Statistical thinking in the simulation design: a continuing conversation on the balancing intercept problem

Boyi Guo, MS, PhD, Linzi Li, MSPH, Jacqueline Rudolph, PhD

As the field of epidemiology evolves, there are growing interests to employ more computational approaches to solve analytic problems. Among them, simulation is one of the most accessible concepts. Previous literature argues the importance of simulation in epidemiology education and research. [TODO: add citations] In this commentary, we review a series of recent discussions on the balancing intercept problem published in previous issues of the American Journal of Epidemiology [TODO: add citations]. Specifically, we explain the balancing intercept problem in greater detail, describing the growing complexity when generalizing to a wider class of data generating mechanisms, e.g. multinomial exposure, and covariate adjustment. We hope to help the audience to connect the simulation problem to the most commonly used statistics concepts, including variable enumeration, link function and design matrix. We also refine the closed-form calculation of balancing intercepts for commonly used log-linear models.

The balancing intercept problem was first introduced by Rudolph et al. (2021). [1] The objective is to control the marginal mean of simulated outcomes at the desired level when limited information is available. To explain with a toy example, we are interested in simulating normally distributed outcomes () for two groups of samples, in other words a binary exposure () with known group sizes. In an ideal scenario where the corresponding group (conditional) means are known, we can simply sample the outcomes for each group using these group means and augment the simulated data to form the overall dataset. Nevertheless, without any calculation, this simulation approach fails if we only know the marginal mean () and the mean difference () between the two groups. Acknowledging the degree of freedom is fixed, we can use the marginal mean-mean difference pair to calculate the group means. This calculation is referred to as the balancing intercept problem. Specifically, the balancing intercept () is the conditional/group mean of the reference group when the group membership/binary exposure is enumerated using the reference coding system, . The reference coding system has been the implicit default variable enumeration system in the previous balancing intercept discussions. Rudolph et al. (2021) provided a closed-form equation to calculate the balancing intercept, equivalently the reference group mean, by arranging the mean function,

Practically, most, if not all, simulation designs are more complex than a two-sample normal outcome design, for example, considering various types of the outcome and the estimand of interest, adjusting cofounding in graphical causal models, and expanding the exposure variable from binomial to multinomial. Equation 1 does not generalize to these complex designs, first noticed by Robertson, Steingrimsson, and Dahabreh (2021). In the following paragraphs, we provide the statistical rationales deciphering these design complications and derive a generalized solution for log-linear models.

Firstly, we discuss why generalizing the type of the outcome and the estimand of interest complicates the calculation of the balancing intercept. Similar to fitting a generalized model, a link function, , is required to describe the expected mathematical relationship between the exposure and the mean of the outcome in the simulation design. The choice of the link function is highly relevant to the type of outcome and dictated by the estimand of interest. For example, we can simulate Gaussian outcomes with a log function (link function) to study the mean ratio (estimand) or a binary outcome with a logit function to study the odds ratio. When the link function is nonlinear, e.g. log function, the equality between the expectation of a link function and the link function of an expectation doesn’t hold any more, . Therefore, the same mathematical manipulation in Equation 1 does not apply when the link function is nonlinear. For interested readers, we defer to the [Supporting Information](https://github.com/boyiguo1/Manuscript-Balance_Intercept/blob/master/Manuscript/appendix.pdf) for the complete mathematical reasoning. Is it possible to derive a closed-form equation for non-linear link function? Yes, but only for limited number of link functions. In the supporting information, we show the derivation of the balancing intercept for log link function,

.

Another popular link function, logit function, does not have a tractable solution. Hence, we recommend to use the numeric approximation approaches proposed by Robertson, Steingrimsson, and Dahabreh (2021) and Zivich and Ross (2022). [2, 3]

Secondly, we look at how to calculate the balancing intercept when covariates are included in the simulation design. Covariate adjustment is one of the most indispensable concepts in quantitative analysis. It is highly relevant to quantify and test causal mechanisms. Recent research articles emphasize how to simulate causal relationships embedded in a directed acyclic graph where confounders are indisguised as covariates in a regression model. The complication of including covariates in the simulation design co-exits with the nonlinear link function complication. The closed-form equation for log link function when including covariates () generalizes to

.

Only if the variables, including the exposure, are pairwise independent, the closed-form equation can simplify to

.

For commonly used distributions, we can replace the expectation of exponential function, with a moment generating function, a well-defined equation, to simplify the calculation. For example, if a normally distributed variable with mean and variance , the moment generating function is . It is not uncommon that the variables are not pairwise independent or the moment generating function doesn’t exist. In these situations, we can apply the Monte Carlo technique to derive Specifically, one can sample the vector of variables with replacement for a large number of iterations (say 1000), and average the exponential function of the created randomly sampled data.

One topic that was not explicitly discussed in the previous balancing intercept literature is how to generalize a binomial variable, both as the exposure or as the covariate, to a multinomial variable. This generalization can be also seen as allowing the interaction of two categorical variables, either as exposure-covariate interaction or covariate-covariate interaction, in the simulation design, which is closely related to effect modification modeling. The challenge here is the enumeration of the categorical variable. When enumerating a binary variable, we create a data column containing zeros and ones to represent the two levels of a variable (implicitly under the referencing coding scheme). Calculating the mean of this enumeration () is straightforward, and hence, Equation 1 is mathematically well-defined. Nevertheless, when the variable is multinomial, the mean is not obvious and Equation 1 fails. Specifically, when enumerating a multinomial variable with levels, we need to create columns in the data matrix to indicate these levels. Each column marks the membership in a corresponding level using either 1 or 0. It would be possible to treat each column as a binary variable and follow the previous equations. But this approach lacks accuracy and is insensible. The such practice treats each column as an independent binary variable and ignores the grouping structure and correlation among columns, resulting in an inflated approximation of the balancing intercept. To circumvent this problem, we recommend using the moment generating function to calculate the mean (of a function). For example, if we have a three-level multinomial exposure with the probability () and the coefficients on the log scale (), the moment generating function gives to replace in the Equation 2. In the presence of statistical interactions, one can simply treat the statistical interaction as a special case of a multinomial variable by enlisting all possible combinations.

To demonstrate the closed-form equation (Equation 2), we conduct a simulation study motivated by Robertson, Steingrimsson, and Dahabreh (2021). The simulation follows a log-normal model with two independent variables, as the exposure and as the covariate. We assume our exposure is a three-level categorical variable with the probability 0.5, 0.35, 0.15. We examine difference distributions of the covariate , including a Bernoulli distribution with probability 0.8, a continuous uniform distribution bounded between -1 and 3, a standard normal distribution and a gamma distribution with shape 1 and rate 1.5. We also examine different magnitudes of covariate coefficient ranging from 1 to 3 with 0.5 increments while fixing the coefficients for the exposure at 0.2, -0.2. The targeted marginal mean considers a sequence of values, from 0.1 to 0.9 with 0.1 increments. For each combination of these parameters, we use the Equation 2 to calculate the balancing intercept and simulate a dataset that consists of 10,000 observations. We calculate the deviation of the observed mean from the target mean, referred to as bias. The process iterates 10,000 times and calculates the averaged bias. Figure 1 shows that the closed-form equation produces accurate balancing intercepts. While we were conducting the simulation study, we observed some interesting numeric problems that we would like to highlight here. When using an unbounded link function, e.g. log link function, to simulate a bounded outcome, e.g. probabilities or binary outcomes, it is difficult to control the marginal mean with an analytic solution, particularly when the effect size is large. For example, if we run the previously described simulation study with a binary outcome instead of the normal outcome, the same described process would produce a dataset that has a marginal probability that is lower than the target (See supporting information Figure 1).

During the investigation, we are also interested in if changing the variable enumeration would have an impact on the calculation of the balancing intercept. The previous balancing intercept literature defaults to the reference coding scheme without mentioning other systems, for example, another commonly used scheme, effect coding. Instead of using 0s and 1s to indicate membership, effect coding uses 0, 1, and -1, emphasizing the deviation from the grand mean, i.e. the average of level means. Under the effect coding scheme, the balance intercept describes the grand mean. When a study is balanced with respect to the exposure variable, the grand mean coincides with the marginal mean. Hence, the marginal mean can be directly used as the balancing intercept and requires no further calculation. In case of an unbalanced design, one can use weighted effect coding instead. However, this simplification has limited utilities - the complications due to or related to the nonlinear link function persist.

In this commentary, we provide a statistical analysis of the balancing intercept problem, describing the consideration of various estimands, the inclusion of covariates, and the generalization to multinomial variables. We also derived a close-form equation to calculate the balancing intercept for simulation designs with the log link function. Simulation studies are conducted to demonstrate the close-form equation.

We also review some fundamental statistical concepts, for example, coding scheme, link function, expectation calculation and moment generating functions. These concepts appear daunting and distanced in the introductory statistical training that is required in most epidemiology education programs. The balancing intercept problem provides an adequate platform to exemplify these concepts and improve students’ understanding. Meanwhile, we would like to advocate the necessity/importance of fundamental statistics training in epidemiology, even in the new era of computation. The growing computation power can greatly reduce the technical burden of deriving analytic solutions. Nevertheless, the accuracy of numeric solutions greatly depends on the perfection of their implementations, and could be easily overlooked. In addition, statistical knowledge and theory “provides a shortcut to computation” and substantially stimulate new approaches to analyze data.

Reference

1. Rudolph, J.E., et al., *SIMULATION IN PRACTICE: THE BALANCING INTERCEPT.* American Journal of Epidemiology, 2021. **190**(8): p. 1696-1698.

2. Robertson, S.E., J.A. Steingrimsson, and I.J. Dahabreh, *Using Numerical Methods to Design Simulations: Revisiting the Balancing Intercept.* American Journal of Epidemiology, 2022. **191**(7): p. 1283-1289.

3. Zivich, P.N. and R.K. Ross, *RE: “USING NUMERICAL METHODS TO DESIGN SIMULATIONS: REVISITING THE BALANCING INTERCEPT”.* American Journal of Epidemiology, 2022.

