Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 935621, 10 pages
doi:10.1155/2012/935621

Research Article

Sample Size Growth with an Increasing Number of Comparisons

Chi-Hong Tseng[1] and Yongzhao Shao[2]

[1] *Department of Medicine, UCLA School of Medicine, Los Angeles, CA 90095, USA*
[2] *Division of Biostatistics, NYU School of Medicine, New York, NY 10016, USA*

Correspondence should be addressed to Chi-Hong Tseng, tseng.ch@gmail.com

Received 30 March 2012; Accepted 8 June 2012

Academic Editor: Wei T. Pan

An appropriate sample size is crucial for the success of many studies that involve a large number of comparisons. Sample size formulas for testing multiple hypotheses are provided in this paper. They can be used to determine the sample sizes required to provide adequate power while controlling familywise error rate or false discovery rate, to derive the growth rate of sample size with respect to an increasing number of comparisons or decrease in effect size, and to assess reliability of study designs. It is demonstrated that practical sample sizes can often be achieved even when adjustments for a large number of comparisons are made as in many genomewide studies.

1. Introduction

With the recent advancement in high-throughput technologies, simultaneous testing of a large number of hypotheses has become a common practice for many types of genomewide studies. Examples include genetic association studies and DNA microarray studies. In a genomewide association analysis, a large number of genetic markers are tested for association with the disease [1]. In DNA microarray studies, the interest is typically to identify differentially expressed genes between patient groups among a large number of candidate genes [2].

The challenges for designing such large-scale studies include the selection of features of scientific importance to be investigated, selection of appropriate sample size to provide adequate power, and choices of methods appropriate for the adjustment of multiple testing [3–7]. There exist recent methodological breakthroughs on multiple comparisons, such as in the frontier of controlling the false discovery rate (FDR) [8, 9], which is particularly useful for the study of DNA microarray and protein arrays. It is also increasingly used in genomewide association studies [10]. On the other hand, the Bonferroni type adjustment is still surprisingly useful. For example, Klein et al. [1] successfully identified two SNPs which are associated with the age-related macular degeneration disease (AMD) using a Bonferroni adjustment. Witte et al. [11] provided an interesting observation that the relative sample size, based on Bonferroni adjustment, is approximately in a linear relationship to the logarithm of the number of comparisons.

An appropriate sample size is crucial for the success of studies involving a large number of comparisons. However, optimal and reliable sample size is extremely challenging to identify, as it typically depends on other design parameters that often have to be estimated based on preliminary data. Preliminary data are often limited at the design stage of studies, which lead to unreliable estimates of design parameters and create extra uncertainty in sample size estimation. Thus, it is of great practical interest to examine the relationship between sample size and other design parameters, such as the number of comparisons to be made. In this paper, we analyze this problem beyond witte et al.'s [11] observation by providing explicit sample size formulas, examining various genomic analyses, and deriving sample size formula for FDR control. The explicit sample size formulas are desirable because they elucidates how the change in other design parameters would affect sample size. This is of fundamental importance for understanding the reliability of study designs.

2. Sample Size Formulas

For testing a single hypothesis, the sample size problem is typically formulated as finding the number of subjects needed to ensure desired power $1-\beta$ for detecting an effect size Δ at a prespecified significance level α. Consider an one-sided test for equality of two normal means assuming known variances σ_1^2 and σ_2^2, respectively. The sample size per group (n) is as follows [12]:

$$n = \frac{\left(z_\alpha + C z_\beta\right)^2}{\Delta^2}, \tag{2.1}$$

where $\Delta = |\mu_1 - \mu_2| / \sqrt{\sigma_1^2 + \sigma_2^2}$, $C = 1$, $\Phi(z_t) = 1 - t$, and $\Phi(z)$ is the distribution function (CDF) of the standard normal distribution.

Many of the most widely used statistical tests have similar sample size formulas as in (2.1). For example, the commonly used Mann-Whitney test for comparing two continuous distributions without normality assumption has the same form of sample size formula as in (2.1). Similarly, for testing equality of two binomial proportions, using independent samples or using correlated samples as in McNemar's test, the sample size formulas are also of form (2.1) as discussed in Rosner [12].

For testing a single hypothesis, the influences of α, β, and Δ on the sample size n can be inferred easily from the above sample size formula (2.1), and are well known. When testing multiple hypotheses, one must guard against an abundance of false-positive results. The traditional criterion for error control in such situations is the familywise error rate (FWER), which is the probability of rejecting one or more true null hypotheses. The simplest and most commonly used method for controlling FWER is the Bonferroni correction, which is discussed in the next subsection.

2.1. FWER Control

In this section, we present sample size formulas for multiple comparisons in the context of controlling the familywise error rate (FWER). Suppose we make multiple comparisons with Δ being the same. If we wish to retain a familywise error rate α, and power $(1-\beta)$, then with the Bonferroni adjustment, $\alpha_{\text{bon}} = \alpha/M$, the sample size corresponding to (2.1) becomes

$$n_M = \frac{\left(z_{\alpha/M} + Cz_{\beta}\right)^2}{\Delta^2}. \tag{2.2}$$

To see how n_M changes as M increases, we can use the following well-known fact: when $\alpha < 0.5$, $\phi(z_\alpha)(1/z_\alpha - 1/z_\alpha^3) \leq 1 - \Phi(z_\alpha) \leq \phi(z_\alpha)/z_\alpha$. Since $\alpha/M = 1 - \Phi(z_{\alpha/M})$, we can approximate $z_{\alpha/M}$ by $z^*_{\alpha/M}$, where

$$z^{*2}_{\alpha/M} \equiv 2\log\left(\frac{M}{\alpha}\right) - \log(2\pi)\log\log\left(\frac{M}{\alpha}\right). \tag{2.3}$$

The explicit approximation of $z^2_{\alpha/M}$ in (2.3) works extremely well for M ranging from 10 to 10^{10}. Putting (2.3) into (2.2) yields the following approximation of the required sample size n_M:

$$n^*_M = \frac{\left(z^*_{\alpha/M} + Cz_{\beta}\right)^2}{\Delta^2}. \tag{2.4}$$

Then, for fixed (α, β, Δ), from (2.3) and (2.4), we have

$$n_M \approx n^*_M \approx \frac{2}{\Delta^2}\log\frac{M}{\alpha}, \quad \text{as } M \longrightarrow +\infty. \tag{2.5}$$

A few facts are self-evident from the above approximation. First, n_M is an approximately linear function of $\log M$ (base 10) with slope $2/\Delta^2$. Second, the impact of β on n_M (or n^*_M) is negligible when M is large. Third, a decrease in α is equivalent to an increase in M on n_M (or n^*_M). The impact of Δ on n_M (or n^*_M) is demonstrated in Figure 1 with $\alpha = 0.05$, $1-\beta = 0.90$, and $\Delta = 0.5, 1$, and 2, respectively. It shows that n_M (open circles) can indeed be approximated well by a linear function of $\log M$. The lines are calculated based on approximate normal quantiles (2.4) for n^*_M. Moreover, when Δ is large (e.g., $\Delta = 2$), the slope is very small.

The simple Bonferroni correction is very useful, when the number of true alternatives is small. This often occurs, for example, in candidate gene association studies. The Bonferroni approach is easy to apply, for example, it is convenient when the hypotheses involve many covariates and nuisance parameters, whereas the permutation approaches may not be applicable, because they require some symmetry or exchangeability on the null hypotheses [13, 14]. Next, we give two practical examples to illustrate the growth rate of sample size relative to the number of tests M to be performed.

Table 1: An SNP from Klein et al. [1].

Attribute	rs1329428 (C/T)
Risk allele	C
OR (dominant)	4.7
Freq in HapMAP CEU	82%
OR (recessive)	6.2
Freq in HapMAP CEU	41%

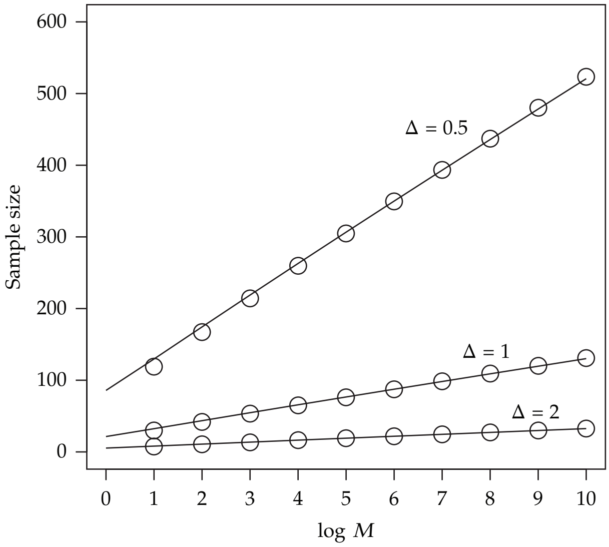


Figure 1: Sample size versus log M (base 10) to detect effect sizes $\Delta = 0.5, 1$ or 2 with $1 - \beta = 90\%$ power at the familywise significance level $\alpha = 5\%$, when Bonferroni adjustment is used. The open circles represent the sample sizes calculated based on exact normal quantiles (2.2).

The AMD Example

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. Klein et al. [1] reported a genomewide screen of 96 cases and 50 controls for polymorphisms associated with AMD. They examined 116,204 single-nucleotide polymorphisms (SNPs). Two of the SNPs are found to be strongly associated with the disease phenotype. This is an example to test equality of two binomial proportions of two independent groups (cases and controls). The required sample size for each marker is given in (2.2) or (2.4) with $\Delta^2 = 2(p_1 - p_2)^2 \overline{p}\,\overline{q}$, $C = \sqrt{(p_1\, q_1 + p_2 q_2)/(2\overline{p}\,\overline{q})}$, and $\overline{p} = (p_1 + p_2)/2$. Illustration for sample size growth with the Bonferroni correction is plotted in Figure 2 against log M using the SNP rs1329428 (Table 1) identified in Klein et al. [1]. Using Bonferroni adjustment, the sample sizes are calculated to provide 90% power to detect the association at the familywise significance level $\alpha = 5\%$. The open circles and plus signs are sample sizes n_M using (2.2) according to the dominant and recessive odds ratios, respectively. The corresponding lines are sample sizes n_M^* based on (2.4).

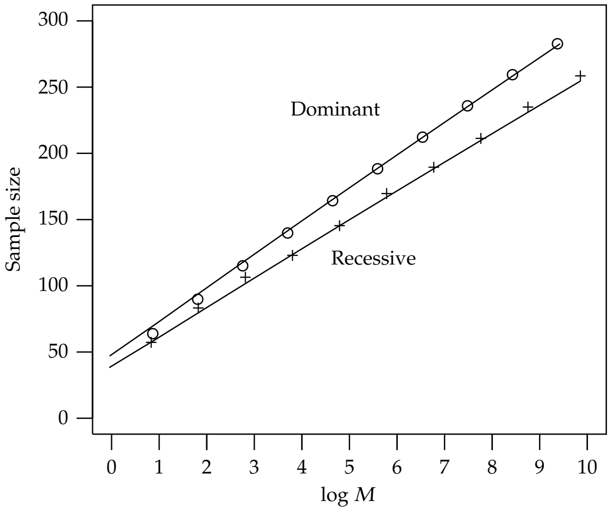


Figure 2: Sample sizes to detect the association at rs1329428 versus numbers of SNPS in genome wide screen of the AMD study.

The TDT Example

To test for linkage or association in family-based studies, the transmission/disequilibrium test (TDT) of Spielman et al. [15] examines the transmission of an allele from heterozygous parents to their affected offspring. If an allele is associated with the disease risk, its transmission may occur more than 50% of the times. Risch and Merikangas [16] studied the required sample size for TDT in affected sib pairs. TDT is equivalent to McNemar's test for two correlated proportions with the hypothesis $H_0 : p = 0.5$ versus $H_1 : p > 0.5$, for the specified alternative $p = p_A$, where p_A is the probability that an A/B parent transmits allele A to an affected offspring. The sample size (matched pairs) needed is given in (2.1) with $C = 2\sqrt{p_A(1 - p_A)}$, $\Delta^2 = 2(p_A - 0.5)^2 p_D$, and p_D is the projected proportion of discordant pairs among all matched pairs. If we assume that each family used in the analysis has only one marker heterozygous parent, then n is the number of families required. Demonstration of sample sizes for TDT is plotted in Figure 3 using the setup given in Risch and Merikangas [16]. Using Bonferroni adjustment, the sample sizes are calculated to provide $1 - \beta = 90\%$ power to identify a disease gene at the familywise significance level $\alpha = 5\%$. The plus signs and open triangles are the sample size n_M calculated based on (2.2) corresponding to disease frequencies equal to 0.1 and 0.5, respectively. The corresponding lines are for n_M^* based on (2.4).

2.2. FDR Control

For the test of multiple hypotheses, such as the analysis of many genes using microarray, the outcomes can be described in Table 2.

It is likely that many genes are differentially expressed in a microarray study [7]. A natural way to control the overall false positives is to control the expected proportion of false

Table 2: Possible outcome for testing M hypotheses.

Truth	Test decision		Total
	Reject H_0	Accept H_0	
H_0	V	$m_0 - V$	m_0
H_1	U	$m_1 - U$	m_1
Total	R	$M - R$	M

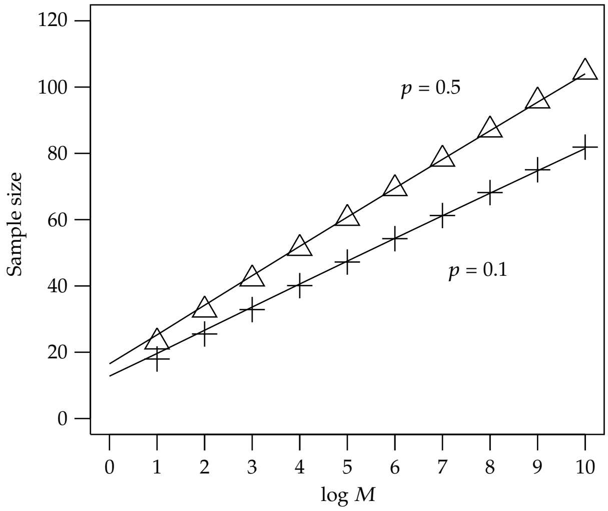


Figure 3: Number of families needed versus log M (base 10). Sample size for the TDT in the example of Risch and Merikangas [16], with disease frequencies of 0.1 (plus signs) and 0.5 (open triangles).

positives. Benjamini and Hochberg [8] defined the false discovery rate (FDR), using Table 2, as

$$\text{FDR} = P(R > 0)E\left[\frac{V}{U} \mid R > 0\right], \quad \text{FDR} = 0 \text{ for } R = 0. \tag{2.6}$$

Storey [9] defines positive FDR (pFDR) as pFDR = FDR$/P(R > 0)$. When M is large as assumed next, $P(R > 0) \approx 1$, unless the power $1 - \beta$ is too small, then FDR $\approx$ pFDR.

The required sample size for multiple testing depends on α, $(1 - \beta)$, M, and Δ of each individual gene. For easy exposition, we assume an equal effect size Δ for all differentially expressed genes, say m_1 genes; thus, the power $(1 - \beta)$ of detecting any individual differentially expressed gene is the same for all of the m_1 genes between samples of two conditions of sizes n_1 and n_2. The expected outcomes in multiple testing can be expressed as functions of α, β, m_0, and m_1 and are summarized in Table 3.

By law of large numbers, from Table 3, FDR $= E(V/R) = m_0\alpha/(m_0\alpha + m_1(1 - \beta))$. Denote the desired FDR level by f. Then from the above equation, we have

$$\alpha_{\text{fdr}} = \frac{f}{1-f}\left[\left(1 - \frac{m_1}{M}\right)^{-1} - 1\right](1 - \beta). \tag{2.7}$$

Table 3: Expected outcome for testing M hypotheses.

Truth	Test decision		Total
	Reject H_0	Reject H_a	
H_0	αm_0	$(1-\alpha)m_0$	m_0
H_1	$(1-\beta)m_1$	βm_1	m_1
Total	$\alpha m_0 + (1-\beta)m_1$	$(1-\alpha)m_0 + \beta m_1$	M

To account for the dependence among tests, we follow Shao and Tseng [17]. Let T_i be the test statistic of an one-sided two sample z-test for the ith alternative hypothesis, let p_i be its P value, and let $u_i = I(p_i < \alpha)$ be the rejection status at the level α; $u_i = 1$ if the ith test result is a rejection and 0 otherwise. Furthermore, if we denote the pairwise correlation coefficient between two tests by $\rho_U^{ij} = \operatorname{Corr}(T_i, T_j)$, then it can be shown that the correlation between u_i and u_j, $\theta_U^{ij} = \operatorname{Corr}(u_i, u_j)$ can be derived from the correlations of test statistics as follows:

$$\theta_U^{ij} = \frac{F\left(\tilde{z}_\alpha, \tilde{z}_\alpha; \rho_U^{ij}\right) - (1-\beta)^2}{\beta(1-\beta)}, \tag{2.8}$$

where F is the CDF of the standard bivariate normal distribution, and $\tilde{z}_\alpha = -z_\alpha + \Delta / \sqrt{n_1^{-1} + n_2^{-1}}$ [18]. Under local dependence assumptions, the total number of true discoveries, $U = \sum_{i=1}^{m_1} u_i$, has an approximately normal distribution: $U \sim N(m_1(1-\beta), \sigma_U^2)$, where $\sigma_U^2 = m_1\beta(1-\beta)[1+\overline{\theta}_U(m_1-1)]$, and $\overline{\theta}_U = (m_1(m_1-1))^{-1}\sum_{i\neq j}\theta_U^{ij}$ is the average correlation among true discoveries. The local dependence assumption can be viewed in a simplified formulation of the central limit theorem under the "strong mixing" given in Theorem 27.4 of Billingsely [19]. "Mixing" means, roughly, that random variables temporally far apart from one another are nearly independent. We think that the local dependence assumption is reasonable in many genetic studies. For example, linkage disequilibrium can result in local dependence of genetic markers. In biomarkers study, biomarkers of the same pathway are often correlated and result in local dependence.

It is often desirable to find sample size to ensure a familywise power Ψ of identifying at least a given fraction $r \in (0,1)$ out of m_1 true discoveries: $\Psi = P(U \geq [m_1 r])$. The above normal approximation of U allows a closed form solution for the comparison-wise β:

$$\beta_{\text{fdr}} = 1 - r - \frac{1 - 2r + \sqrt{4m_1^* r(1-r) + 1}}{2m_1^* + 2}, \tag{2.9}$$

where $m_1^* = m_1 / \{[1+\overline{\theta}_U(m_1-1)]z_{1-\Psi}^2\}$. When m_1 is large, to have a family-wise power Ψ in detecting at least $100r\%$ out of m_1 true alternatives, and with an FDR f, the sample size needed for a one-sided z-test is given by (2.1), with α and β determined by (2.7) and (2.9) iteratively.

A Microarray Example.

We now consider a well-known dataset from a study of leukemia in Gloub et al. [2] to demonstrate the relationship between sample size and number of multiple comparisons when

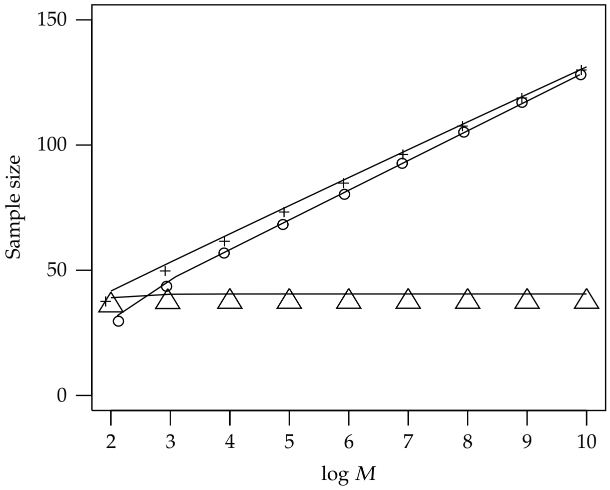


Figure 4: Sample size versus $\log M$ (base 10) for controlling FDR $f = 5\%$ with $\Psi = 90\%$. The open circles represent the sample sizes needed when the number of true alternatives m_1 stays as constant ($m_1 = 40$), the plus signs give the sample sizes when $m_1 = 2 \log M$, and the triangles are the sample sizes when the proportion of true alternatives is constant ($m_1 = M/10$).

controlling FDR. The original purpose of the experiment described in Gloub et al. [2] is to identify the susceptible genes related to clinical heterogeneity in two subclass of leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The dataset contains 7129 attributes from 47 patients with ALL and 25 patients with AML. We can apply (2.1), (2.7), and (2.9) iteratively to obtain the required sample size when controlling FDR. Figure 4 provides 3 different settings for controlling FDR $f = 5\%$ with $\Psi = 90\%$. Based on the top 100 most differentially expressed genes in Gloub et al. [2], $\overline{\theta}_U = 0.07$ (see (2.9)). The open circles represent the sample sizes n_M needed when the number of true alternatives m_1 stays constant ($m_1 = 40$). In this case, we observe that the sample size is a linear function of $\log M$ as M increases. The "plus" signs denote the sample sizes n_M when the number of true alternatives increases in a slower pace than M ($m_1 = 2 \log M$); the sample size is also approximately a linear function of $\log M$. The triangles denote the sample sizes n_M when the proportion of true alternatives is constant ($m_1/M = 10\%$), and the sample sizes roughly remain constant as the number of tests increases which is expected from (2.7). The lines in Figure 4 represent sample sizes n_M^* based on (2.4).

3. Discussion

In this short paper, we have shown that a large increase in the number of comparisons often only requires a small increase in the sample size. We further demonstrated that when controlling FDR, the sample size may even sometimes stay constant as the number of comparisons increases (Figure 4). The sample size required for testing M hypotheses is generally not growing faster than a linear function of $\log M$, even when a simple Bonferroni adjustment is used, and the slope of the linear growth rate (in $\log M$) is small when detecting a large effect size. These results have important implications in practice due to the wide use of multiple comparisons.

In this paper, we discuss the sample size formulas based on fixed effect size in alternative hypotheses. In reality, the effect sizes may follow a distribution, and simulation method may be useful in determining the sample size. We used z-test to derive the sample size formula, because large sample size is usually required for studies with multiple comparisons. If the effect size is large and sample size is small, t-test may be more appropriate. However, we expect the relationship between sample size and the logarithm of number of comparisons made is still linear.

In practice, if feasible, using a conservative sample size can reduce the chance of obtaining false-positive results and ensure reproducibility [6]. The simple sample size formulas provided in this paper might be used to select a suitable sample size by varying other design parameters and by taking into consideration the reliability of the proposed designs. While FDR is very useful and is increasingly used in multiple comparisons, our experience in helping biomedical investigators and the analysis in this paper indicate that the simple Bonferroni approach can often provide conservative but useful sample sizes in many situations.

References

[1] R. J. Klein, C. Zeiss, E. Y. Chew et al., “Complement factor H polymorphism in age-related macular degeneration,” *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[2] T. R. Golub, D. K. Slonim, P. Tamayo et al., “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] P. H. Westfall and S. S. Young, *Resample-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, John Wiley & Sons, New York, NY, USA, 1993.

[4] J. C. Hsu, *Multiple Comparisons: Theory and Methods*, Chapman & Hall, London, UK, 1996.

[5] P. H. Westfall and R. D. Wolfinger, “Multiple tests with discrete distributions,” *American Statistician*, vol. 51, no. 1, pp. 3–8, 1997.

[6] N. J. Risch, “Searching for genetic determinants in the new millennium,” *Nature*, vol. 405, no. 6788, pp. 847–856, 2000.

[7] J. D. Storey and R. Tibshirani, “Statistical significance for genomewide studies,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[8] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society, Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[9] J. D. Storey, “A direct approach to false discovery rates,” *Journal of the Royal Statistical Society, Series B*, vol. 64, no. 3, pp. 479–498, 2002.

[10] Q. Yang, J. Cui, I. Chazaro, L. A. Cupples, and S. Demissie, “Power and type I error rate of fales discovery rate approaches in genome-wide association studies,” *BMC Genetics*, vol. 6, supplement 1, article S134, 2005.

[11] J. S. Witte, R. C. Elston, and L. R. Cardon, “On the relative sample size required for multiple comparisons,” *Statistics in Medicine*, vol. 19, no. 3, pp. 369–372, 2000.

[12] B. Rosner, *Fundamentals of Biostatistics*, Duxbury, Los Angeles, Calif, USA, 2006.

[13] Y. Ge, S. Dudoit, and T. P. Speed, “Resampling-based multiple testing for microarray data analysis,” *Test*, vol. 12, no. 1, pp. 1–77, 2003.

[14] Y. Huang, H. Xu, V. Calian, and J. C. Hsu, “To permute or not to permute,” *Bioinformatics*, vol. 22, no. 18, pp. 2244–2248, 2006.

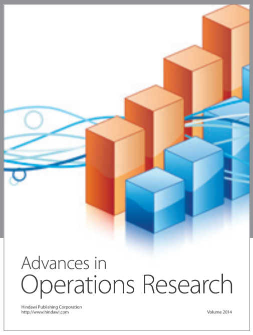
[15] R. S. Spielman, R. E. McGinnis, and W. J. Ewens, “Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM),” *American Journal of Human Genetics*, vol. 52, no. 3, pp. 506–516, 1993.

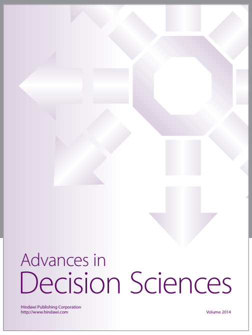
[16] N. Risch and K. Merikangas, “The future of genetic studies of complex human diseases,” *Science*, vol. 273, no. 5281, pp. 1516–1517, 1996.

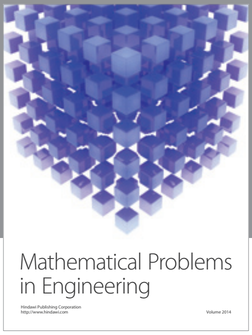
[17] Y. Shao and C. H. Tseng, “Sample size calculation with dependence adjustment for FDR-control in microarray studies,” *Statistics in Medicine*, vol. 26, no. 23, pp. 4219–4237, 2007.

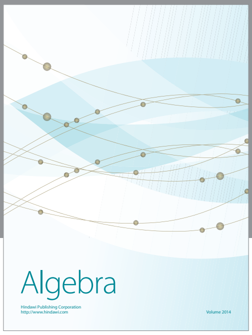
[18] H. Ahn and J. J. Chen, “Generation of over-dispersed and under-dispersed binomial variates,” *Journal of Computational and Graphical Statistics*, vol. 4, no. 1, pp. 55–64, 1995.

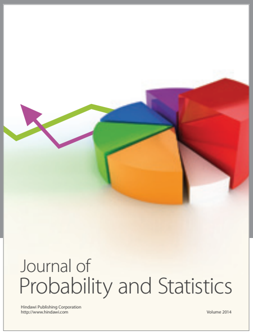
[19] P. Billingsley, *Probability and Measure*, John Wiley & Sons, New York, NY, USA, 1995.


Advances in
Operations Research

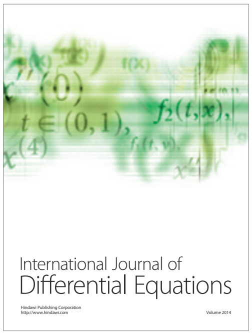

Advances in
Decision Sciences

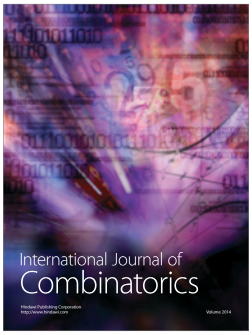

Mathematical Problems
in Engineering

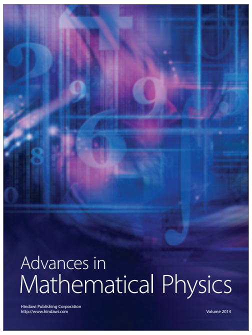

Algebra

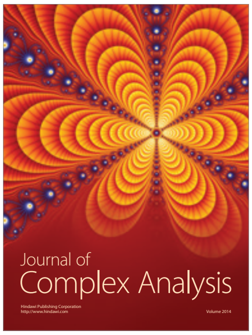

Journal of
Probability and Statistics

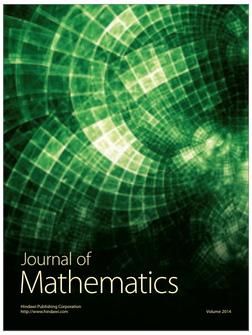
The Scientific
World Journal

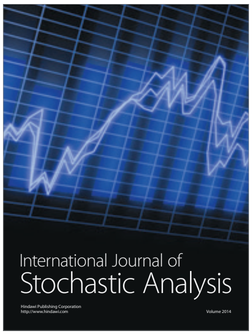

International Journal of
Differential Equations

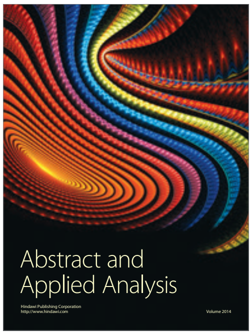

International Journal of
Combinatorics

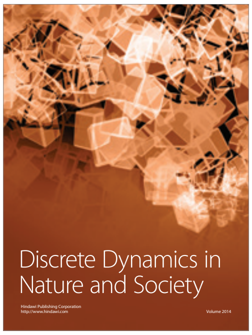

Advances in
Mathematical Physics

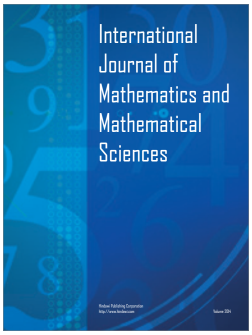

Journal of
Complex Analysis


Journal of
Mathematics


International Journal of
Stochastic Analysis

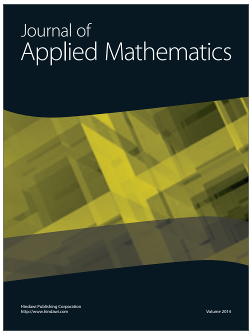

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


Journal of
Applied Mathematics

Journal of
Optimization