Statistical Science
2011, Vol. 26, No. 4, 598–600
DOI: 10.1214/11-STS356A
Main article DOI: 10.1214/11-STS356

Discussion of "Multiple Testing for Exploratory Research" by J. J. Goeman and A. Solari

Ruth Heller

Abstract. Goeman and Solari [*Statist. Sci.* **26** (2011) 584–597] have addressed the interesting topic of multiple testing for exploratory research, and provided us with nice suggestions for exploratory analysis. They defined properties that an inferential procedure should have for exploratory analysis: the procedure should be *mild*, *flexible* and *post hoc*. Their inferential procedure gives a lower bound on the number of false hypotheses among the selected hypotheses, and moreover whenever possible identifies elementary hypotheses that are false. The need to estimate a lower bound on the number of false hypotheses arises in various applications, and the partial conjunction approach was developed for this purpose in *Biometrics* **64** (2008) 1215–1222 (see also *Philos. Trans. R. Soc. Lond. Ser. A* **367** (2009) 4255–4271 for more details). For example, in a combined analysis of several studies that examine the same problem, it is of interest to give a lower bound on the number of studies in which the finding was reproduced. I will first address the relation between the method of Goeman and Solari and the partial conjunction approach. Then I will discuss possible extensions and address the issue of exploration in more general settings, where the local test may not be defined in advance or where the candidate hypotheses may not be known to begin with.

1. RELATION TO THE TESTING OF PARTIAL CONJUNCTION HYPOTHESES

Let $H_1, \ldots, H_n$ be the elementary hypotheses. The idea of giving a lower bound on the number of false elementary hypotheses (or equivalently an upper bound on the number of true elementary hypotheses) appears in [1], and is closely related to the tests of partial conjunction hypotheses. The partial conjunction null hypothesis $H^{u/n}$ in [1] asks whether fewer than u of the elementary hypotheses are false, and the alternative hypothesis is that at least u of the elementary hypotheses are false. Testing whether $H^{u/n}$ is false at a significance level α in order (i.e., for $u = 1, 2, \ldots$) results in a $1-\alpha$ confidence lower bound on the number of false elementary hypotheses:

Ruth Heller is Senior Lecturer, Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel (e-mail: ruheller@post.tau.ac.il).

THEOREM 1.1. *Let $p^{u/n}$ be a partial conjunction p-value for testing $H^{u/n}$. Let $u_{\max} = \max\{u : p^{i/n} \le \alpha \ \forall i = 1, \ldots, u\}$. Then with $1-\alpha$ confidence, the true number of false hypotheses is in $[u_{\max}, n]$.*

PROOF. Let k be the true number of false elementary hypotheses. If $k = n$, that is, all elementary hypotheses are false, there is nothing to prove. If $k < n$,

$$\begin{aligned}\Pr(k \ge u_{\max}) &= 1 - \Pr(k < u_{\max}) \\ &= 1 - \Pr\bigl(P^{(k+1)/n} \le \alpha\bigr) \ge 1 - \alpha. \qquad \square\end{aligned}$$

The lower bound $u_{\max}$ above is identical to the lower bound of Goeman and Solari (denoted by $f_\alpha\{1, \ldots, n\}$ in their paper), when the full set of elementary hypotheses is considered. Moreover, the shortcuts suggested by Goeman and Solari are equivalent to the tests of partial conjunction hypotheses suggested in [1], that do not require examination of all $\binom{n}{n-u+1}$ intersection hypotheses for the test of $H^{u/n}$, but rather require only testing the subset of $n-u+1$ intersection hypotheses

that correspond to the $n-u+1$ least significant elementary hypotheses p-values. Specifics follow.

Reference [1] suggested methods for combining the p-values for testing $H^{u/n}$ that are based on *sufficient combining functions*.

DEFINITION 1.1. $f(U_1,\ldots,U_m)$ is a sufficient combining function from $\Re^m \to \Re$ if it has the following properties:

1. If $U_i' \geq U_i$, then $f(U_1,\ldots,U_{i-1},U_i',U_{i+1},\ldots,U_m) \geq f(U_1,\ldots,U_{i-1},U_i,U_{i+1},\ldots,U_m)$, that is, f is an increasing function of its components.
2. If U_i is uniformly distributed or stochastically larger than the uniform, that is, $U_i \underset{\text{st}}{\geq} U(0,1)\ \forall i = 1,\ldots,n$, then $f(U_1,\ldots,U_m) \underset{\text{st}}{\geq} U(0,1)$.

Let $p_{(1)} \leq \cdots \leq p_{(n)}$ be the sorted p-values. The following lemma gives the guiding principle for the p-values suggested in [1] for testing the partial conjunction hypothesis:

LEMMA 1.1. *Let $f(U_1,\ldots,U_{n-u+1})$ be a sufficient combining function from $\Re^{n-u+1} \to \Re$. Let $p^{u/n}$ be the result of combining the largest $n-u+1$ p-values using the function f, that is, $p^{u/n} = f(p_{(u)},\ldots,p_{(n)})$. Then $\Pr(P^{u/n} \leq \alpha) \leq \alpha$ if $H^{u/n}$ is true.*

For example, if the p-values are independent the p-value motivated by the Fisher method for testing $H^{u/n}$ is

$$p^{u/n} = \Pr\left(\chi^2_{2(n-u+1)} \geq -2\sum_{i=u}^{n} \log p_{(i)}\right).$$

Finding $u_{\max}$ using the partial conjunction test p-values based on Fisher's method will give the same result as the procedure in Section 4.1 of Goeman and Solari, when the full set of elementary hypotheses is considered.

Similarly, if a set $R \subset \{1,\ldots,n\}$ is selected a priori, then the lower bound on the number of false hypotheses may be found by testing in order the partial conjunction hypotheses $p^{u/|R|}, u = 1,2,\ldots,$ where $|R|$ is the cardinality of R. If the set R is selected post hoc, then the lower $1-\alpha$ confidence bound on the number of false hypotheses may be lower than the bound resulting from the above procedure because of the selection effect, and the procedures suggested by Goeman and Solari can be used to adjust for the selection effect.

2. MULTIPLE FAMILIES OF HYPOTHESES IN EXPLORATORY RESEARCH

In [1], the partial conjunction approach was used to estimate the lower bound on the number of false hypotheses when a large number of such lower bounds need to be estimated simultaneously. In multiple testing for exploratory research, a similar problem may arise. Consider, for example, a large genomics study, where the signal in many genes (or SNPs) are measured simultaneously. In order to select genes (or SNPs) for follow-up, the researcher may want to select a subset of promising genes from prespecified regions in the genome. In such a problem, in each region a subset of promising genes (or SNPs) may be selected by exploration of that region.

When exploring multiple families of hypotheses, in order to limit the total number of false leads, the decision about the subset of hypotheses selected for follow-up in each family may be affected by the estimated lower bounds on the number of false null hypotheses in the subsets selected in other families of hypotheses. Moreover, the researcher may be interested in a lower bound on the number of false leads at the level of families rather than at the level of elementary null hypotheses. These are natural extensions to the problem addressed by Goeman and Solari, where multiple testing may be applied to multiple families of hypotheses in an exploratory manner.

3. THE CHOICE OF THE LOCAL TEST

The approach of Goeman and Solari assumes that the test of each intersection hypothesis is known in advance. However, it may be difficult to decide which local test is best without first looking at the data.

In some applications, we may not always have a good statistic in mind for evaluating an elementary null hypothesis. We may need to explore the data in order to decide on a good test statistic for testing the null hypothesis. However, when testing the elementary hypothesis on the data explored to decide on the test, the test is no longer a valid test in the sense that there is no guarantee it preserves the level of the test.

Moreover, when we have several elementary hypotheses of interest and we want to test their intersection hypothesis, how should the test statistic be chosen? Different tests will have power against different alternatives. Even if we limit ourselves to tests that are based on combining functions of the elementary hypotheses p-values, different functions are better capable of detecting different patterns of evidence against

the intersection null hypothesis, and the differences among them can be large (see, e.g., [7] and [4]). Because no single combining function can be best under all circumstances, in exploratory analysis the researcher may choose a combining function by exploring different combining methods. The chosen method may then be used on data from follow-up studies. However, for testing the intersection hypotheses on the data explored, the test is no longer a valid test.

Therefore, if the data are explored to select which local test to use, the confidence sets may no longer have the correct level and may be misleading. Nevertheless, the use of multiple testing for selecting hypotheses for follow-up is still valuable as a tool, even though it is not possible to quantify the number of false leads in the selected subset of hypotheses for follow-up.

4. THE PRACTICE OF EXPLORATORY RESEARCH

Even when multiple comparisons issues are addressed, still studies are too often not reproducible (see [6]) and scientists follow too many false leads. This may be because together with advances in multiple comparisons over the years, there have been many advances in how data can be explored. The multiple comparisons correction is possibly done only on a subset of hypotheses without intention. From sophisticated (and even simple) graphical displays, a hypothesis may be generated. But how can one quantify then how many potential hypotheses have actually been tested before selecting the particularly interesting one based on the picture? If the user cannot quantify how many hypotheses may be looked at in the exploratory stage, how should the data be analyzed to select promising hypotheses to follow up on while still quantifying the error in terms of a lower bound on the number of false null hypotheses?

One possibility is to define the hypotheses on part of the data by creative exploratory analysis and then apply the multiple testing procedure on the rest of the data (see [3]). The problem is that by testing only part of the data we lose power. Therefore, a modest change in current practice may be the following: to set aside only the amount of data that the investigator is willing to spare for the purpose of generation of hypotheses and in order to decide what local test to use for each hypothesis. So, for example, from a study of 500 subjects the investigator may be willing to set aside 100 subjects, and from a sample size of 100 perhaps only 15 subjects may be set aside for hypothesis generations. Once the hypotheses and tests of hypotheses have been decided upon, the procedure of Goeman and Solari may be applied. This process is *mild, flexible* and *post hoc* without losing all ability to quantify the confidence on the estimated number of false positives among the selected hypotheses.

ACKNOWLEDGMENTS

Supported by the Israel Science Foundation (ISF) Grant no. 2012896.

REFERENCES

[1] BENJAMINI, Y. and HELLER, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64** 1215–1222. MR2522270

[2] BENJAMINI, Y., HELLER, R. and YEKUTIELI, D. (2009). Selective inference in complex research. *Philos. Trans. R. Soc. Lond. Ser. A* **367** 4255–4271. MR2546387

[3] COX, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444. MR0378189

[4] DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81. MR1157714

[5] GOEMAN, J. and SOLARI, A. (2011). Multiple testing for exploratory reserach. *Statist. Sci.* **26** 584–597.

[6] IOANNIDIS, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* **2** e124.

[7] LOUGHIN, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Comput. Statist. Data Anal.* **47** 467–485. MR2086483