THE DISTRIBUTION OF THE NUMBERS OF SIGNIFICANT RESULTS IN STUDIES WITH MANY HYPOTHESES

Gerhard Hommel[1] and Andreas Faldum[1]

[1] *Institute of Medical Biostatistics, Epidemiology and Informatics*

When in a study with many hypotheses (e.g. a microarray study) a lot of significant results has been obtained, it may often occur that none or only few of these significances will be confirmed when the study is replicated. Publication bias can be one reason for this effect. However, it is also worth to investigate the distribution of the random variables R_1 = number of all rejections and of R_2 = number of true rejections. When the dispersion of R_1 or R_2 is large, this is an additional argument that a replication can lead to completely different results. Though R_2 is more relevant, we focus in particular on the observable random variable R_1.

We considered the dispersion of R_1 and R_2 for three common multiple test procedures: Bonferroni, Bonferroni-Holm (Holm, 1979), and the explorative Simes procedure by Benjamini and Hochberg (1995). In a simulation study we assumed that the test statistics were multivariate normal with common correlation coefficient ρ. Different values for n = number of all hypotheses, n_0 = number of true null hypotheses and ρ were chosen. As a measure of dispersion the standard deviation $\sigma(R_1)$ ($\sigma(R_2)$) was computed.

We found that Holm's procedure had a slightly larger standard deviation than the Bonferroni procedure. In contrast, the standard deviation of the explorative Simes procedure was for realistic situations substantially larger than that of the other two procedures. Since superiority in power could be the reason for this effect, the three procedures were also compared when the same expectations $E(R_1)$ ($E(R_2)$) were assumed. Nevertheless, the tendency of the results remained the same as without this standardization.

One concludes that the reproducibility of significant results is much worse for the explorative Simes procedure. Stepwise procedures have a larger dispersion than non-stepwise procedures. It is discussed whether these properties can be accepted since procedures with a large $\sigma(R_1)$ ($\sigma(R_2)$) have usually much more power. It could also be interpreted as an "injustice" when two researchers performing the same experiment obtain completely different numbers of positive findings.