A Method with Flexible and Balanced Control of False Negatives and False Positives for Hit Selection in RNA Interference High-Throughput Screening Assays: A Statistical Terminology

EDNA SCHECHTMAN

Zhang suggests a new method that is flexible and controls the balance between false negatives and false positives for hit selection in RNA high-throughput screening assays. The author shows that the same decision rules and balances can be expressed by familiar statistical terms such as type I error and power and hence connects the new method to known statistical tools. (Journal of Biomolecular Screening 2008:309-311)

Key words: strictly standardized mean difference, restricted false-positive rate, high-throughput screening, hit selection

INTRODUCTION

In a recent paper, Zhang1 suggests a new method for hit selection in RNA interference (RNAi) HIT assays with flexible and balanced control of false negatives and false positives. The method "allows the differentiation between siRNAs with large and small effects on the assay output and maintains flexible and balanced control of both the false-negative rate and the restricted false-positive rate."1

In this note, we suggest a different formulation of Zhang's method, using statistical terminology that will relate Zhang's suggestion to known concepts such as the power of a test.

Zhang, in his article, states, "First, we do not want siRNA with large effects to be selected as nonhits, and second, we do not want the siRNA with small effects to be selected as hits. In terms of statistics, we want to control the false-negative rate in which the siRNA with large effects are not selected as hits as small as possible. Meanwhile, we also want to control the false-positive rate in which the siRNA with small effects are selected as hits as small as possible."1(p648) Next, Zhang suggests a new decision process or, as he says, a new hit selection method that maintains a balanced control of both error rates.

The objective of this short note is to help the researcher identify the relationship of the above terms to the relevant statistical concepts. In this short note, we suggest to “translate” the problem to statistics language. We shall illustrate the idea on 1 case—large positive effect. The second case (large negative effect case) can be treated similarly. Zhang1 defines the problem as “we want to select siRNA with large positive effect represented by $\beta \geq c$, where $c (c > 0)$ is a preset level of minimal magnitude of difference that we want to achieve.” We shall show that once the problem is stated as a hypothesis-testing problem, Zhang’s false-negative level (FNL) and the restricted false-positive level (RFPL) are the ordinary type I and II rates, and the flexibility can be translated into choosing a tolerance level and a significance level in the power function. The advantage of this presentation is that one can use any statistical software to calculate the power function and use it to find the appropriate balance between the 2 types of errors.

THE OBJECTIVE AND THE SUGGESTED FORMULATION

Let $P_1$ and $P_2$ be 2 populations with means $\mu_1$ and $\mu_2$ and variances $\sigma_1^2$ and $\sigma_2^2$, respectively. The new measure that was introduced by Zhang2 for measuring the magnitude of difference between 2 populations is called strictly standardized mean
difference (SSMD). This measure was first introduced in Zhang, and its properties were further investigated in Zhang and used in Zhang et al.

Assuming that the 2 populations are independent, SSMD is defined by

$$\beta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

We note in passing that the square of this measure has been used by others to measure the overlap between distributions and is called Fisher’s discriminant ratio.

To draw inference about the measure, we first introduce its estimator. Assuming that the populations are normally distributed, as was assumed by Zhang, the maximum likelihood estimator of $\beta$ is given by

$$\hat{\beta} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}}$$

where $\bar{X}$ and $S^2$ are the sample mean and sample variance, respectively.

Let us formulate the problem as a hypothesis-testing problem. Let $c$ ($c > 0$) be a preset level of minimal magnitude of difference that we want to achieve, as suggested in Zhang. Then we can state our hypotheses as

- $H_0$: $\beta \geq c$
- $H_1$: $\beta < c$

Rejecting $H_0$ in favor of $H_1$ means declaring that there is no HIT when in fact there is a difference at least as big as $c$. In the statistics world, this means making a type I error, and its rate of $H_1$ if $\beta = \beta^*$, where $\beta^*$ is the solution of the following equation, based on the type I error rate of $\alpha$:

$$\alpha = P(\hat{\beta} < \beta^* | H_0).$$

The solution is $\beta^* = c - Z_{\alpha} \hat{\sigma}_\beta$, where $\hat{\sigma}_\beta$ is the estimator of the standard deviation of $\hat{\beta}$ and $Z_{\alpha}$ is the upper $\alpha$th percentile from the standard normal distribution (see the appendix for details). Note that the same decision rule is suggested in Zhang as decision rule I. It is important to emphasize that, as mentioned by Zhang, these results are asymptotic results. They are based on the fact that the maximum likelihood estimator of $\beta$ is asymptotically normally distributed. Therefore, the results hold when the sample size is large.

We now move to type II error. In our case, a type II error means declaring a difference at least as big as $c$ (i.e., declaring a hit) when in fact there is no hit. When dealing with type II error, one has to specify a value in the alternative. Zhang says, “Let us assume that we can tolerate the false positive with $\hat{\beta} > c$, where $0 \leq c \leq c$ for selecting the siRNA with large positive effects.” To be consistent with Zhang’s notation, we shall choose $c < c$ as the value in the alternative. In statistical terminology, we can calculate the probability of making a type II error (usually denoted by $\beta$, but we will use the word $\beta$ to avoid confusion), under the assumption that $\hat{\beta} = c$.

$$\beta = \beta (c_2) = P(\hat{\beta} > c | \beta = c_2).$$

Note that $\beta$ is a function of the chosen value of $c_2$.

Zhang calls this probability the restricted false-positive level.

The solution is $\beta = \Phi \left( \frac{c - c_2}{\sigma_\beta} - Z_{\alpha} \right) = \Phi \left( \frac{c - c_2}{\sigma_\beta} \right)$.

where $\Phi$ is the cumulative distribution function of the standard normal distribution.

The power of the test (i.e., of the decision rule) is

$$\text{Power} (c_2) = 1 - \beta (c_2) = \Phi \left( \frac{c - c_2}{\sigma_\beta} \right).$$

Obviously, the power is a function of $c_2$—the smaller the value of $c_2$, the larger the power. Intuitively, this means that the smaller the difference you are willing to tolerate, the fewer type II errors you will make, and the more the power will go up.

It is well known that it is not possible to control both type I and type II errors simultaneously for a given data set—there is a trade-off. The practice used is to draw a power function, for a given $\alpha$, and let the researcher choose $c_2$ that he or she can live with (in terms of type II error) or, alternatively, increase $\alpha$ (which will in turn decrease $\beta$ and increase the power). Another option is to change the value of $c_2$.

We demonstrate it in Figure 1. For the illustration, we chose $c = 3$ and $\sigma = 0.707$ (same values as in Zhang). The plots illustrate 2 main points. The first point is the dependence on $\alpha$—the bigger the $\alpha$, the higher the power. Intuitively, if one agrees to have a higher type I error, his or her type II error will be smaller and hence higher power. The second point is that the power decreases as $c_2$ increases, as was explained above. The plot illustrates the trade-off, and it is up to the researcher to choose the configuration he or she wants to work with.

**CONCLUDING REMARKS**

Zhang suggests a new method that is flexible and controls the balance between false negatives and false positives for hit selection in RNA high-throughput screening assays and illustrates...
Flexible and Balanced Control for Hit Selection

APPENDIX

The calculation of $\beta^*$

Let

$H_0$: $\beta \geq c$

$H_1$: $\beta < c$

The decision rule for this 1-sided alternative is of the following form: reject $H_0$ in favor of $H_1$ if $\hat{\beta} < \beta^*$, where $\beta^*$ is the solution of the following equation, based on the type I error rate of $\alpha$:

$$\alpha = P(\hat{\beta} < \beta^* | H_0) = P\left(z < \frac{\beta^* - c}{\hat{\sigma}_\beta}\right).$$

Therefore,

$$\frac{\beta^* - c}{\hat{\sigma}_\beta} = -Z_{\alpha},$$

and the solution is $\beta^* = c - Z_{\alpha} \hat{\sigma}_\beta$, where $\hat{\sigma}_\beta$ is the estimator of the standard deviation of $\hat{\beta}$, and $Z_{\alpha}$ is the upper $\alpha$th percentile from the standard normal distribution. Note that, as mentioned by Zhang,$^{1,2}$ these results are asymptotic results and require a large sample size.

For completeness, we give the expression of $\hat{\sigma}_\beta$, taken from Zhang.$^{1,2}$

$$\hat{\sigma}_\beta = \frac{(n_1 - 1)S^2_1}{n_1} + \frac{(n_2 - 1)S^2_2}{n_2} + \left[\frac{(n_1 - 1)S^4_1}{n_1} + \frac{(n_2 - 1)S^4_2}{n_2}\right]$$

Also,

$$\hat{\sigma}_\beta^2 = \frac{(n_1 - 1)S^2_1}{n_1} + \frac{(n_2 - 1)S^2_2}{n_2}.$$