Discovery (FD)	m_0
H_0 False	False Negative (FN)	True Discovery (TD)	m_1
	N = TN+FN	D = FD+TD	m

The simplest strategy is to carry out the individual tests of hypothesis at a much lesser level of significance $\frac{\alpha}{m}$. This is known as the Bonferroni method. If the overall probability of rejecting a true hypothesis is to be restricted

* Indian Institute of Management Ahmedabad, Gujarat, India. Email: arnab@iima.ac.in

to 5% i.e. α = 0.05 and the study involves testing 10 hypotheses then each hypothesis is tested at 0.5% level of significance. In other words in Bonferroni method only those hypotheses for which q_i < 0.005 are rejected. To understand why this approach works one needs to recall the Boole's inequality which states that for any set of events $A_1, \ldots, A_k$, $P(\bigcup_{i=1}^{m_0} E_i) \leq \sum_{i=1}^{k} P(A_k)$. Let E_i be the event that $q_i \leq \frac{\alpha}{m}$. Then $P(FD > 0) =$ *P(a least one true hypothesis is rejected*)

$$= P\bigcup_{i=1}^{m_0} E_i \leq \sum_{i=1}^{m_0} P(E_i) = m_0 \frac{\alpha}{m} \leq \alpha$$

The Bonferroni procedure is extremely simple but it suffers from a major disadvantage that it has a very high rate of false negatives.

We examine this further using a simulated example. We generate 50 datasets each of size 10 from a normal distribution with mean(μ) = 0 and standard deviation(σ) = 1 and another 50 datasets each of size 10 from a normal distribution with μ=1 and σ=1. We want to test H_0 : $\mu = 0$ against H_1 : $\mu \neq 0$ by applying the t-test to each of the datasets maintaining P(FD > 0) = α = 0.05. The Bonferroni method requires us to carry out the individual tests at $\frac{0.05}{100}$ = 0.0005 level of significance. When this method is carried out we get the following:

Table 2

	H_0 retained	*H_0 rejected*	
H_0 True	49	1	50
H_0 False	47	3	50
	96	4	100

From Table 2 we find that only 3 of the 50 false hypotheses are rejected by the Bonferroni procedure substantiating the conservativeness of the procedure.

Benjamini and Hochberg (1995) introduced the idea of controlling the expected proportion of falsely rejected hypotheses which they call the False Discovery Rate (FDR). They developed a procedure now widely known as the Benjamini-Hochberg (BH) procedure the controls the FDR = $E\left(\frac{FD}{D}\right)$. The BH-procedure ensures that FDR ≤ α. The mechanics of the BH-procedure is given below:

- We first sort the p-values of the m individual tests in ascending order.
- Next, we assign ranks corresponding to the p-values with the smallest p-value having a rank of 1, the second smallest p-value having a rank of 2 etc.
- Calculate the Benjamini-Hochberg (BH) critical value corresponding to each individual p-value, using the formula $\frac{1}{m}Q$, where i = the individual p-value's rank, m = total number of tests, Q = the false discovery rate.
- The original p-values are now compared with the BH critical values obtained in Step 3. Let p^* be the largest p value that is smaller than the BH critical value.
- All hypotheses having p-values less than p^* are rejected.

Let us now examine what happens when the BH-method with FDR control of 5% is applied to the p-values obtained in the simulated example discussed above. Table 3 below provides a summary. We find that there has been a substantial increase in the number false hypotheses that are rejected by BH method compared to that rejected by the Bonferroni method while the number of true hypothesis that are rejected by both the methods remain the same. Thus the BH-method successfully overcomes a critical shortcoming of the Bonferroni method.

Table 3

	H_0 retained	*H_0 rejected*	
H_0 True	49	1	50
H_0 False	15	35	50
	64	36	100

Reference

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.