Lean on your statistics: The generalization and simplification of the balance intercept problem

Boyi Guo, Jacqueline Rudolph

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 discussion on the balance intercept problem published on previous issues of American Journal of Epidemiology [TODO: add citations]. Specifically, we explain the balance intercept problem from a statistical perspective and derive a closed-form solution for some commonly used data generating mechanisms. In addition, we provide some tips and tricks to simply the simulation process.

The balance intercept problem was first introduced by Rudolph et al. (2021). The objective is to control the marginal means of simulated outcomes at a desired level when limited information is presented. To start with a simple example, we are interested in simulating normally distributed outcomes for two groups of samples, in other words a binary exposure, with fixed group sizes. In an idea scenario where the corresponding group (conditional) means are known, we can simply sample the outcomes for each group using these group means and aggregate the simulated data to form the overall dataset. Nevertheless, this approach fails if we only know the marginal mean and the mean difference, which retains the same amount of statistical information. This is because the model degree of freedom is fixed. The process of deriving the conditional means based on marginal mean and the mean differences is the balance intercept problem. Specifically, the balance intercept is the conditional mean of the reference group/group when the categorical covariates are enumerate using the reference coding system. The closed-form equation in Rudolph et al. (2021) can be used to calculate the balance intercept in this toy example.

[TODO: insert the equation here]

Most simulation designs are more complex than the two-sample normal outcomes design, for example the consideration of estimands and outcomes, covariate adjustment and graphical causal model, and multiple-level exposures. Hence, calculating the balance intercept is more complicated and requires further discussion in the following paragraphs.

To generalize the balance intercept problem to other outcomes and estimands, we need to first review a simple statistics concept, the link function. Similar to the generalized linear model, we need a link function to describe the mathematical relationship between linear predictors and the mean of a outcome in the simulation design. 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 calculating the balance intercept, the choice of the link function in the simulation design dominates the complexity of the calculation. For example, when a nonlinear link function is used, the closed-form equation in Rudolph et al. (2021) becomes an approximation of balance intercept and fails to control the marginal mean with accuracy, due to the inequality between the expectation of a link function and the link function of an expectation. [TODO: add math notation] Meanwhile, the choice of link functions limits if a closed-form equation of the balance intercept is possible. We derive the the closed-form equation for linear link function and log link function in the appendix. The balance intercept is mathematically intractable with logit link function, and hence, we recommand to use numeric approximation proposed by Robertson, Steingrimsson, and Dahabreh (2021) and Zivich and Ross (2022).

Similar to generalized linear models, the choice of estimands decides the link function in the data generating model. For example, if the estimands of observation is risk difference, an identify function is preferred instead

of a logistic function in the data generating model. The choice of the link function in the data generating model has impact on the accuracy of the imbalance intercept approximation and will be elaborate more later.

control the marginal probability of binary outcomes when constructing a simulation study. The authors proposed to calculate a "balance intercept" to replace the "standard intercept" in simulation procedures. Robertson, Steingrimsson, and Dahabreh (2021) discovered that the analytic solution of the balance intercept produces inaccurate control when the data generating model employs a logistic link function. They proposed to use root finding algorithms to find the balance intercept numerically. Later, Zivich and Ross (2022) generalized the numeric solutions to several other outcome distributions.

To better understand the balance intercept problem, we first review two basic statistics concepts, coding schemes and link functions. A coding scheme describes how a categorical variable is enumerated in a regression system. Most commonly used coding schemes include the reference coding (also known as dummy coding) and the effect coding. Both coding schemes create $p-1$ columns for a categorical variable with p levels. The reference coding employs 0 and 1 to denote the level an individual belongs, while the effect coding employs 0, 1, and -1. For example, a binary variable would be coded as a vector of 0 for the reference level and 1 for the comparing level using the reference coding and a vector of 1 and -1 using the effect coding. (See an example of three-level categorical variable in Table 1). As the enumerations are different for the two coding schemes, they offer different interpretations of regression coefficients. The reference coding emphasizes the change relative to a reference level of preference; the effect coding emphasizes the deviation from the grand mean (here refer to as the mean of the level means). Nevertheless, the two schemes translate to each other one-on-one and provide the same statistical inference. To note, both coding schemes can be applied ubiquitously in any regression system regardless the outcome distribution and the link function of choice. In addition, a link function describes the mathematical relationship between linear predictor and the mean of the outcome. Similar to generalized linear models, the choice of estimands decides the link function in the data generating model. For example, if the estimands of observation is risk difference, an identify function is preferred instead of a logistic function in the data generating model. The choice of the link function in the data generating model has impact on the accuracy of the imbalance intercept approximation and will be elaborate more later.

The original proposal (Rudolph et al. 2021) of the balance intercept defaults to the reference coding system without explicit mentioning, and carries along in the subsequent studies. The intercept term in the reference coding scheme describes the conditional probability of the reference level of a categorical variable. Forcing the intercept term to the target marginal probability (referred to as the standard intercept in Rudolph et al. (2021)) would naturally fail due to the definition discrepancy. Statistically, to derive the balance intercept is to derive the conditional probability of the reference level and can be achieved with analytic and numeric approximations. Nevertheless, the accuracy could be vulnerable to the quality of numeric analysis and programming proficiency.

The primary reason that the analytic approximation (Rudolph et al. 2021) of the balance intercept may fail is because of the inequality between a function of the expectation and the expectation of a function when the function is not linear. Derive the analytic approximation on the linear predictor scale as in Rudolph et al. (2021) will introduce this problem when link functions is nonlinear. (Robertson, Steingrimsson, and Dahabreh 2021) Hence, the derivation of the balance intercept should be conducted on the response scale, e.g., the probability scale for binary outcomes regardless of the link function. Following the mathematical derivation in the Supporting Material [TODO: add hyperlink], it is clear Rudolph's balance intercept equation is the exact solution for linear link functions. We also provide a closed-form solution for logarithm link function. Another commonly used link function, the logit function, is not easily tractable, and hence is recommended to use Monte Carlo algorithms or root-find algorithms to derive numerically.

Using the reference coding system in data generating models is very convenient when the comparison between the group means are the known information. Nevertheless, it requires additional calculation when conditional probabilities are given. In this case, it is more convenient to use effect coding system, where the coefficients describe the discrepancy of the level means. Meanwhile, the intercept term in the effect coding scheme describes the mean of the group means, which coincides with the marginal probability when the sample sizes are balanced across levels. In other words, the balance intercept problem does not exist for simulations with balanced designs when using the effect coding. When the underlying design is not balanced, the intercept is simply to calculate the grand mean by averaging the conditional probabilities, $\delta_0 = \sum_{j=1}^{p} \pi_j / p$.

In this report, we provide a statistics perspective to the balance intercept problem. Specifically, we clarify the simulation procedure when the reference coding scheme and the effect coding scheme are used as the foundation of the procedure. With an statistical analysis of the simulation procedure, we provide a generalized solution to calculate the balance intercept that addresses all forms of outcomes and link functions. We show that the balance intercept problem drastically simplifies when applying effect coding scheme. Notably, when the design of the study is balanced, there is minimum calculation is needed.

In addition to the statistical solution for the balance intercept problem, we want to emphasize that the fundamental skills of statistics can not be ignored, even in the erra of computation. The growing computation power can greatly reduce the technical burden to derive analytic solutions with numeric devices. Nevertheless, the accuracy of numeric solution greatly depends on the perfection of the implementation, and can be easily overlooked. The fundamental statistics skills can provide a shortcut to the correct solution, and provide great translatability across programming languages in comparison to numeric solutions. The author doesn't arugula if analytic approach and computational approach is superior. Instead, we advocate for a balanced emphasis on both computational skills and analytic thinking.

References

Robertson, Sarah E, Jon A Steingrimsson, and Issa J Dahabreh. 2021. "Using Numerical Methods to Design Simulations: Revisiting the Balancing Intercept." *American Journal of Epidemiology*, November, kwab264. https://doi.org/10.1093/aje/kwab264.

Rudolph, Jacqueline E, Jessie K Edwards, Ashley I Naimi, and Daniel J Westreich. 2021. "SIMULATION IN PRACTICE: THE BALANCING INTERCEPT." *American Journal of Epidemiology* 190 (8): 1696–98. https://doi.org/10.1093/aje/kwab039.

Zivich, Paul N, and Rachael K Ross. 2022. "RE: "Using Numerical Methods to Design Simulations: Revisiting the Balancing Intercept"." *American Journal of Epidemiology*, May. https://doi.org/10.1093/aje/kwac083.