A New Method for Detecting Significant p-values with Applications to Genetic Data

Una nuevo método para la detección de valores p significativos y su aplicación a datos genéticos

Jorge Iván Vélez[1,2,3,a], Juan Carlos Correa[3,4,b], Mauricio Arcos-Burgos[1,2,c]

[1]Genomics and Predicitive Medicine Group, Genome Biology Department, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia

[2]Group of Neurosciences, University of Antioquia, Medellín, Colombia

[3]Research Group in Statistics, National University of Colombia, Medellín, Colombia

[4]Department of Statistics, National University of Colombia, Medellín, Colombia

Abstract

A new method for detecting significant p-values is described in this paper. This method, based on the distribution of the m-th order statistic of a $U(0,1)$ distribution, is shown to be suitable in applications where $m \to \infty$ independent hypothesis are tested and it is of interest for a fixed type I error probability to determine those being significant while controlling the false positives. Equivalencies and comparisons between our method and others methods based-on p-values are also established, and a graphical representation of the distribution of the test statistic is depicted for different values of m. Finally, our proposal is illustrated with two microarray data sets.

Key words: Extreme values theory, p-value, Type I error probability, Multiple testing, Genetic data.

Resumen

Se describe una nuevo método para la detección de valores p significativos. Este método, basado en el m-ésimo estadístico de orden de la distribución $U(0,1)$, es adecuado en casos en los que se realizan $m \to \infty$ pruebas de hipótesis independientes y es de interés determinar aquellas que son significativas, controlando los falsos positivos, para una probabilidad de error tipo I predeterminada. Adicionalmente, se realiza una comparación con algunas

[a]Ph.D Scholar. E-mail: jorge.velez@anu.edu.au

[b]Associate professor. E-mail: jccorrea@unal.edu.co

[c]Associate professor. E-mail: mauricio.arcos-burgos@anu.edu.au

pruebas clásicas y se grafica la distribución del estadístico de prueba para diferentes valores de m. Finalmente se ilustra el uso de la metodología con dos conjuntos de datos provenientes de estudios con microarreglos.

Palabras clave: teoría de valores extremos, valor-p, probabilidad de error tipo I, comparaciones múltiples, datos genéticos.

1. Introduction

Genome-wide association studies (GWAS) are aimed at identifying genetics variants associated with a trait (Manolio 2010). For this, hundred of thousands participants with and without a particular disease (or trait) are required, and hundred of thousand of genetic variants, i.e., single nucleotide polymorphisms (SNPs), are read using SNPs arrays. Associated variants are further determined after performing (not necessarily) independent statistical tests comparing either the allele frequency or the distribution of the genotypes of these SNPs between cases and controls. Further, the correspondent p-value for each SNP is used to determine whether it is associated with the disease.

As a total of $m \to \infty$ independent SNPs are being tested in a typical GWAS, the problem of determining which variants are associated with the specific trait can be reduced to a multiple testing problem (for a review see Shaffer 1995) and so the family-wise error rate (FWER), i.e., the probability that one or more of the significance tests results in a type I error, must be controlled at level α. For such purpose, several methods can be applied (Bonferroni 1935, Shaffer 1995, Benjamini & Hochberg 1995, Nyholt 2004, Liu et al. 2010). In general terms, these methods use the p-values for each SNP and compare with a (adaptative) threshold, such that the SNPs associated with the trait are those for which the p-value is grater (or lower) than that threshold.

Here we describe a new method to detect p-values while controlling the FWER at level α. This method is heavily based on extreme values theory and considers the distribution of m-th order statistic of a $U(0,1)$. We derive the test statistic, show its equivalency with Bonferroni's method, and provide asymptotic results for its limiting distribution. In addition, we report preliminary results of a simulation study in which, under the null hypothesis, i.e., $p \sim U(0,1)$, the limiting distribution and the simulated values are depicted for different values of m. Finally, we apply our method to two well-known microarray data sets (Golub et al. 1999, Mootha et al. 2003).

2. Describing the Method

2.1. Background

Suppose that $m \to \infty$ independent hypotheses of the form

$$H_{0,i} : \theta_i \in \Theta \quad \text{vs.} \quad H_{1,i} : \theta_i \notin \Theta \qquad i = 1, 2, \ldots, m \tag{1}$$

are tested, with θ_i some parameter of interest and Θ the parameter space. Let $\alpha \in (0,1)$ be the type I error probability at which the ith hypothesis is tested and

$$P_i = 1 - G(T_i) \qquad i = 1, 2, \ldots, m \tag{2}$$

be its P-value. In (2), T_i is the test statistic for the ith hypothesis and G its cumulative distribution function (cdf). Under H_0, $P_1, P_2, \ldots, P_m$ is a random sample from a $U(0,1)$ (Sackrowitz & Samuel-Cahn 1999, Murdoch, Tsai & Adcock 2008).

Let V be a random variable with cdf F, and let $V_{(m)} = \max\{V_1, V_2, \ldots, V_m\}$ be its maximum in a random sample of size m. The exact distribution of $V_{(m)}$ is given by Casella & Berger (2001):

$$P(V_{(m)} \leq t) = \{F(t)\}^m \tag{3}$$

Note that if F is not known, (3) cannot be calculated. However, Serfling (1980, pp. 89) presents an alternative using extreme values theory and asymptotic results. As in a GWAS $m \to \infty$ independent hypothesis are being tested, to build up our methodology on such results seems intuitive.

2.2. The Test

Consider the random variable

$$D_m = (V_{(m)} - a_m)/b_m \tag{4}$$

with $V_{(m)}$ as previously defined. For some choices of constants $\{a_m\}$ and $\{b_m\}$, the limiting distribution of D_m is known (Serfling 1980, pp. 89). It follows from the $U(0,1)$ null distribution of the p-values that $-\log(p)$ has a standard exponential distribution with parameter $\lambda = 1$, and choosing $a_m = \log(m)$ and $b_n = 1$ yields (Serfling 1980, pp. 90)

$$
\begin{aligned}
F_{D_m}(t) &= P(D_m \leq t) \\
&= P(V_{(m)} - \log(m) \leq t) \\
&\to e^{-e^{-t}}, \quad m \to \infty
\end{aligned}
\tag{5}
$$

making possible the calculation of (3). It is straightforward to show that the limiting density function of D_m is given by

$$
\begin{aligned}
f_{D_m}(t) &= \frac{d}{dt} F_{D_m} \\
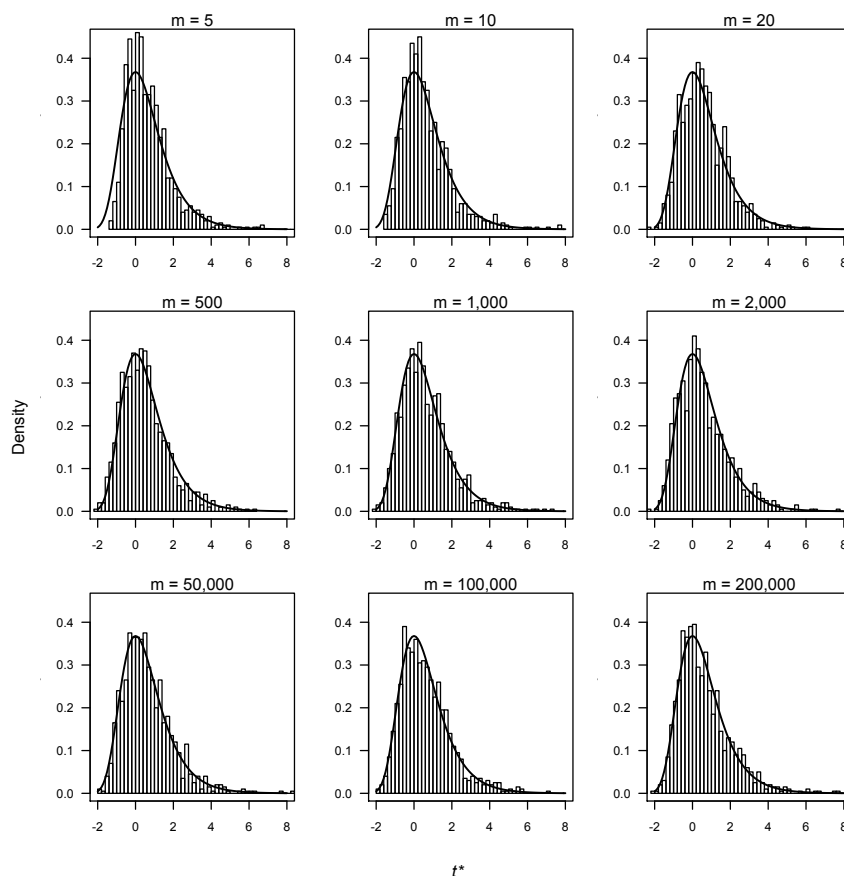&\to \exp\{-(t + \exp(-t))\}, \quad m \to \infty
\end{aligned}
\tag{6}
$$

FIGURE 1: Simulation-based distribution of t^* for different values of m when the p-values come from a $U(0,1)$ and $\alpha = 0.05$. Here, the black line corresponds to $f_{D_m}(t)$ in (6).

We shall say that the ith p-value is significant at level α if

$$t_i^* > t_c \qquad i = 1, 2, \ldots, m \tag{7}$$

where

$$t_i^* = -\log(-\log(1 - P_i)) \tag{8}$$

is the test statistic and t_c the critical value of the test at level α, e.g., t_c is such that

$$P(V_{(m)} - \log(m) \geq t_c) = \alpha \tag{9}$$

Combining (5) and (9), and solving for t leads to

$$t_c = -\log(-\log(1 - \alpha)) \tag{10}$$

In Figure 1 we depict the simulation-based distribution of t^* when $P_1, P_2, \ldots, P_m \overset{iid}{\sim} U(0,1)$ for different values of m.

It is also possible to establish some equivalencies between our proposed method and others. For instance, if the Bonferroni (1935) method is to be applied to control by multiple testing (Shaffer 1995), the critical value

$$t_c^* = t_c + \log(m) \tag{11}$$

should be used instead of (10). This result is particularly useful in situations where a stringent control of the FWER (and hence the false positives) is required.

2.3. Using the Test

The following steps are suggested for detecting those p-values being statistically significant:

1. For each p-value, calculate t_i^* as in (8) and denote them as $t_1^*, t_2^*, \ldots, t_m^*$. Here, higher values of t^* indicate strong evidence against H_0 in (1).

2. Determine which $t_i^{*\prime}$s are greater than t_c (or t_c^*).

3. Define the p-values from step 2 as potential candidates.

In order to facilitate the use of our proposal, an implementation of the afore-mentioned steps in R (R Core Team 2013) is provided in 4. This function takes a vector of p-values as the main argument, calculates the test statistic and the critical value, and prints the number of rejected p-values as well as the rejection rate. Furthermore, an invisible object (a list) with three components is returned; this list contains the actual p-value, the test statistic and the correspondent decision (significant: `TRUE`; not significant: `FALSE`). If necessary, such an object can be used for further analyses.

3. Examples

In this section, we consider two gene expression data sets to illustrate the usefulness of our proposed method for the identification of significant p-values.

3.1. Tumor Data

Golub et al. (1999) present a generic approach to cancer classification based on gene expression monitoring by DNA microarrays. As a test case, the authors use gene expression data from 3,051 genes in 38 tumor mRNA samples from patients with leukemia; 27 samples come from patients with lymphoblastic leukemia (ALL)(cases) and 11 from patients with acute myeloid leukemia (AML)(controls). For analysis, the processed data was obtained from the `multtest` package (Pollard, Gilbert, Ge, Taylor & Dudoit 2011).
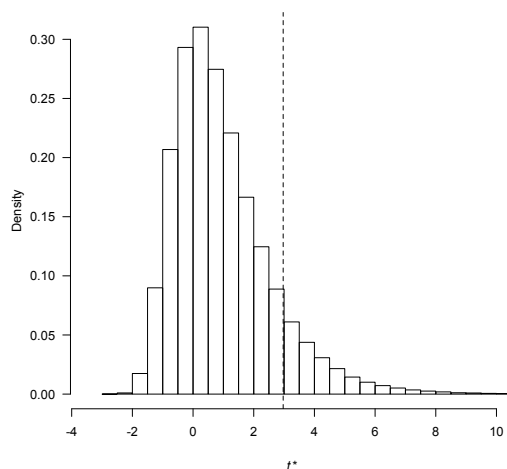
FIGURE 2: Distribution of t^* for the microarray data in Golub et al (1999). The vertical dotted line represents the critical value of the test for $\alpha = 0.05$ when no correction for multiple testing is applied.

We tested whether the ith gene ($i = 1, 2, \ldots, m = 3,051$) was differentially expressed (DE), i.e., if there was any statistical difference between the expression levels in cases and controls. This is equivalent to test

$$H_{0,i} : \mu_{\text{ALL},i} = \mu_{\text{AML},i} \quad \text{vs.} \quad H_{1,i} : \mu_{\text{ALL},i} \neq \mu_{\text{AML},i} \tag{12}$$

As implemented in the `genefilter` package (Gentleman, Carey, Huber & Hahne 2011), we used a two-sample t-test for testing (12) and calculated the p-value for each gene. Further, these p-values were used to calculate (10) and (11).

In Figure 2 we present the distribution of t^* using equation (8) for the m genes. When no correction for multiple testing is applied on the p-values, a total of 1,045 (34.3%, $t_c = 2.97$) genes were found to be DE, which were reduced to 98 (3.2%, $t_c^* = 10.99$) when a Bonferroni correction was applied. On the other hand, when the p-values were FDR-corrected before applying our methodology, 681 (22.3%, $t_c = -5.05$) were found to be DE. Equivalent results were obtained using built-in R function `p.adjust()`.

3.2. Type 2 Diabetes Data

Mootha et al. (2003) presented an analytical strategy for detecting modest but coordinate changes in gene expression using DNA microarray data. This data consists of 22,283 gene expression levels measured in 43 age-matched males skeletal muscle biopsy samples, 17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance (IGT) and 18 with type 2 diabetes (T2D).

After randomly selecting 1,000 gene expression levels for T2D samples from the original data, the linear correlation coefficient ρ for each pair of genes was

calculated. ρ might be seen as a «proxy» of the potential interacting effects between pair of genes.

TABLE 1: Significant correlation coefficients for pairs of genes in 1,000 randomly selected gene expression levels (Mootha et al. 2003) when only T2D samples are included. Bonferroni correction was applied. CI: Confidence Interval.

Genes	$\hat{\rho}$	95%CI	t-statistic	t_c^*	Raw P-value
G12-G720	0.939	(0.840, 0.977)	10.899	18.621	8.16×10^{-9}
G291-G350	0.938	(0.837, 0.977)	10.777	16.643	9.60×10^{-9}
G490-G698	0.927	(0.812, 0.973)	9.903	17.274	3.14×10^{-8}
G108-G434	-0.921	(-0.971, -0.797)	-9.459	16.642	5.91×10^{-8}
G210-G720	0.920	(0.795, 0.970)	9.409	16.570	6.36×10^{-8}
G293-G308	0.917	(0.787, 0.969)	9.196	16.257	8.69×10^{-8}

A total of $m = 499,500$ hypothesis of the form

$$H_{0,i} : \rho_i = 0 \quad \text{vs.} \quad H_{1,i} : \rho_i \neq 0 \qquad i = 1, 2, \ldots, m \qquad (13)$$

were tested. For $\alpha = 0.05$, 52,576 (10.53%, $t_c = 2.97$) correlation coefficients were significant when no correction for multiple testing was applied, which reduced to 319 (0.06%, $t_c = 2.97$) and 6 ($\sim 0\%$, $t_c^* = 16.09$), respectively, when the FDR and Bonferroni corrections were used. Results for the latter are presented in Table 1.

4. Discussion

In this paper, we propose a new method to determine whether a p-value is significant under a multiple testing setting while controlling (or not) the FWER. Our proposal, based on the m-th order statistic of a $U(0,1)$ distribution, has been shown to give equivalent results to Bonferroni's method while controlling the FWER, and to classical methods while not. Furthermore, under the null hypothesis, the proportion of true null hypothesis being rejected is close to the nominal level α. Observe that, by no means, we are stating that our method is improving any of the other alternatives available in the literature to correct by multiple testing, and which have extensively been applied in the genetics field.

The contribution of this paper can be seen under two perspectives. First, it offers a graphical alternative to represent p-values and the cutoff value beyond which, in the genetic context, we consider that a SNP (or gene in a microarray) is statistically significant. Second, the use of asymptotic statistics and extreme values theory in genetics. In a review of the literature previous to the writing of this paper, we found no mention or application of these two important concepts in genetics. The main advantages of this new approach are the direct calculation of the cutoff value labelling a p-value as significant, the simplicity of its calculations, and how easy it is to graphically represent the results. Computationally, our approach is better than the FDR (Benjamini & Hochberg 1995) as it does not require to store all the p-values.

Although in our applications section we showed how to use our approach to determine significant p-values with GWAS and microarray data, it is not limited,

under any circumstance, to these type of data. The main reason for this is that our approach uses the p-values of the hypotheses tested regardless of the type(s) of data on which they have been tested. Future extensions of this methodology include considering correlated tests as those proposed by Benjamini & Yekutieli (2001).

Ackowledgements

References

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), 389–300.

Benjamini, Y. & Yekutieli, D. (2001), 'The control of the false discovery rate in multiple testing under dependency', *Annals of Statistics* **29**(4), 1165 – 1188.

Bonferroni, C. E. (1935), 'Il calcolo delle assicurazioni su gruppi di teste', *Studi in Onore del Professore Salvatore Ortu Carboni,* pp. 13–60.

Casella, G. & Berger, R. (2001), *Statistical Inference*, 2 edn, Duxbury Press, United States of America.

Devroye, L. (1986), *Non-Uniform Random Variate Generation*, New York: Spring-Verlang.

Gentleman, R., Carey, V., Huber, W. & Hahne, F. (2011), *genefilter: Methods for filtering genes from microarray experiments*. R package version 1.34.0.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–537.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G. & Macgregor, S. (2010), 'A versatile gene-based test for genome-wide association studies', *The American Journal of Human Genetics* **87**(1), 139 – 145.

Manolio, T. A. (2010), 'Genomewide association studies and assessment of the risk of disease', *New England Journal of Medicine* **363**(2), 166–176.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. & Groop, L. C. (2003), 'Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes', *Nature Genetics* **34**(3), 267–73.

Murdoch, D., Tsai, Y. & Adcock, J. (2008), 'P-values are random variables', *The American Statistician* **62**(3), 242–245.

Nyholt, D. R. (2004), 'A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other', *The American Journal of Human Genetics* **74**(4), 765 – 769.

Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S. & Dudoit, S. (2011), *multtest: Resampling-based multiple hypothesis testing.* R package version 2.8.0.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*http://www.R-project.org/

Sackrowitz, H. & Samuel-Cahn, E. (1999), 'P Values as Random Variables-Expected P Values', *The American Statistician* **53**(4), 326–331.

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, United States of America.

Shaffer, J. P. (1995), 'Multiple hypothesis testing', *Annual Review of Psychology* **46**, 561–584.

Appendix. Detect significant p-values in R using the proposed method

```
#-------------------------------------------------------------------------------
#                               ARGUMENTS
#   p           vector of p-values
#   plot        histogram of transformed p-values? (default: TRUE)
#   corrected   correction by multiple testing? (default: TRUE)
#   line        add vertical line indicating critical value?  (default: TRUE)
#   alpha       type I error probability (default: 0.05)
#   ...         additional arguments passed to hist()
#-------------------------------------------------------------------------------
pvaltest <- function(p, plot = TRUE, corrected = TRUE,
                            line = TRUE, alpha = 0.05, ...){
m <- length(p)
ti <-  -log(-log(1 - p))
tc <- -log(-log(1 - alpha))
tcstar <- tc + 1 * corrected * log(m)
total <- sum(ti > tcstar)

if(plot){
hist(ti, breaks = 50, prob = TRUE, las = 1,
   xlab = expression(italic(t*"*")), ylab = "Density", ...)
   abline(v = 1* line * tcstar, col = 1, lty = 2)
}
cat("Number of tests = ", m, "\n")
cat("Critical value = ", round(tcstar, 2), "\n")
cat("Total rejected = ", total, "(", round(100*total/m, 2), "%)", "\n")
invisible(list(p.value = p, statistic = ti, reject = ti > tcstar))
}

## Example
set.seed(123)
p <- c(runif(100, 0, 1e-4), runif(5000))
res <- pvaltest(p, main = "")
# Number of tests =  5000
# Critical value =  11.49
# Total rejected =  7 ( 0.14 %)

str(res)
# List of 3
#  $ p.value  : num [1:5100] 2.88e-05 7.88e-05 4.09e-05 8.83e-05 9.40e-05 ...
#  $ statistic: num [1:5100] 10.46 9.45 10.1 9.33 9.27 ...
#  $ reject   : logi [1:5100] FALSE FALSE FALSE FALSE FALSE TRUE ...
```