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 recommend to use numeric approximation proposed by Robertson, Steingrimsson, and Dahabreh (2021) and Zivich and Ross (2022).

Another layer of complex comes from the covariate adjustment in the simulation design, particularly when the covariates are continuous and have non-zero means. Covariate adjustment is one of the most indispensable

concept in quantitative analysis and it is highly relevant to test and quantify causal mechanisms. Recent research articles emphasize how to simulate causal relationship embedded in a directed acyclic graph. [TODO: add citation] The complication in calculating balance intercept mainly coexits when the linear link is not linear., which causes difficulties to calculate the expectation of the link function. When the link function is nonlinear, we can not simply considers the balance intercept as a linear function of $E(X)$ due to the inequality between g^{-1}(EX) = E(g^{-1}(X). we provide an closed form equation of balance intercept with log link function for estimand on the multiplicative model. To simplify this equation, we can replace the expectation $E(\exp(\beta X))$ with the moment generating function. The moment generating function is a short cut to calculate these expectation of exponential functions, and hence and simplify the calculation. Neverthesless, moment generating function only works when the underlying distribution of the covariate is known. For situations when the underlying distribution is not unknown, e.g.synthesizing outcome using existing data and effect sizes. we can apply Monte Carlo technique to derive $E(exp(betaX)$. Specifically, one can sample the covariates with replacement for a large amount of iterations (stay 1000), and apply the function to the sampled data, and take average.

One of the complications that was not explicitly discussed in previous balance intercept literature is the generalization of binary exposure and comparatives to multiple-level, which can also include the statistical interaction of two categorical variables, i.e. exposure-covariate interaction or covariate-covariate interaction, as an special case. When enumerating a binary variable in the simulation design, we normally create a data column containing zeros and ones to represent the two levels of the variable (with refernce coding). In contrast, we create $p-1$ columns for a p-level variable, where each column is marked with 1 if a subject belongs to a level while 0 otherwise. This enumeration nullifies the closed-form proposed by Rudolph et al. (2021) even with a linear link function. It is unclear if one should calculate the mean of the categorical variable as a multinomial variable. It would also unsensible to calculate the expecation of each column of the data matrix spanning the categorical variable. This treats each column independent binary variables and ignores the grouping strucutre and collinearity of these columns. This will result in for the balance intercept. In these situations we advise to use our proposed model, by applying the moment generating function, it intrinsically handles xxx and provide an more accurate solution. Even though each column of the design matrix can be seen as individual variables, the previous close equation would not work because the variables are not pairwise independent. Nevertheless, we can rely on the mgf of multinomial distribution or Monte Carlo approach to calculate xxx. When statistical interaction exists, one can simply treat it as an special case of multi-level categorical variable by enlist all possible combinations and derive the corresponding risk ratios.

To demonstrate the closed-form equation we provide works, we conduct a simulation studies. The simulation study is motivated by Robertson, Steingrimsson, and Dahabreh (2021). The simulation follows a log-normal model with two independent variables, X_1 serving as the exposure and X_2 as the covariate. We assume our exposure X_1 is a three-level categorical variable with unbalanced group size, with the probability of each level 0.5, 0.35, 0.15. We examine difference distributions of the covariate X_2, 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 magnitude of covariate coefficient β_2 ranging from 1 to 3 with 0.5 increments, while fixing the coefficients β_1 for the exposure X_1 at 0.2, -0.2. The target marginal expectation consider a sequence of valules, from 0.1 to 0.5 with 0.1 increments. For each combination of these parameters, we use the equation xx to calculate the balance 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 to derive the Monte Carlo standard error. The results (Figure yy) shows the close form equation produce accurate balance intercept.

While we were conducting the simulation study, we observed some other numeric problems that we would like to highlight here. When the simulated outcome is binary and the link function is not bounded, it is possible to produce an dataset that not possible to control the marginal probability if following the previous described process. For example, if we run the previously described simulation study with a log function that is lower bounded by 0 but not upper bounded. Following the same described process would produce a dataset that underestimate the marginal probability. (See supporting information Figure xxx). In this case, we advise people to invert the coding of reference level and modify the coefficients accordingly by inverting the coefficients for binary exposures. Hence, this is possible to bound differently.

Other tricks exists. For example, we can leverage different coding schemes to simply the calculation. For example, when the study is balance across exposure, we can set the marginal mean to be the intercept following the definition of intercept in effect coding. When the study is not balanced across exposure, we need to adjust the weighting of group sizes and lever weighted effect coding. Similarly, the marginal mean would be the intercept.

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.