A Univariate Statistical Parameter Assessing Effect Size for Multivariate Responses

Xiaohua Douglas Zhang
Early Development Statistics – Asian Pacific, BARDS, Merck Research Laboratories,
Beijing, China. Email: xiaohua_zhang@merck.com

Abstract

The statistical significance has been intensively criticized in medical and social sciences because of many issues that it has. Effect sizes have been proposed as an alternative to statistical significance. Recently, strictly standardized mean difference (SSMD) has been proposed for the comparison of two groups with applications in a univariate-response setting. There is a need to extend this type of effect size from a univariate-response setting to a multivariate-response setting. In this paper, based on SSMD and Mahalanobis distance, I construct a novel parameter called dimension-adjusted squared Mahalanobis distance (DSMD). The concept of DSMD can be applicable to both univariate- and multivariate-response settings. Moreover, the criterion of DSMD to assess the differentiation between two groups can also be applicable to both univariate- and multivariate-response settings. Thus, DSMD may have the potency of being applicable to a variety of situations in medical and social sciences.

KEY WORDS: dimension-adjusted squared Mahalanobis distance, dimension-adjusted Mahalanobis distance, strictly standardized mean difference, d^{+}-probability, effect size.

1. Introduction

The widely used method for the comparison of two groups is statistical significance or *p*-value of *t*-test. However, the use of statistical significance for the comparison of two groups has been intensively criticized in medical and social sciences [1], leading to various effect sizes as alternatives [2-4]. Recently, strictly standardized mean difference (SSMD) has been proposed for the comparison of two groups [5-8]. The SSMD-based criterion for assessing effect size has a probabilistic basis. In practice, SSMD has been used for quality control and hit selection in high-throughput screening experiments [8-12]. However, SSMD is applicable only to a univariate-response setting, not to a multivariate-response setting yet.

Mahalanobis distance for the means between two groups with multivariate responses can serve as a parameter for measuring the degree of differentiation [13]. However, there are flaws for the use of Mahalanobis distance, two of which are, (1) its value increases as the number of dimensions increases even though the degree of differentiation in each dimension is the same, and (2) it lacks a criterion for using Mahalanobis distance to quantifying effect size. In this paper, based on the concept of Mahalanobis distance, I extend SSMD from a univariate-response setting to a multivariate-response setting and provide a criterion to quantify effect sizes which is derived from the SSMD-based criterion.

2. Parameters for quantifying group differentiation

2.1. In univariate settings

Suppose we are interested in the comparison of two groups with a single response. The first group has a distribution F_1 with mean μ_1 and variance σ_1^2 and the second group has a distribution F_2 with mean μ_2 and variance σ_2^2. The covariance between these two groups is σ_{12}. The magnitude of difference between two groups can be assessed by a parameter SSMD. Let a random variable *D* denote the difference between two random values from two groups respectively, i.e., $D = P_1 - P_2$. SSMD (denoted as β) is defined as the ratio of mean to standard deviation of the difference *D*, namely $\beta = \frac{\mu_D}{\sigma_D}$ where μ_D and σ_D are the mean and standard deviation of *D* respectively. Expressing SSMD as a formula of means, variances and covariance in the two groups, we get $\beta = \frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2+\sigma_2^2-2\sigma_{12}}}$.

SSMD has a direction of the mean difference. When we are only interested in distance but not direction, we may use SSMD absolute value which equals Mahalanobis distance between the two means in a univariate case, i.e., $|\beta| = \sqrt{\frac{(\mu_1-\mu_2)^2}{\sigma_D^2}}$. With this consideration, we may use the squared SSMD, namely $\beta^2 = \frac{\mu_D^2}{\sigma_D^2}$, which equals the squared Mahalanobis distance between two means in the two groups. The Mahalanobis distance between the two random variable *P₁* and *P₂* is $\sqrt{\frac{(P_1-P_2)^2}{\sigma_D^2}}$. Assume these two groups are normally distributed, then $D \sim N(\mu_1-\mu_2, \sigma_D^2)$ and $\frac{D}{\sigma_D} \sim N(\frac{\mu_1-\mu_2}{\sigma_D}, 1)$. Thus $\frac{D^2}{\sigma_D^2} \sim \chi^2(1, \beta^2)$ where β^2 is a non-central parameter of the χ^2-distribution.

2.2. In multivariate settings

Suppose we are interested in the comparison of two groups with multiple responses (i.e., two *k*-variate groups). Considering the two *k*-variate groups $\mathbf{P}_1$ (with sample size n_1, mean $\boldsymbol{\mu}_1$, covariance matrix $\boldsymbol{\Sigma}_1$) and groups $\mathbf{P}_2$ (with sample size n_2, mean $\boldsymbol{\mu}_2$, covariance matrix $\boldsymbol{\Sigma}_2$), where $\mathbf{P}_1 = \begin{pmatrix} P_{11} \\ \vdots \\ P_{1k} \end{pmatrix}$, $\mathbf{P}_2 = \begin{pmatrix} P_{21} \\ \vdots \\ P_{2k} \end{pmatrix}$, $\boldsymbol{\mu}_1 = \begin{pmatrix} \mu_{11} \\ \vdots \\ \mu_{1k} \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} \mu_{21} \\ \vdots \\ \mu_{2k} \end{pmatrix}$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} \sigma_{111}^2 & \cdots & \sigma_{11k} \\ \vdots & \ddots & \vdots \\ \sigma_{1k1} & \cdots & \sigma_{1kk}^2 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} \sigma_{211}^2 & \cdots & \sigma_{21k} \\ \vdots & \ddots & \vdots \\ \sigma_{2k1} & \cdots & \sigma_{2kk}^2 \end{pmatrix}, \ \boldsymbol{\Sigma}_{12} = \begin{pmatrix} \sigma_{1211} & \cdots & \sigma_{121k} \\ \vdots & \ddots & \vdots \\ \sigma_{12k1} & \cdots & \sigma_{12kk} \end{pmatrix}.$$

$\mathbf{\Sigma}_{12}$ is the covariance matrix between $\mathbf{P}_1$ and $\mathbf{P}_2$. Let $\mathbf{\Sigma} = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 - 2\mathbf{\Sigma}_{12}$. Denote the difference between the two groups as $\mathbf{D}$ (namely, $\mathbf{D} = \mathbf{P_1} - \mathbf{P_2}$), and its mean and variance as $\boldsymbol{\mu}_{\mathrm{D}}$ and $\mathbf{\Sigma}_{\mathrm{D}}$, respectively. Then $\boldsymbol{\mu}_{\mathrm{D}} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\mathbf{\Sigma}_{\mathrm{D}} = \mathbf{\Sigma}$.

For the comparison of two groups with multiple responses, we may want to use a single parameter to capture the separation between these two groups based on all responses. One way to address it is that we may explore each response separately as in the univariate case described in previous section and then pool the results of all individual responses to a single statistical parameter. It is nontrivial to integrate different directions in different responses. Thus, we may focus on non-direction parameters first.

For a single response (say the i^{th} response), the Mahalanobis distance between two means is a non-direction parameter $r_i = \sqrt{\frac{(\mu_{1i} - \mu_{2i})^2}{\sigma_{Di}^2}}$. Then, if the k responses are independent, we may pool the result of all k-responses by $r = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\frac{(\mu_{1i} - \mu_{2i})^2}{\sigma_{Di}^2}}$. The adjustment efficient k is needed to adjust the increment of a parameter purely due to the increment of the k value. The r can readily be extended to non-independent cases with the same form as $r = \sqrt{\frac{1}{k}\lambda}$ where $\lambda = \boldsymbol{\mu}_{\mathbf{D}}{}' \, \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_{\mathbf{D}}$ which is the non-central parameter of a squared Mahalanobis distance. The Mahalanobis distance between two values from groups $\mathbf{P}_1$ and $\mathbf{P}_2$ respectively is $\sqrt{(\mathbf{P_1} - \mathbf{P_2})' \, \mathbf{\Sigma}^{-1}(\mathbf{P_1} - \mathbf{P_2})}$. Correspondingly the squared Mahalanobis distance is $\mathrm{d}^2(\mathbf{P_1}, \mathbf{P_2}) = (\mathbf{P_1} - \mathbf{P_2})' \, \mathbf{\Sigma}^{-1}(\mathbf{P_1} - \mathbf{P_2})$. Assume the two groups are normally distributed, then $\mathbf{D} = \mathbf{P_1} - \mathbf{P_2} \sim \mathrm{N_k}(\boldsymbol{\mu}_{\mathbf{D}}, \mathbf{\Sigma})$ and $\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{D} \sim \mathrm{N_k}(\mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}_{\mathbf{D}}, \mathbf{I_k})$. Thus, the squared Mahalanobis distance between the two groups has a non-central chi-square distribution, that is, $\mathrm{d}^2(\mathbf{P_1}, \mathbf{P_2}) = \left(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{D}\right)' \left(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{D}\right) \sim \chi_k^2(\lambda)$, where $\chi_k^2(\lambda)$ is a non-central chi-squared distribution with degree of freedom k, non-central parameter λ, mean $k + \lambda$ and variance $2(k + 2\lambda)$.

2.3. Properties and interpretation of distance parameters

In the above section, I extend SSMD from a univariate-response setting to a multi-response setting, resulting in two new parameters, r and r^2. For convenience, we may call the new parameter r "dimension-adjusted Mahalanobis distance" (DMD) and its square r^2 "dimension-adjusted squared Mahalanobis distance" (DSMD) between two group means. That is, for DMD, we have

$$r = \sqrt{\frac{1}{k}\boldsymbol{\mu}_{\mathbf{D}}'\mathbf{\Sigma}_{\mathrm{D}}{}^{-1}\boldsymbol{\mu}_{\mathbf{D}}} = \sqrt{\frac{1}{k}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)} \, .$$

For DSMD, we have

$$r^2 = \frac{1}{k}\boldsymbol{\mu}_{\mathbf{D}}'\mathbf{\Sigma}_{\mathrm{D}}{}^{-1}\boldsymbol{\mu}_{\mathbf{D}} = \frac{1}{k}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

It is obvious that, when $k = 1$ (i.e., the univariate-response case), DMD becomes the absolute value of SSMD and DSMD become the squared value of SSMD. The properties and interpretation of the two distance parameters can be also extended from SSMD for the univariate-response setting as follows.

SSMD has two clear and meaningful interpretations when it is used for the comparison of two groups. The first interpretation is that it is the ratio of mean to standard deviation of a random variable representing the difference between two groups, and the second interpretation is that it reflects the probability that a random value from the first group is larger than a random value from the second group, namely d^+-probability [8]. When the data are normally distributed in both groups, d^+-probability = $\Phi(\beta)$, where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution. When the data are not normally distributed, there is still relationship between d^+-probability and β [8]. Because of clear and meaningful interpretations of SSMD, we can construct meaningful and interpretable SSMD-based criteria for classifying the magnitude of difference between two groups, namely $|\beta| \geq 1.645$ for "huge", $1.645 > |\beta| \geq 1$ for "large", $1 > |\beta| > 0.25$ for "medium", and $|\beta| \leq 0.25$ for "small" under normality assumption [8]. These SSMD-based criteria can be interpreted using probability as described in [7]. Equivalently, we can have DSMD-based criteria for classifying the degree of differentiation between two groups in a multivariate-response setting, namely $\text{DSMD} \geq 2.706$ for "huge", $2.706 > \text{DSMD} \geq 1$ for "large", $1 > \text{DSMD} > 0.0625$ for "medium", and $\text{DSMD} \leq 0.0625$ for "small".

3. Statistical inference of DSMD

When the two groups to be compared are independent, assume $\mathbf{x}_i$, $\mathrm{i} = 1, \ldots, n_1$, $\mathbf{y}_\mathrm{j}$, $\mathrm{j} = 1, \ldots, n_2$, are the samples from the two groups, respectively. Let $\bar{\boldsymbol{x}}, \mathbf{S}_\mathbf{x}^2$, and $\bar{\boldsymbol{y}}, \mathbf{S}_\mathbf{y}^2$ denote sample mean and sample covariance matrix for the two samples from the two groups $\mathbf{P}_1$ and $\mathbf{P}_2$, $\mathrm{S}_{\mathrm{xy}}^2$ denotes the sample covariance matrix between those samples. Let $N = n_1 + n_2$. The estimates of DSMD for independent and paired groups are given below, assuming that the two groups are normally distributed.

Consider the situation where the two groups are paired. Assume that $\mathbf{d}_i = \boldsymbol{x}_i - \mathbf{y}_i$, $\mathrm{i} = 1, \ldots, \mathrm{n}$ is the sample from $\mathbf{D}$ and that $\bar{d}, \mathrm{S}$ are sample mean vector and sample covariance matrix of $\mathbf{D}$, respectively. Then the estimate of r^2 is $\hat{r}^2 = \frac{1}{k}\bar{d}'S^{-1}\bar{d}$.

Consider the situation where the two groups are independent. When the covariance matrices of two groups are unequal, by applying the maximum likelihood estimates (MLE) of the mean and covariance matrix we can obtain the estimate of r^2. That is, $\hat{r}^2 = \frac{1}{k}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})'(\frac{n_1-1}{n_1}\boldsymbol{S}_x^2 + \frac{n_2-1}{n_2}\boldsymbol{S}_y^2)^{-1}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})$. When the covariance matrices of two independent groups are equal, we can pool the two groups to get the estimate of their covariance matrices. Let $\mathrm{S}_{\mathrm{pool}}^2$ be the pooled sample covariance matrix, then,

$$\hat{r}^2 = \frac{1}{k}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})'(\frac{2(n_1 + n_2 - 2)}{n_1 + n_2}\boldsymbol{S}_{\mathbf{pool}}^2)^{-1}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})$$

4. Conclusion

To address the need for quantifying effect size in a multivariate-response setting, here I derive a univariate parameter DSMD from a recently developed parameter SSMD.

Subsequently, I explore the estimation of DSMD. Moreover, I construct a DSMD-based criterion for classifying the degree of differentiation between two groups, namely $\text{DSMD} \geq 2.706$ for "huge", $2.706 > \text{DSMD} \geq 1$ for "large", $1 > \text{DSMD} > 0.0625$ for "medium", and $\text{DSMD} \leq 0.0625$ for "small" under normality assumption. DSMD and its criterion can be applied to uniformly to both univariate- and multivariate-response settings, thus having the potency of being applicable to a variety of situations in medical and social sciences.

References

1. Harlow LL, Mulaik SA, Steiger JH: What if there were no significance tests? Mahwah, NJ: Lawrence Erlbaum Associates; 1997.
2. Cohen J: The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology* 1962, 65: 145-153.
3. Glass GV: Primary, secondary, and meta-analysis of research. *Educational Researcher* 1976, 5: 3-8.
4. Vacha-Haase T, Thompson B: How to estimate and interpret various effect sizes. *Journal of Counseling Psychology* 2004, 51: 473-481.
5. Zhang XHD: A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics* 2007, 89: 552-561.
6. Zhang XHD: A new method with flexible and balanced control of false negatives and false positives for hit selection in RNA interference high-throughput screening assays. *Journal of Biomolecular Screening* 2007, 12: 645-655.
7. Zhang XHD: Strictly standardized mean difference, standardized mean difference and classical t-test for the comparison of two groups. *Statistics in Biopharmaceutical Research* 2010, 2: 292-299.
8. Zhang XHD: *Optimal High-Throughput Screening: Practical Experimental Design and Data Analysis for Genome-Scale RNAi Research*. New York: Cambridge University Press; 2011.
9. Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E *et al.*: Statistical methods for analysis of high-throughput RNA interference screens. *Nature Methods* 2009, 6: 569-575.
10. Zhang XHD, Heyse JF: Determination of sample size in genome-scale RNAi screens. *Bioinformatics* 2009, 25: 841-844.
11. Zhang XHD: Assessing the size of gene or RNAi effects in multifactor high-throughput experiments. *Pharmacogenomics* 2010, 11: 199-213.
12. Zhang XHD, Lacson R, Yang RJ, Marine SD, McCampbell A, Toolan DM *et al.*: The Use of SSMD-Based False Discovery and False Nondiscovery Rates in Genome-Scale RNAi Screens. *Journal of Biomolecular Screening* 2010, 15: 1123-1131.
13. Mahalanobis PC: On the generalised distance in statistics. *Proceedings National Institute Sciences, India* 1936, 2: 49-55.