GSSMD: A new standardized effect size measure to improve robustness and interpretability in biological applications

Seongyong Park and Shujaat Khan
Department of Bio and Brain Engineering
Korea Advanced Institute of Science and Technology
Daejeon, South Korea
{sypark0215, shujaat}@kaist.ac.kr

Muhammad Moinuddin and Ubaid M. Al-Saggaf
Center of Excellence in Intelligent Engineering Systems
King Abdulaziz University
Jeddah, Saudi Arabia.
{mmsansari, usaggaf}@kau.edu.sa

Abstract—In many biological applications, the primary objective of study is to quantify the magnitude of treatment effect between two groups. Cohens’ d or strictly standardized mean difference (SSMD) can be used to measure effect size however, it is sensitive to violation of assumption of normality. Here, we propose an alternative metric of standardized effect size measure to improve robustness and interpretability, based on the overlap between two sample distributions. The proposed method is a non-parametric generalized variant of SSMD (Strictly Standardized Mean Difference). We characterized proposed measure in various simulation settings to illustrate its behavior. We also investigated finite sample properties on the estimation of effect size and draw some guidelines. As a case study, we applied our measure for hit selection problem in an RNAi experiment and showed superiority of proposed method.

Index Terms—Standardized effect size measure, Overlap statistics, Robustness, Interpretability, Biological applications

1. Introduction

In typical biological experiments, we often divide independent groups into controls and cases to quantify the treatment effect. The effect size is a statistical metric that provides an estimate of the difference between the two groups, and it is an indicator that is calculated as a standardized mean difference divided by pooled estimate of sample standard deviation. Variants of Cohen’s d such as Glass’s δ and Hedge’s g can be used for the same purpose and they have the same form but use slightly different way to estimate the sample variance. The strictly standardized mean difference (SSMD) [1] is another variant of Cohen’s d and it is popular in bioassay quality control and hit selection problems. Different bioassay quality control measures for effect size estimation are available e.g., Z’-factor and SSMD, and robust SSMD etc. The primary difference in robust variants is that they utilize median and MAD instead of mean and variance.

The drawbacks of aforementioned effect size measures are sensitivity toward underlying distribution and interpretability. Since these measures are only utilize statistical parameters such as sample mean and variance, they are susceptible to noise in the measurement or distribution transformation. In addition, these measures can be difficult to interpret since the range of values are (−∞, ∞) and due to the distributional differences and measurement noise the meaning of the value changes with application domain.

To address aforementioned issues, we propose a new effect size measure called generalized strictly standardized mean difference (GSSMD). Since the GSSMD is defined by the non-overlap proportion between two distributions, it avoids the complexity of other effect size measure calculations in non-standard or transformed distributions. By definition, it is intuitive enough and easy to interpret. In this paper, we used simulations and real biological data sets to support our claims. Mathematical definitions and experimental results are described in the following sections.

2. Methods

2.1. Definition of overlap statistics and GSSMD

In general, distribution overlap between two continuous probability density functions (PDFs) is defined as

\[
OVL = \int_{\mathbb{R}^n} \min[f_A(x), f_B(x)] dx
\]

(1)

where \(f_A(x)\) and \(f_B(x)\) represent two sample distributions. If samples are drawn from two normal distributions \(f_A(x)\) and \(f_B(x)\), then \(OVL\) is equivalent to

\[
OVL = \int_{\mathbb{R}^n} \min[N(x|\bar{x}_A, s_A^2), N(x|\bar{x}_B, s_B^2)] dx.
\]

(2)

This is equivalent to

\[
OVL = 2\Phi\left(-\frac{\bar{x}_{pos} - \bar{x}_{neg}}{2s_{pooled}}\right).
\]

(3)
where, $\Phi(\cdot)$ is the cumulative distribution function. To use overlap statistic as differential measure, herein we defined GSSMD as

$$GSSMD = sign(\bar{x}_{pos} - \bar{x}_{neg}) \times (1 - OVL) \quad (4)$$

where $sign(\cdot)$ is sign operator and the GSSMD is in the range $-1 \leq GSSMD \leq 1$.

2.2. GSSMD and detection probability

Let us assume two probability density functions $P_{pos}(x)$ and $P_{neg}(x)$ representing positive and negative population distributions of an assay and a random variable $x$ representing assay values as shown in supplementary Figure S1. Here we assumed that $P_{pos} > P_{neg}$. The type I (False Positive Rate, FPR) and type II (False Negative Rate, FNR) errors or mis-classification risks of a classifier $h(x)$ with a given threshold $\epsilon$ can be formulated by

$$R_0(h|\epsilon) = P(h(x|\epsilon) \neq Y|Y = 0)$$
$$R_1(h|\epsilon) = P(h(x|\epsilon) \neq Y|Y = 1)$$

The formulae can be obtained by integrating two distributions

$$R_0(h|\epsilon) = \int_{-\infty}^{\infty} p_{neg}(x)dx$$
$$R_1(h|\epsilon) = \int_{-\infty}^{\infty} p_{pos}(x)dx.$$

Hence, a probability of error or total risk of mis-classification can be calculated as

$$R(h|\epsilon) = P(h(x|\epsilon) \neq Y)$$
$$= P(Y = 0) R_0(h|\epsilon) + P(Y = 1) R_1(h|\epsilon) \quad (5)$$

If the prior probability $P(Y = 0) = P(Y = 1) = 0.5$, then the risk becomes

$$R(h|\epsilon) = \frac{R_0(h|\epsilon) + R_1(h|\epsilon)}{2} \quad (6)$$

We can find optimal threshold $\epsilon_0$ of the classifier $h(x)$ from $P_{min} = min, R(h|\epsilon)$ by combining integral forms with this equation at optimal threshold, we can get

$$P_{min} = \frac{1}{2} \left( \int_{-\infty}^{\epsilon_0} p_{pos}(x) dx + \int_{\epsilon_0}^{\infty} p_{neg}(x) dx \right) \quad (7)$$

which is identical to $P_{min} = OVL/2$ or equivalently, $P_{max} = 1 - OVL/2$. The GSSMD considers directional effect of this measure to adapt biological experimental condition. Therefore, the analytical relationship between GSSMD and the maximum probability of success can be defined by

$$GSSMD = sign(\bar{x}_{pos} - \bar{x}_{neg}) \times (2P_{max} - 1) \quad (8)$$

If the prior probability $P(Y = 0) \neq P(Y = 1)$, GSSMD still correlated with $P_{max}$ but the optimal threshold $\epsilon_0$ may differ from $P(Y = 0) = P(Y = 1)$ case. We may apply Neyman Pearson paradigm to get the optimal threshold $\epsilon_0$ for a given false discovery rate $\alpha$ as recently proposed by Xin Tong et. al [2].
showed similar performance as GSSMD. The dot represents mean of grey line is inserted to notice scale difference.

The data are unknown and in addition to this the value of SSMD has no intuitive meaning when distributions of outlier samples is greater than 10. However, the values to outliers and approaches to specific values when the mean converged to the percentage of outlier when \( \mu_{\text{diff}} \geq 5 \). The horizontal grey line is inserted to notice scale difference.

\frac{\text{mean of outlier samples}}{\text{mean of outlier samples}} \geq \frac{5}{10}

In comparison to \( \text{Z}'\)-factor which is adversely sensitive, SSMD and robust version of SSMD or SSMD\text{robust} measures are less sensitive to outliers and approaches to specific values when the mean of outlier samples is greater than 5. However, the values of SSMD has no intuitive meaning when distributions of the data are unknown and in addition to this the value of SSMD\text{robust} varied greatly depending on different sampling scheme (Fig. 2(B)). Thus, GSSMD is the reliable measure for detecting outliers at small mean differences because its values have intuitive meaning, indicating the percentage of outliers. For the robustness analysis of the proposed method against different noise conditions see Figure S4 and S5 in supplementary material.

3.3. Finite sample property of GSSMD

In this experiment, we estimated the lower-bound of GSSMD for two sample sets. In particular, we tested normal and log-normal distribution cases both with \( \mu = 0 \) and \( \sigma = 1 \). Theoretically, the expected GSSMD value for both scenarios is 0, however due to sampling variations the calculated value might be different due to the variations in estimation of PDFs. Therefore, we performed an 10,000 independent trials and calculated the mean, variance and extreme values of GSSMD for each sample size.

Fig. 3 show the estimated GSSMD for various sample size (\( N = 3 \times 10^6 \)). As expected, as the sample size increase, the distance between the mean of the estimated GSSMD and 0 decrease. We also found that the shape of distributions of estimated GSSMD with the sample size of more than 100 is stabilized. The lower-bound of GSSMD for a specific sample size can be seen in the figure. Here we showed only normal distribution case, but the overall behavior is quite similar for log-normal distribution as well. (see supplementary material: Figure S6 and S7).

In fact, for log-normal distribution the lower-bound is even smaller than the normal distribution case, however for simplicity one can choose the maximum of the normal distribution to avoid possible false positives. This robust and intuitive nature of GSSMD measure make it a suitable choice for reliable assessment of assay quality. To reject null hypothesis, throughout our simulation study, we chose 5% as the threshold of GSSMD for sample size \( N = 1,000 \).

3.4. Case study

For the case study, we selected RNAi screening of cell viability in Drosophila Kc167 cells [5] and performed the plate aggregated type of hit selection. Fig. 4 shows the calculated measures on the dataset, where red dots represent negative controls (GFP, Rho, no RNAi treatment) and blue dots represent positive controls (RNAi for D-IAP1 which inducing time-dependent cell death in Drosophila Kc167 cells). Grey dots are samples treated with RNAi, and it is unknown how much reduction in cell viability indicates targeted RNAi binding. As discussed earlier, like SSMD, the proposed GSSMD can also be used to identify thresholds for hit selection. Since the risk of misclassification \( R_0(h) \) dependent on the overlap between two distributions, we can find the optimal threshold of GSSMD for a given FDR based on the value of overlap in a certain range.

Fig. 4(A) shows upper and lower bounds of GSSMD as well as threshold values. The upper and lower bounds of each measure are calculated as discussed earlier. In short, we read the value of luciferase activity \( y \) at the \( OVL \) on each ECDF (positive and negative control distributions, see supplementary material: Figure S12). The threshold was selected by \( x \) at maximum \( OVL \) occur. In order to evaluate the performance of each measure for finding a correct threshold of hit selection, we trained a logistic regression classifier based on positive and negative control samples. The logistic
Figure 4. GSSMD can be used to find hit thresholds in RNAi screening. (A) Upper and lower bounds of GSSMD as well as threshold values. (B) Upper and lower bounds of SSMD and robust SSMD as well as threshold values. (C) Corresponding threshold values for each case. We set the GSSMD threshold at the point where the value of overlap is maximum. Both recommended SSMD and GSSMD threshold were similar to the threshold of optimal logistic regression classifier (Accuracy = 0.956, Type I error = 0.018).

4. Discussion and Conclusion

Effect size measures are used in many domains including bioassay as we described above. We showed that non-parametric estimation of effect size based on overlap statistics may provide robust and interpretable estimate of effect size compared to conventional measures. As described in the case study, SSMD may be too loose for suitable selection thresholds for hits even in experiments for which the metric is designed. Moreover, even if there is a guideline to choose alternative threshold, the categorization of positive control is subjective due to the qualitative nature of the guideline. In contrast, the proposed GSSMD measure is relatively sensitive to the detect changes in biological assays. Non-parametric characteristics may cause deviations in the estimates of the measurements, but we have shown that the appropriate number of samples and bin size can minimize this variability.

It may be infeasible to use histogram based overlap estimation in extremely small sample size. In those cases, we can use parametric distribution fitting method based on the understanding of experiment (see supplementary material: Figure S8, S9, and S11). It is also possible to use KDE based estimation instead of histogram based method. We found that when we have enough sample number \( n \geq 100 \), histogram with adaptive bin-size showed better estimates compared to those KDE based method (see supplementary material: Figure S10).

We believe that the proposed method can be easily implemented in many biological applications such as bioassay, gene expression meta analysis etc. The supplemental material and implementation code are available at the author’s Github page (https://github.com/psychemistz/gssmd).

Acknowledgment

This research was funded by Center of Excellence in Intelligent Engineering Systems (CEIES), King Abdulaziz University, Jeddah, Saudi Arabia. We thank Professor Alfonso Rodriguez-Molares, Norwegian University of Science and Technology for sharing a MATLAB implementation of GCNR, a similar task performed in ultrasound image quality assessment. We also thank Yoonhyeok Lee for helpful discussion about applying GSSMD measure on the small-scale scenario.

References

[1] X. D. Zhang, “A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays,” Genomics, vol. 89, no. 4, pp. 552–561, 2007.
[2] X. Tong, Y. Feng, and J. J. Li, “Neyman-pearson classification algorithms and NP receiver operating characteristics,” Science advances, vol. 4, no. 2, p. eaao1659, 2018.
[3] J.-H. Zhang, T. D. Chung, and K. R. Oldenburg, “A simple statistical parameter for use in evaluation and validation of high throughput screening assays,” Journal of biomolecular screening, vol. 4, no. 2, pp. 67–73, 1999.
[4] C. J. Cellucci, A. M. Albano, and P. E. Rapp, “Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms,” Physical Review E, vol. 71, no. 6, p. 066208, 2005.
[5] M. Boutrous, L. P. Bräs, and W. Huber, “Analysis of cell-based RNA screens,” Genome biology, vol. 7, no. 7, p. R66, 2006.