Illustrations on Using the Distribution of a P-value in High Dimensional Data Analyses

Xiaojun Hu[1], **Gary L. Gadbury**[2], **Qinfang Xiang**[1], and **David B. Allison**[3]

Xiaojun Hu: ; Gary L. Gadbury: gadbury@ksu.edu; Qinfang Xiang: ; David B. Allison:

[1] Endo Pharmaceuticals, Chadds Ford, Pennsylvania 19317

[2] Department of Statistics, Kansas State University, Manhattan, KS 66506

[3] Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama 35294

Abstract

Several statistical methods have recently been developed that use the distribution of P-values from multiple tests of hypotheses to analyze data from high-dimensional experiments. These methods are only as valid as the P-values that were derived from test statistics. If an incorrect distribution for a test statistic was used, the P-value will not be valid and the distribution of P-values from multiple test statistics could give misleading results. Moreover, if the correct distribution of a test statistic is used, a distribution of P-values may still give misleading results if P-values are correlated. A primary focus of this paper is on the distribution of a P-value under a null hypothesis, and the test statistic that is considered is the number of rejected null hypotheses. Two issues are demonstrated using six data examples, two that are simulated and four from actual microarray experiments. The results provide some insight into how much of an effect might be introduced into a distribution of P-values if invalid P-values are computed or if P-values are correlated. Additional illustration is given regarding the distribution of a P-value under an alternative hypothesis and some approaches to modeling it are presented.

Keywords

correlation; FDR; microarray; multiple testing; type I error

1. Introduction

This paper focuses on two issues that can arise when analyzing high dimensional data using the distribution of P-values from tests of multiple hypotheses. Both issues are discussed in a context where a "global null hypothesis" is true, that is, the null hypothesis is true for all tests. This context is important because, if a global null hypothesis is true, then a histogram of valid P-values from multiple tests is expected to be approximately uniform on the interval from 0 to 1, for example, as shown in Figure 1(A). Many proposed methods that use the distribution of P-values from high dimensional experiments (HDEs) have two tasks (cf., Broberg 2004; Yang and Yang 2006): estimating the proportion of true null hypotheses (hereafter denoted π_0), and using a model (empirical or parametric) for the distribution of P-values to estimate desired quantities such as the false discovery rate (FDR, Benjamini and Hochberg, 1995). If an observed distribution of P-values appeared uniform then one might

Correspondence to: Gary L. Gadbury, gadbury@ksu.edu.

expect that π_0 is close to 1. If there is a clustering of P-value near zero, then the distribution would suggest there is evidence that some genes are differentially expressed and that π_0 is less than 1. Many methods estimate π_0 using algorithms that assess how much an observed distribution of P-values deviates from a uniform.

The first issue that is considered is the effect of an incorrect statistical test on the distribution of a P-value. When the incorrect statistical test is used to compute a P-value, the distribution of a P-value is no longer expected to be uniform under a null hypothesis, and the observed distribution of P-values from multiple tests may then be misleading. We show some illustrations of a non-uniform distribution of P-values under the global null hypothesis (i.e., all null hypotheses are true) and discuss the implications in the context of the distributions of P-values shown in Figure 1. We specifically consider a two treatment comparison study where a two sample t-test is typically used to compute a P-value.

A second issue is then discussed that arises when the correct statistical test is used to compute the P-value but groups of P-values are correlated. The distribution of P-values under the global null hypothesis is expected to be uniform but may vary widely from experiment to experiment. Since sample sizes are usually small relative to the large number tests from an HDE, the correlation structure cannot be estimated from observed data. The effect of certain assumed correlation structures on results from a statistical method has been evaluated using simulation experiments (e.g., Allison et al., 2002; Gadbury et al., 2003; Nettleton et al., 2006), and some analytic modeling of a correlation parameter in a high dimensional experiment presented by Efron (2007). We show here that the effect of an assumed correlation structure on key aspects of the distribution of P-values under the global null hypothesis can be assessed analytically using a simple statistic, the number of P-values below a threshold. The implications of this are discussed in the context of the P-value distributions in Figure 1.

Finally, some discussion is given regarding some properties of the P-value under the alternative hypothesis. In particular, approaches to modeling the alternative distribution and interpreting results are discussed.

Topics in this paper have been discussed before in other contexts (cf., Westfall and Young 1993; Donahue, 1999; Sackrowitz and Samuel-Cahn, 1999). However, recent research activity in using the distribution of a P-value for analyzing high dimensional data suggest that it would be useful to re-explore these issues and their consequences as they occur in HDEs. The type of HDE specifically considered here is a microarray experiment where tests for differential genetic expression across two or more treatment conditions are carried out on thousands of genes simultaneously (see, for example, Allison et al. 2006b for more information on microarrays and related genomics techniques). Some related discussion on the distribution of a P-value as a random variable under the null and alternative hypotheses was given in Hu et al., (2005). In Section 2, the data that produced the six plots in Figure 1 are described. In section 3, the effect of an incorrect statistical test and correlated P-values on the observed distribution of P-values is considered. Section 4 discusses the alternative distribution of a P-value. We summarize with some observations and conclusions.

2. Data Descriptions

Figure 1(A) shows a distribution of P-values that would be expected if a global null hypothesis is true. Figures (C) – (F) show distributions for which some genes may be differentially expressed. The signal appears strongest in (F) and weakest in (E), yet all four of these plots show a clustering of P-values near zero. On some occasions, however, we have seen distributions like that in Figure 1(B) (see Page et al. 2006 for other examples).

This shape is difficult to interpret because fewer P-value are clustering near zero than would even be expected if the global null hypothesis were true.

In (A), 10,000 P-values were independently simulated from a uniform distribution. In (B), 10,000 test statistics were generated from a 10,000 dimensional multivariate normal distribution with marginal distributions that were standard normal but with a block diagonal correlation structure that was used by Allison et al. 2002 and Gadbury et al. 2003. The correlation matrix is of a type $R = [J_b\rho + (1 - \rho)I_b] \otimes I_m$ where J_b is a square matrix of all ones of size b, and I_b is a b-dimensional identity matrix. Thus, R is block diagonal matrix with m blocks of size b. Within each block, the correlation coefficient between all pairs is ρ and test statistics in different blocks are uncorrelated. The values used to simulate the data were $\rho = 0.6$, $b = 300$, $m = 33$ so that 9900 statistics had a correlation matrix of the form given by R and the remaining 100 statistics were in a remainder block of size 100. Two-tailed P-values were then obtained using the standard normal distribution as a reference distribution.

The P-values in (C) – (F) were obtained from actual data sets where, in each, two treatments were being tested for differential expression. In all plots, P-values were computed using two sample pooled variance t-tests. In (C), human rheumatoid arthritis synovial fibroblast cell line samples were stimulated with tumor necrosis factor-α where one group ($n = 3$) had the Nf-κB pathway taken out by a dominant negative transiently transfected vector and the other group ($n = 3$) had a control vector added. An objective of the experiment was to determine what genes were differentially expressed across the two groups. Further details of the experiment are in Zhang et al. (2004). Gene expression measurements were obtained for 12,625 genes.

In (D), the study sought genes that are differentially expressed between CD4 cells taken from five young and five old male rhesus monkeys (Kayo T, Prolla TA, Weindruch R. unpublished data). There were 12,625 genes in the experiment and expression levels were normalized using quantile-quantile normalization. In (E), mammary gland tissue was dissected from 50 day old rats where one group had a standard diet (10 rats) and another group had a diet supplemented with resveratrol (10 rats). The data were obtained from the Center for Nutrient-Gene Interaction (CNGI) at University of Alabama at Birmingham (www.uab.edu/cngi). Resveratrol is found in red grapes and their products such as red wine. It is thought that since the compound is similar to estrogen, it may affect the response of gene networks to physiologic estrogens and hence alter the development of a cancer gene. Expression levels were measured for 31,042 genes. In (F), the study sought differences in gene expression between adipocytes from lean (19 subjects) and obese (19 subjects) humans; 63,149 gene expression levels were measured for all subjects (Lee et al., 2005).

Tests for differential expression are declared significant if a P-value falls below a set threshold, τ. A simple statistic that is considered here is the number of P-values from an HDE that are less than or equal to τ. We define $N = \{\#\ pvalues \leq \tau\}$. Values of N for three different τ are given in Table 1 for the six sets of P-values in Figure 1.

3. Illustrations on the null distribution of P-values

Suppose that the test statistic T has a known continuous distribution under a null hypothesis H_0 and let its cumulative distribution function (CDF) be denoted by $G_T(t)$. Consider testing H_0: $\delta = 0$ versus H_a: $\delta \neq 0$ where δ represents an "effect size", e.g., a difference between two (or more) means. The two-tailed P-value is defined by

$$p=p(t)=\begin{cases} 2prob(T \le t) & if \quad t \le 0 \\ 2(1-prob(T \le t)) & if \quad t>0 \end{cases} = \begin{cases} 2G_T(t) & if \quad t \le 0 \\ 2(1-G_T(t)) & if \quad t>0 \end{cases} \tag{1}$$

Its CDF is

$$\begin{aligned} &G_p(p|H_0)=prob\left(G_T \le \tfrac{p}{2}|H_0\right) I_{(-\infty,0]}(T)+prob\left(G_T>1-\tfrac{p}{2}|H_0\right) I_{(0,\infty)}(T) \\ &=prob\left(T \le G_T{}^{-1}\left(\tfrac{p}{2}\right)|H_0\right)+1-prob\left(T \le G_T{}^{-1}\left(1-\tfrac{p}{2}\right)|H_0\right) \\ &=G_T\left(G_T{}^{-1}\left(\tfrac{p}{2}\right)\right)+1-G_T\left(G_T{}^{-1}\left(1-\tfrac{p}{2}\right)\right) \\ &=\tfrac{p}{2}+1-\left(1-\tfrac{p}{2}\right)=p \end{aligned} \tag{2}$$

which is the CDF of a uniform distribution. A P-value with this null distribution we refer to as a valid P-value. Let $N^0=\{\#\, pvalues \le \tau\}$ when the global null hypothesis is true. Then the expected value of N^0, i.e., $E(N^0)$, in a study of K hypothesis tests when the correct statistical test is used is equal to $K\tau$.

3.1 The expected value of N^0 when an incorrect test is used

Suppose that the t-distribution with v degrees of freedom is chosen as $G_T(t)$ and a two-tailed P-value is defined as in (1) using this reference distribution. Values of v will be chosen corresponding to degrees of freedom for a two sample pooled variance t-test that would be used for the data sets shown in Figure 1. The two sample t-test has been a common test for differential expression across two treatment conditions. Suppose that the actual distribution of the test statistic has CDF $H_T(t)$. The derivation for $G_P(p|H_0)$ is similar to (2) above, except that the CDF is,

$$\begin{aligned} &G_p(p|H_0)=prob\left(G_T \le \tfrac{p}{2}|H_0\right)+prob\left(G_T>1-\tfrac{p}{2}|H_0\right) \\ &=H_T\left(G_T{}^{-1}\left(\tfrac{p}{2}\right)\right)+1-H_T\left(G_T{}^{-1}\left(1-\tfrac{p}{2}\right)\right). \end{aligned} \tag{3}$$

The density is,

$$g_p(p|H_0)=\frac{\partial}{\partial p}G_p(p|H_0)=\frac{h_T(G_T^{-1}(p/2))}{2g_T(G_T^{-1}(p/2))}+\frac{h_T(G_T^{-1}(1-p/2))}{2g_T(G_T^{-1}(1-p/2))}. \tag{4}$$

Lower case letters represent a density and the notation in (3), $G_T^{-1}(a)$, the a^{th} quantile of $G_T(\cdot)$. This distribution is not uniform if $H(\cdot)$ and $G(\cdot)$ are different distributions. How "far off" from a uniform is illustrated by using the $G_T(t)$given above and two different distributions for $H(\cdot)$: the logistic and the double exponential distribution, both with location at the origin. These two distributions were chosen because, like the t-distribution, they are symmetric about zero. The other parameter for these two distributions was chosen so that the variance of the distribution matched that of the corresponding t-distribution that would be used for each data set in Figure 1. Thus the first three moments of $H(\cdot)$ matched the corresponding moments of the t-distribution, so H and G differ only on moments equal to or

higher than the 4th. This choice then illustrates the effect of the lesser understood 4th moment (Ruppert 1987) or higher moments on the distribution of the P-value. For reference, the second and fourth moments of the distributions that were used, representing the degrees of freedom associated with each data set, are given in Appendix A, Table A1. One can see from table A1 that the fourth moment of the logistic distribution is less than that of the t-distribution for lower degrees of freedom but exceeds that of the t-distribution for larger degrees of freedom, at which the t-distribution becomes closer to a standard normal distribution. The fourth moment of the double exponential matches that of the t-distribution with 6 degrees of freedom but exceeds it for other degrees of freedom. The fourth moment of a t-distribution does not exist for $v = 4$.

3.2 Interpreting results from a correct test versus an incorrect test

The first entry in each cell in Table 2 shows the expected value of N^0, providing the correct test (i.e., the t-test) was used to compute all of the P-values in Figure 1, (A) – (F). Data were simulated for data sets (A) and (B) and, so, $v = 6$ degrees of freedom was chosen for those. For the other data sets there would be 4 degrees of freedom for (C), 8 for (D), 18 for (E), and 36 for data set (F). The other two entries in each cell in Table 2 show the expected value of N^0 if the actual distribution of the test statistic was logistic (the second entry) or double exponential (third entry), where parameter values were chosen as described above.

There are two key results in Table 2. The first is that expected values of N^0 may be either inflated or deflated if the true reference distribution of the test statistics is not t_v. The results show the rather extreme case where the test statistics had the same distribution for every test. Suppose that $v = 6$ and there are $K = 10{,}000$ tests, as was the case for data sets (A) and (B). If a test is to be declared significant at $\tau = .01$, one would expect 100 type I errors when the global null hypothesis is true. If one rejected 138 null hypotheses at this threshold, one could interpret this as evidence that some genes are differentially expressed since the value N exceeds what would be expected. However, if the true reference distribution is double exponential, then this number would be expected even if no genes are differentially expressed. If the true reference distribution was logistic, then one would only expect 82 type I errors, so 138 rejected null hypotheses would suggest that the global null hypothesis is not true. For larger degrees of freedom, the effect of an incorrect reference distribution on expected type I errors becomes more pronounced at smaller thresholds, with expected numbers of type I errors being several orders of magnitude higher than what would be expected if the correct distribution was used. At smaller degrees of freedom, the choice of a t-distribution is fairly robust (and even conservative in some cases) in expected number of type I errors when the true distribution is one of the other two. It is also reasonably robust at the larger threshold of 0.05.

The second result from Table 2 is that actual observed counts in Table 1 can be compared to what would be expected under a global null hypothesis and under the three different reference distributions. For example, the counter-intuitive distribution in Figure 1(B) resulted from correlated test statistics and the numbers of P-values below the threshold (Table 1) was smaller than what would be expected. However, these smaller numbers of 60 and 3 for thresholds 0.01 and 0.001 are closer to what would be expected if a logistic distribution was the true distribution of the test statistic. However, this is not the case for the larger threshold of 0.05, where the observed number of 346 is not consistent with any of the three distributions, suggesting perhaps that there is some other explanation for the shape in Figure 1(B). The distribution in Figure 1(E) suggests that some genes might be differentially expressed, and the numbers of P-values below the three thresholds (Table 1) do exceed what would be expected if the correct reference distribution was the t-distribution and if the global null hypothesis is true. However, at thresholds of 0.01 and 0.001, the numbers in Table 1 are closer to what would be expected if the true distribution was logistic and even

below expected values if the true distribution was double exponential. But at the larger threshold of 0.05, the observed number exceeds what would be expected for either of the other two distributions. Nevertheless, the relatively weak signal in Figure 1(E) makes it difficult to determine whether it is due to some genes being differentially expressed, or an effect of the choice of reference distribution when computing the P-values. The strong signals for data sets (C) and (F) are apparent in the tables as the numbers in Table 1 far exceed what would be expected for any of the distributions in Table 2. A signal (or lack thereof) in a distribution of P-values can also be produced when P-values are correlated.

3.3 Correlated P-values

In microarray data, test statistics for gene specific tests of differential expression are likely to be correlated, a fact that was investigated by Gadbury et al. (2003), Qui et al. (2005), and Efron (2007). Correlation among genes in a microarray experiment can greatly affect the variance of estimates of quantities important to the investigator (Owen 2005; Qiu et al., 2006).

Suppose that the correct test was used to compute a P-value so that the null marginal distribution of a P-value is uniform[0,1]. Consider, as in Schweder and Spjøvoll 1982, a Bernoulli variable

$$D_i = \begin{cases} 0 & if \quad p_i \le \tau \\ 1 & if \quad p_i > \tau \end{cases}$$

that is, $D_i \sim$ Bernoulli$(1-\tau)$ where p_i is a P-value from the i^{th} hypothesis test and τ is a threshold. Thus the i^{th} test will be declared statistically significant when $D_i = 0$.

Suppose that K test statistics for gene specific tests for differential expression have some specified correlation structure given by a K by K matrix, and that, again, the global null hypothesis is true. The same block diagonal correlation structure described in section 2 for data set (B) is considered here. It is a block diagonal matrix with m blocks of size b plus a remainder block for r genes so that $K = bm + r$. Within each block, the correlation coefficient between all pairs is ρ and test statistics in different blocks are uncorrelated. Define N_b to be the number of tests declared not significant in one block of size b, and N_r the analogous number for the remainder block. The expected value of N_b is

$\mathrm{E}(N_b) = \mathrm{E}(\sum_{i=1}^{b} D_i) = b(1-\tau)$ and the variance of N_b is

$\mathrm{Var}(N_b) = \mathrm{Var}(\sum_{i=1}^{b} D_i) = b\tau(1-\tau) + b(b-1)\mathrm{Cov}(D_1, D_2)$. These quantities for N_r are similar with r replacing b.

Denote T_1 and T_2 as test statistics for two hypotheses that are bivariate normal with standard normal marginal distributions and with correlation equal to ρ. Then,

$$\begin{aligned} \mathrm{Cov}(D_1, D_2) &= P(D_1 = 1, D_2 = 1) - (1-\tau)^2 \\ &= P(|T_1| \le z_{\tau/2}, |T_2| \le z_{\tau/2}) - (1-\tau)^2 \\ &= \Phi(z_{\tau/2}, z_{\tau/2}) - \Phi(-z_{\tau/2}, z_{\tau/2}) - \Phi(z_{\tau/2}, -z_{\tau/2}) + \Phi(-z_{\tau/2}, -z_{\tau/2}) - (1-\tau)^2 \end{aligned}$$

where $\Phi(z_1, z_2)$is a bivariate normal CDF of (T_1, T_2) evaluated at (z_1, z_2). As before, let N^0 be the number of genes out of a total K that are declared differentially expressed when the global null hypothesis is true. The expected number of significant genes is,

$$E(N^0)=K-\mathrm{E}\left(\sum_{j=1}^{m}N_{bj}\right)-E(N_r)=K-(1-\tau)(mb+r)=K-K(1-\tau)=K\tau$$

and the variance of N^0 is

$$\mathrm{Var}(N^0)=\mathrm{Var}(K-\sum_{j=1}^{m}N_{bj}-N_r)=\sum_{j=1}^{m}\mathrm{Var}(N_{bj})+\mathrm{Var}(N_r)=m\mathrm{Var}(N_b)+\mathrm{Var}(N_r)$$

The standard deviation of N^0 is then $\sigma_{N^0}=\sqrt{\mathrm{Var}(N^0)}$.

3.4 Interpreting the effect of correlated P-values on the number of significant results

Table 3 illustrates the effect of $\sigma_N 0$ for varying block sizes and correlation values on the interpretation of N^0. Since the correct statistical test was assumed, the expected numbers are the same as the first line at each threshold in Table 2. The standard deviation of the numbers increases with increasing correlation within blocks and/or increasing size of the blocks. The strongest dependence structure uses blocks of size 500 with correlation between all pairs within a block equal to 0.8. When $\rho = 0$, all tests are uncorrelated.

Referring to the numbers in Table 1, one can see that the values of N for data set (A) are within one standard deviation (SD) of their expected values under the global null hypothesis at the three thresholds (just beyond 1 SD at $\tau = 0.01$). The numbers in Table 1 for data set (B) are below what would be expected if the global null hypothesis is true; however, the numbers are plausible with a medium to strong correlation structure among P-values. For example, at $\tau = 0.01$the observed number of 60 is less than 2 SD below the expected number if ρ =0.4 and the block size is 100. The numbers in Table 1 for data set (C) are way above what would be expected if the global null hypothesis is true even with a strong dependence structure among P-values. In other words, a strong correlation structure would not likely explain the strong signal seen in data set (C), a similar result seen in Gadbury et al., (2003) for this same data set by using simulations. This is also even truer for the strong signal seen in data set (F). The signal seen in data set (D) would require a relatively strong correlation structure to explain if the global null hypothesis is true, and the numbers in Table 1 for data set (E) are around 2 SD above the expected values if the global null hypothesis is true and ρ = 0.4 in block sizes of 500.

One objective in a high dimensional experiment is to determine if the number of statistically significant results (“discoveries”) is higher (or lower) than expected if the global null hypothesis is true. A key point here is that, if one is willing to conjecture a “plausible” correlation structure, then it is relatively straightforward to assess the effect of such correlation on the number of discoveries from a given experiment, rather than resorting to computer simulations. This provides some added insight into how strong a signal is relative to what could be seen due to correlation. The block diagonal structure for the correlation matrix was not necessary to assume but it did allow for some simplification of the above

formulas, and this particular structure illustrated here is likely to be too simplistic for real microarray data (Allison et al., 2006a).

4. Illustrating the alternative distribution of a P-value

Some methods for analyzing data from high dimensional experiments (e.g., Allison et al. 2002; Liao et al. 2004; Pounds and Cheng 2004) have used a mixture model of the form,

$$f(p)=\pi_0 f_0(p)+(1-\pi_0)f_1(p) \tag{5}$$

where p is a P-value, f_0 is the uniform density function, and f_1 is the density function under the alternative hypothesis, and π_0 is the weight on f_0, interpreted here as the proportion of true null hypotheses. Resulting from the mixture distribution is a formula for the false discovery rate, *FDR* (Benjamini and Hochberg 1995; Storey 2002), computed at specified threshold $\tau \in (0,1)$,

$$FDR=\frac{\pi_0 F_0(\tau)}{\pi_0 F_0(\tau)+(1-\pi_0)F_1(\tau)} \tag{6}$$

where F_0 is the CDF of a uniform distribution, and F_1 is the CDF for the alternative hypothesis. A local version of *FDR* (Efron 2004) is available using (6) but replacing the CDFs with corresponding densities if they exist. Hung et al. (1997) showed that F_1 depends on the sample size, effect size, and the distribution of the test statistic used to compute the P-value. Parker and Rothenberg (1988) and Allison et al. (2002) chose a standard two parameter beta distribution for F_1 because of its flexibility in modeling shapes on the unit interval, and Pounds and Morris (2003) did the same but set the second shape parameter equal to 1. The following illustrates a "true distribution" for a P-value under the alternative hypothesis and the suitability of the standard two parameter beta distribution in approximating the shape.

Assume there are two samples $X_1, \cdots, X_{n_1} \sim (\mu_X, \sigma_X^2)$ and $Y_1, \cdots, Y_{n_2} \sim (\mu_Y, \sigma_Y^2)$, and hypotheses H_0: $\mu_X = \mu_Y$ versus H_1: $\mu_X \neq \mu_Y$ are being tested. If $\sigma_X^2=\sigma_Y^2=\sigma^2$, $n_1=n_2=n$, and the two distributions are normal, then under H_0 the test statistic $T=\frac{\overline{X}-\overline{Y}}{S_p\sqrt{2/n}} \sim G_T(t)$ which is a central t-distribution with degrees of freedom $\nu=2n-2$, where S_p is the pooled standard deviation. If the alternative hypothesis H_1 is true, then T has a non-central t-distribution with non-centrality parameter $\delta\sqrt{n/2}$ where $\delta=\frac{\mu_X-\mu_Y}{\sigma}$ is the effect size. For the two-tailed test, the P-value is computed as equation (1) and the results for its CDF and pdf, that were also discussed in Hung et al. (1997), and Donahue (1999), are the same as given in (3) and (4) except that the noncentral t-distribution is represented by $H(\cdot)$. If we know the effect size which is a fixed value δ, the density of a P-value under H_1 (for details see appendix) is,

$$f_p(p|H_1)=\frac{1}{2}\left(\frac{g_T\left[G_T^{-1}(1-p/2;\nu,0);\nu,\delta\sqrt{n/2}\right]}{g_T[G_T^{-1}(1-p/2;\nu,0);\nu,0]}+\frac{g_T\left[G_T^{-1}(p/2;\nu,0);\nu,\delta\sqrt{n/2}\right]}{g_T[G_T^{-1}(p/2;\nu);\nu,0]}\right) \tag{7}$$

where the notation G^{-1} (a; b, c) represents the ath quantile of a t-distribution with b degrees of freedom and noncentrality parameter c. The lower case g for the density function is similarly denoted.

It is interesting to note that this true density in (7) can be expressed as a function of beta distributions,

$$f_P(p|H_1)=\frac{b_U\left[B_U^{-1}\left(1-p;\frac{1}{2},\frac{\nu}{2},0\right);\frac{1}{2},\frac{\nu}{2},\frac{n\delta^2}{2}\right]}{b_U\left[B_U^{-1}\left(1-p;\frac{1}{2},\frac{\nu}{2},0\right);\frac{1}{2},\frac{\nu}{2},0\right]} \quad (8)$$

where $B_U^{-1}(a;b,c,d)$ is the a^{th} quantile of a noncentral beta distribution (Gupta and Nadarajah 2004) with shape parameters b and c, and noncentrality d. The density of this distribution is denoted by the lower case b_U. A noncentral beta density function contains an infinite sum and probabilities from the beta distribution require evaluation of an integral. So numerical estimation of parameters in the distribution given by (8) pose more challenge than estimating those from a standard beta distribution.

As mentioned earlier, a standard two parameter beta distribution has been used to model f_P ($p|H_1$). The pdf of a standard two parameter beta distribution is given by,

$$\beta(x;r,s)=I_{(0,1)}(x)\frac{x^{r-1}(1-x)^{s-1}}{B(r,s)} \quad (9)$$

where $B(r,s)=\int_0^1 u^{r-1}(1-u)^{s-1}du$ and $I_{(0,1)}(x) = 1$ for $x \in (0,1)$. Analytically relating the shape parameters of the beta density in (9) to the true density in (8), in particular the noncentrality parameter, appears to be intractable. Numerical experimentation reported elsewhere (Hu et al., 2005) have shown that as the sample size and/or effect size increases, the first shape parameter r decreases and the second, s increases.

A question of interest is whether the standard beta distribution has sufficient flexibility to capture the true distribution for this particular case (i.e., the test statistic has a t-distribution). A "best" approximation to the true distribution was determined two ways. One way used a moment method (MME), matching the first two moments of the true P-value distribution given in (8) with those of the beta distribution in (9). The second way involves minimizing the distance between the two distributions, referred to as the least squares method (LSE). See the appendix for details of these two methods. The pdf of the true density and the two approximate beta densities are given in Figure 2 for two different effect sizes $\delta = 1, 3/2$ occurring in a two sample experiment with n = 4 in each group. The two estimated standard beta distributions using the two methods generally capture the shape of the true density used in this example, and the distribution can be fit to data relatively easily using the "optim" function in R. Xiang et al., (2006) evaluated the precision of estimates of parameters from fitting a mixture of a uniform and a standard beta distribution to a distribution of P-values. A mixture of a uniform and a standard beta distribution fitted to the P-values in Figure 1(c), for example, using maximum likelihood estimation resulted in the estimates for the model parameters in equations (5) and (9),$(\hat{\pi}_0, \hat{r}, \hat{s}) = (0.60,0.54,1.84)$. The further use of this model to compute quantities of interest to investigators was presented in Gadbury et al., (2004).

5. Summary and conclusions

Some consequences of a non-uniform null distribution of a P-value were illustrated here. The illustrations highlight the criticality of obtaining a valid P-value for each test in an HDE situation. Methods have been proposed that borrow information from all test statistics when estimating a reference distribution (e.g., Efron 2004). Adopting such methods can be helpful in improving the validity of inferences versus using normality assumptions and conducting a t-test for each hypothesis. But it may not be reasonable to assume that test statistics under the null hypothesis are identically distributed, and each hypothesis may require its own test to ensure that all P-values are valid. Larger sample sizes that are becoming more common in microarray studies, combined with appropriate pre-processing, transformations, and outlier detection may help in obtaining a correct test statistic for each hypothesis.

The favorable asymptotic (i.e., large sample) properties of tests based on permutations or the bootstrap have been shown (Pollard and van der Laan, 2004), but these tests can produce P-values that are too discrete in small samples to model with continuous distributions in follow-on analysis such as FDR estimation (Gadbury et al. 2003; Pounds and Cheng 2006). In more complex HDEs such as occur in ecological genomics studies (e.g., Travers et al., 2007), a P-value may be computed from a contrast in a mixed effects ANOVA model. Careful consideration of design factors and control of blocking variables then play an important role in ensuring that P-values are valid.

Methods for estimating quantities of interest in high dimensional experiments can be valid on average but could produce estimates with high variance due to correlations among P-values. Evaluating the performance of statistical methods in the presence of a nonestimable correlation structure is challenging, but the effect of a chosen correlation structure can be easily demonstrated on a simple statistic such as the number of rejected null hypotheses. The choice of a plausible correlation structure is difficult, but in some situations results from clustering or knowledge of gene ontology classes (cf., Osier et al., 2004) may give added insight into the structure of correlation that may be present in a high dimensional data set. Efron (2007) recently proposed a method to detect correlation, quantify the strength of correlation, and ways to make adjustments when computing quantities like FDR. The effect of a correlation structure on more complicated statistics such as parameter estimates from a mixture model may require computer simulations (Allison et al., 2002). How to simulate high dimensional data with a realistic correlation structure remains a challenge (Allison et al., 2006a), but one approach was recently proposed in Gadbury et al., (2008).

Finally, the distribution of a P-value under the alternative hypothesis depends on the test statistic that is used to obtain the P-value. For some test statistics, the true distribution can be intractable and difficult to fit to data. A standard beta distribution is relatively easy to use when modeling this distribution and, in the particular illustration considered here, the standard beta distribution captured the shape of the true distribution. The uniform-beta mixture from Allison et al., (2002) had been used extensively to model distributions of P-values and is implemented in an on-line tool described in Trivedi et al., (2005).

New insights being offered by investigators have important implications for high dimensional data analyses and they are opening up new avenues for inquiry and discovery with respect to challenges in multiple testing. We hope our discussion has served to highlight some important considerations when using the distribution of P-values to answer questions of interest in high dimensional investigations.

Acknowledgments

The authors thank John Mountz, Richard Weindruch, Paska Permana, and Grier Page for access to the data used here. This research supported in part by NIH grant U54 CA100949.

References

Allison DB, Cui X, Page GP, Sabripour M. Microarray Data Analysis: from Disarray to Consolidation and Consensus. Nature Reviews|Genetics 2006a;7:55–65.

Allison DB, Gadbury GL, Heo M, Fernanfez JR, Lee C, Prolla TA, Weindruch R. A Mixture Model Approach for the Analysis of Microarray Gene Expression Data. Computational Statistics & Data Analysis 2002;39:1–20.

Allison, DB.; Page, GP.; Mark Beasley, T.; Edwards, JW. DNA Microarrays and Related Genomics Techniques, Design, Analysis, and Interpretation of Experiments. New York: Chapman & Hall/CRC; 2006b.

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Royal Statistical Society B 1995;57:289–300.

Broberg P. A New Estimate of the Proportion Unchanged Genes in a Microarray Experiment. Genome Biolog 2004;5:P10.

Donahue MJ. A Note on Information Seldom Reported via the P-values. The American Statistician 1999;53:303–306.

Efron B. Large-scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. Journal of the American Statistical Association 2004;99:96–104.

Efron B. Correlation and Large-scale Simultaneous Significance testing. Journal of the American Statistical Association 2007;102:93–103.

Gadbury GL, Page GP, Edwards J, Kayo T, Prolla TA, Weindruch R, Permana PA, Mountz JD, Allison DB. Power and sample size estimation in high dimensional biology. Statistical Methods in Medical Research 2004;13(4):325–338.

Gadbury GL, Page GP, Heo M, Mountz JD, Allison DB. Randomization Tests for Small Samples: An Application for Genetic Expression Data. Applied Statistics 2003;52(Part 3):365–376.

Gadbury GL, Xiang Q, Yang L, Barnes S, Page GP, Allison DB. Evaluating statistical methods using plasmode data sets in the age of massive public databases: An illustration using false discovery rates. Plos Genetics 2008;4(6):e1000098. [PubMed: 18566659]

Gupta, AK.; Nadarajah, S. Hand book of beta distribution and its applications. Marcel Dekker, Inc; New York: 2004.

Hu, X.; Gadbury, GL.; Xiang, Q. American Statistical Association Proceedings of the Biometrics Section. Alexandria, VA: American Statistical Association; 2005. Distributional Aspects of P-values and Their Uses in Multiple Testing Applications. [CD - ROM]

Hung HM, O'Neill RT, Bauser P, Köhne K. The Behavior of the P-values when the alternative hypothesis is true. Biometrics 1997;53:11–22.

Lee YH, Nair S, Rousseau E, Allison DB, Page GP, Tataranni PA, Bogardus C, Permana PA. Microarray Profiling of Isolated Abdominal Subcutaneous Adipocytes from Obese vs Non-Obese Pima Indians: Increased Expression of Inflammation-Related Genes. Diabetologia 2005;48:1432–0428.Epub

Liao JG, Lin Y, Selvanayagam ZE, Shih WJ. A mixture model for estimating the local false discovery rate in DNA microarray analysis. Bioinformatics 2004;20:2694–2701. [PubMed: 15145810]

Nettleton D, Hwang JTG, Caldo RA, Wise RP. Estimating the Number of True Null Hypotheses From a Histogram of P-values. Journal of Agricultural, Biological, and Environmental Statistics 2006;11:337–356.

Page GP, Edwards JW, Gadbury GL, Yelisetti P, Wang J, Trivedi P, Allison DB. The PowerAtlas: A Power and Sample Size Atlas for Microarray Experimental Design and Research. BMC Bioinformatics 2006;7:84. [PubMed: 16504070]

Parker RA, Rothenberg RB. Identifying important results from multiple statistical tests. Statistics in Medicine 1988;17:1031–1043.

Pollard KS, van der Laan MJ. Choice of a Null Distribution in Resampling-based Multiple Testing. Journal of Statistical Planning and Inference 2004;125:85–100.

Pounds S, Cheng C. Improving false discovery rate estimation. Bioinformatics 2004;20:1737–1745. [PubMed: 14988112]

Pounds S, Cheng C. Robust Estimation of the False Discovery Rate. Bioinformatics 2006;22:1979–1987. [PubMed: 16777905]

Pounds S, Morris SW. Estimating the occurrence of false positive and false negative in microarry studies by approximating and partitioning the empirical distribution of P-values. Bioinformatics 2003;19(10):1236–1242. [PubMed: 12835267]

Osier MV, Zhao H, Cheung KH. Handling multiple testing while interpreting microarrays with the gene ontology database. BMC Bioinformatics 2004;5:124. [PubMed: 15350198]

Owen AB. Variance of the Number of False Discoveries. Journal of the Royal Statistical Society 2005;Series B 67:411–426.

Qiu X, Klebanov L, Yakovlev A. Correlation Between Gene Expression Levels and Limitations of the Empirical Bayes Methodology in Microarray Data Analysis. Statistical Applications in Genetics and Molecular Biology 2005;4 [PubMed: 16646853]paper 34

Qiu X, Xiao Y, Gordon A, Yakovlev A. Assessing Stability of Gene Selection in Microarray Data Analysis. BMC Bioinformatics 2006;7:50. [PubMed: 16451725]

Ruppert D. What Is Kurtosis?: An Influence Function Approach. The American Statistician 1987;41:1–5.

Sackrowitz H, Samuel-Cahn EP. P Values as Random Variables-Expected P Values. The American Statistician 1999;53:326–331.

Schweder T, Spjøtvoll E. Plots of P-values to Evaluate Many Tests Simultaneously. Biometrika 1982;69:493–502.

Storey JD. A direct approach to false discovery rates. Journal of Royal Statistical Society Series B 2002;64:479–498.

Trivedi P, Edwards JW, Wang J, Gadbury GL, Srinivasasainagendra V, Zakharkin SO, Kim K, Mehta T, Brand JPL, Patki A, Page GP, Allison DB. HDBStat!: A platform-independent software suite for statistical analysis of high dimensional biology data. BMC Bioinformatics 2005;6:86. [PubMed: 15813968]

Travers SE, Smith MD, Bai JF, Hulbert SH, Leach JE, Schnable PS, Knapp AK, Milliken GA, Fay PA, Saleh A, Garrett KA. Ecological Genomics: Making the leap from model systems in the lab to native populations in the field. Frontiers in Ecology and the Environment 2007;5:19–24.

Westfall, PH.; Young, SS. Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment. New York: John Wiley & Sons; 1993.

Xiang Q, Edwards JW, Gadbury GL. Interval estimation in a finite mixture model: modeling P-values in multiple testing applications. Computational Statistics & Data Analysis 2006;52:570–586.

Yang JJ, Yang MCK. An Improved Procedure for Gene Selection from Microarray Experiments Using False Discovery Rate Criterion. BMC Bioinformatics 2006;7:15. [PubMed: 16405735]

Zhang H, Hyde K, Page GP, Brand JPL, Zhou J, Yu S, Allison DB, Hsu H, Mountz JD. Novel tumor necrosis factor α-regulated genes in rheumatoid Arthritis. Arthritis & Rheumatism 2004;50(2): 420–431.

Appendix A

Table A1

Second and fourth moments of the t-distribution (t_v) and chosen logistic (LOG) and double exponential (DE) distributions for each data set (A) – (F) where the corresponding degrees of freedom, v, is determined by the sample sizes for each data set. The second moment is the variance and is the same for all distributions. The entries in the table are the fourth moments. The fourth moment does not exist for a t-distribution with $v = 4$.

Data set,v ***Variance***	**(A),6** **1.5**	**(B),6** **1.5**	**(C),4** **2**	**(D),8** **1.33**	**(E),18** **1.125**	**(F),36** **1.059**
t_v	13.5	13.5	NA	8	4.34	3.57
LOG	9.45	9.45	16.8	7.47	5.32	4.71
DE	13.5	13.5	24	10.67	7.59	6.73

APPENDIX B

Derivation of Equation (7):

$$\begin{aligned} F_P(p|H_1) &= prob(P \le p|H_1) \\ &= prob(1 - G_T(T, n-1) \le p/2|H_1) + prob(G_T(T, n-1) \le p/2|H_1) \\ &= 1 - G_T(G_T^{-1}(1 - p/2, n-1), n-1, \sqrt{n}\delta) + G_T(G_T^{-1}(p/2, n-1), n-1, \sqrt{n}\delta) \end{aligned}$$

$$\begin{aligned} f_P(p|H_1) &= \frac{\partial}{\partial p} F_P(p|H_1) \\ &= \frac{1}{2}\left(\frac{g_T(G_T^{-1}(1-p/2,n-1),n-1,\delta\sqrt{n})}{g_T(G_T^{-1}(1-p/2,n-1),n-1)} + \frac{g_T(G_T^{-1}(p/2),n-1,\delta\sqrt{n})}{g_T(G_T^{-1}(p/2,n-1),n-1)}\right) \end{aligned}$$

APPENDIX C

C.1 The Moment Method

Given the density of a P-value, $g_P(p)$, the first two moments of the distribution are obtained by, $E_P(p) = \int_0^1 p g_P(p) dp$, and $\sigma_P(p) = E_P(p^2) - (E_P(p))^2$, where $E_P(p^2) = \int_0^1 p^2 g_P(p) dp$. The integrals for the two moments were obtained using the integrate function implemented in R (www.r-project.org). The parameters for the standard beta distribution given by equation (9) are obtained by matching these first two moments, $r = \frac{(1 - E(p))(E(p))^2 - E(p)\sigma^2(p)}{\sigma^2(p)}$ and $s = \frac{(1 - E(p))^2(E(p)) - (1 - E(p))\sigma^2(p)}{\sigma^2(p)}$ to the first two sample moments in the distribution of P-values.

C.2 The Least Squares Method

Suppose the density of a P-value distribution is $g_P(p)$ and a density function of a standard beta distribution is $\beta(p; r, s)$. Suppose there are k P-values $p_1, p_2, \cdots, p_k$. Consider the function: $D(P, r, s) = \sum_{i=1}^{k} (g_P(p_i) - \beta(p_i, r, s))^2$. Parameters (r, s) for given a set of P-values

were computed by minimizing the above function using the numerical optimizer “optim” implemented in R with k = 999.

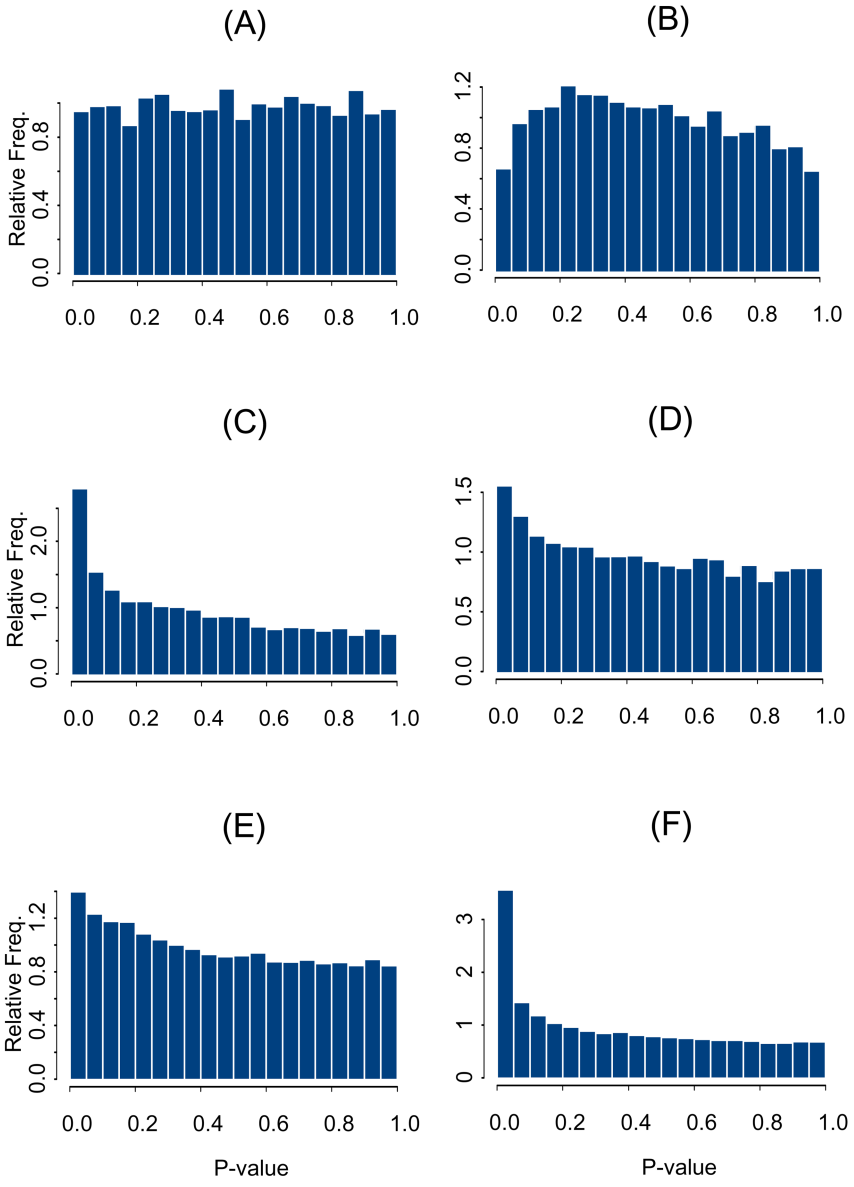

Figure 1.
Six relative frequency histograms of P-values.

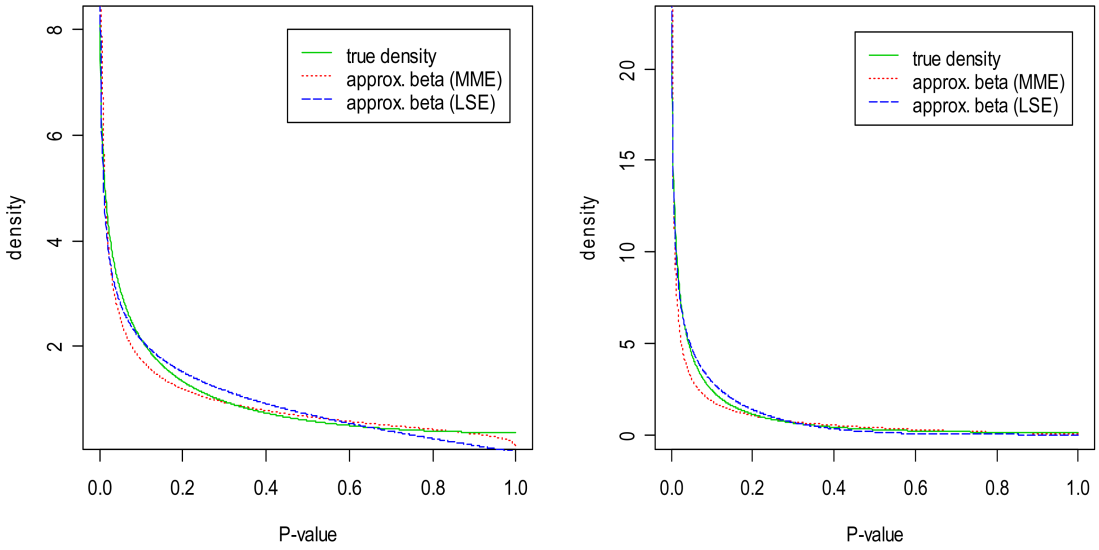

Figure 2.

The density of a P-value given by equation (8) compared with a beta distribution given by (9) that is computed using method of moments (MME) and the least squares minimization (LSE).

Table 1

Actual numbers, N, of P-values below the threshold τ for the six P-value distributions shown in Figure 1, (A) – (F). Data sets (A) and (B) have a total of 10,000 P-values, data sets (C) and (D) have 12,625, data set (E) has 31,042, and data set (F) has 63,149.

Data set	(A)	(B)	(C)	(D)	(E)	(F)
$\tau = 0.05$	488	346	1797	999	2197	11,386
$\tau = 0.01$	89	60	706	202	478	5788
$\tau = 0.001$	8	3	186	28	68	2642

Table 2

The expected value of N^0, i.e., $E(N^0)$, rounded to the nearest integer, for each of the six data sets if the global null hypothesis is true. The first entry in each cell is for the correct reference distribution, that is, the t-distribution with the appropriate degrees of freedom (*DF*) for that data set. The second entry is for the logistic distribution and the third entry for the double exponential. Parameter values for these latter distributions were chosen so that the first three moments matched those of the t-distribution. Data sets (A) and (B) have a total of 10,000 P-values, data sets (C) and (D) have 12,625, data set (E) has 31,042, and data set (F) has 63,149.

Data set **DF**	**(A)** **6**	**(B)** **6**	**(C)** **4**	**(D)** **8**	**(E)** **18**	**(F)** **36**
$\tau = 0.05$	500 520 593	500 520 593	631 698 786	631 657 749	1552 1663 1885	3157 3442 3889
$\tau = 0.01$	100 82 138	100 82 138	126 69 126	126 129 207	310 449 669	631 1038 1504
$\tau = 0.001$	10 3 10	10 3 10	13 0 2	13 9 26	31 76 166	63 228 459

Table 3

Expected number, $E(N^0)$, of P-values below a threshold τ and the standard deviation, $\sigma_{N^0} = \sqrt{Var(N^0)}$, of the number as a function of correlation, ρ, of P-values within m blocks of size $b = 100$ or 500 (plus the remainder block of size $r = K - bm$). Data are shown for the numbers of P-value in the distributions in Figure 1(A) – (F), under an assumption that the global null hypothesis is true. Data sets (A) and (B) have 10,000 P-values, data sets (C) and (D) have 12,625, data set (E) has 31,042, and data set (F) has 63,149.

			$\sqrt{Var(N^0)}$					
			$\rho = 0.0$		$\rho = 0.4$		$\rho = 0.8$	
Data Set	τ	$E(N^0)$	$b = 100$	500	$b = 100$	500	$b = 100$	500
(A) & (B)	0.05	500	22	22	69	147	140	312
	0.01	100	9	9	25	52	58	128
	0.001	10	3	3	5	11	16	35
(C) & (D)	0.05	631	24	24	77	165	158	350
	0.01	126	11	11	28	58	65	144
	0.001	13	4	4	6	12	18	39
(E)	0.05	1552	38	38	121	260	247	550
	0.01	310	18	18	44	92	102	226
	0.001	31	6	6	10	19	28	61
(F)	0.05	3157	55	55	172	370	353	784
	0.01	631	25	25	62	131	145	322
	0.001	63	8	8	14	27	39	87