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

Boyi Guo, Jacqueline Rudolph

2022-05-27

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] Even though computational tools can be very helpful, we caution the excess reliance on the computation in analytic problem solving and the total neglect of fundamental statistics theories. In the article, we demonstrate how basic statistics knowledge can simplify the balance intercept problem and provide a generalized solution. Specifically, we dissect the balance intercept problem from a statistical perspective, formulate the balance intercept calculation for all outcomes and link functions, and provide an alternative solution that require significantly less computation.

The balance intercept concept was first introduced by Rudolph et al. (2021) to 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 $p-1$ binary 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 normally employs 0, 1, and -1. For example, a binary variable would be coded as a vector of 0 and 1 with the reference coding and a vector of 1 and -1 with the effect coding. (See an example of three-level categorical variable in Table). As the enumerations are different for the two codign 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 group means). Nevertheless, the two schemes translate to each other one-on-one, and hence 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 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

Levels	Referece			Effect		
	X_1	X_2	Conditional Probability	X_1	X_2	Conditional Probability
Level 1	0	0	β_0	1	0	$\beta_0 + \beta_1$
Level 2	1	0	$\beta_0 + \beta_1$	0	1	$\beta_0 + \beta_2$
Level 3	0	1	$\beta_0 + \beta_2$	-1	-1	$\beta_0 - \beta_1 - \beta_2$

Table 1: One-on-one translation between the reference and effect coding and their calculation of conditional probability of a three-level categorical variable.

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 efficiency. To note, the balance intercept in data generating models always exists and is unique as the model degree of freedom is fixed.

In many scenarios, the balance intercept is analytically tractable with extra cautious. As previously discovered in Robertson, Steingrimsson, and Dahabreh (2021), the analytic approximation fails because of the inequality between a function of the expectation and the expectation of a function when the function is not linear. For example, when the link function is logistic function, the expectation of log odds, $E(logit(Y))$, does not equal to the log odds of the expectation $logit(E(Y))$. Hence, anchoring the calculation in the response scale is preferred in comparison to the linear predictor scale which Rudolph et al. (2021) employs. Suppose the marginal probability π, the estimand ratio (e.g. risk ratio, odds ratio) against the reference level for the ith level, β_{i-1}, and the sample size for the ith level n_i is given, the marginal probability can be formulated as the weighted summation of the conditional probabilities π_i

$$\pi = \sum_{i=1}^{p} \frac{n_i}{n} \pi_i, \text{ where } n = \sum_{i=1}^{p} n_i.$$

In order to for the reference level, we need to write each conditional probability as a function of the reference level probability. The solution is tractable when link function is log function, $\pi_1 = \frac{n\pi}{n_1 + \sum_{i=2}^{p} n_i \exp(\beta_{i-1})}$, and identify function, $\pi_1 = \pi - \sum_{i=2}^{p} \frac{n_i}{n} \beta_{i-1}$. [TODO: provide the full-on derivations in the supporting information] When the link function becomes more complex, e.g., logistic function, the solution is not algebraically tractable. The closed form equations provides computational convenience when link function is the identify function or log function. [TODO: write a generalizes]

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.