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

Boyi Guo, Jacqueline Rudolph

2022-05-26

As the field of epidemiology evolves, there are growing interest to employ more computational approaches to solve analytic problems. Among them, simulation is one of the most accessible concept. 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 a basic statistics knowledge can simply the balance intercept problem and provide a generalized solution. In the rest of this article, we look into the balance intercept problem with different statistical lenses. Specifically, we summarize what statistical problem the balance intercept manifest/represent, how the original authors set up the problem in a reference coding system, and how to translate the problem into the effect coding scope which provides a much simple solution without algorithmic integration.

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.

This same problem was later revisited by Robertson, Steingrimsson, and Dahabreh (2021) who discovered that the analytic solution of the balance intercept analytically can produce inaccurate controlling of the marginal probability at the desired level. Instead, Robertson, Steingrimsson, and Dahabreh (2021) proposed a numeric solution to solve for the balance intercept for binomial simulation with a logistic link function. Later, Zivich and Ross (2022) did xxx for multi-level categorical variable. [TODO: add citence]

To better understand the balance intercept problem and the proposed simplification, we first introduce a basic statistics concept, coding scheme. A coding scheme describes how a categorical variable is enumerated in the regression system. Most commonly used coding schemes in analysis include 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. As the enumerations are different for the two schemes, they offer different interpretations of regression coefficients. (See example in Table ) 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 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. For more techinical details, we defer to [TODO: add textbook, probably INTRODUCING ANOVA and ANCOVA by andrew rutherford].

The balance intercept problem embeds in the reference coding system in the original proposal to control the marginal probability, even though the authors didn't mention explicitly. The intercept term in the reference coding scheme describes the conditional probability of the reference level. Hence, Setting the intercept term to the target marginal probability (referred to as the standard intercept) would fail to control the empirical marginal probability due to the theoretical discrepancy. The balance intercept is a derivation of the conditional probability of the reference level, and can be calculated with previous solutions. Nevertheless, the accuracy of the simulation study would be vulnerable to the quality of numeric analysis and programming efficiency.

| Levels | Referece | | | Effect | | |
|---|---|---|---|---|---|---|
| | $X_1$ | $X_2$ | Conditional Probability | $X_1$ | $X_2$ | Conditional Probability |
| Level 1 | 0 | 0 | $\beta_0$ | 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.

The balance intercept would be always unique as the degree of freedom is fixed.

The process of finding the balance intercept may not be necessary when the effect coding system, and hence greatly reduces the amount of numeric computation. The intercept term in the effect coding scheme describes the mean of the group means, which coincide with the marginal probability when the sample sizes are balanced across groups. In other words, the balance intercept problem does not exist for simulations with balanced designs. When the underlying design is not balanced, some simple arithmetic calculation is still needed to calculate the grand mean. However, this can be done easily with a closed form equation and applied to any data generating model of interest, following the mathematical definition of the intercept terms. Besides its straightforwardness, the proposed solution should not incur any knowledge burden as the effect coding is commonly introduced in introductory statistics classes. We demonstrate the simulation with the following toy example.

We simulate a data set of $n = 10000$ individuals whose binary outcome has an overall marginal probability of $\pi = 0.4$. The outcome is associated with a three-level covariate on the logit scale (in the generalized linear model framework, the link function is $g(x) = log(\frac{x}{1-x})$). The log odds ratios between the second level and the third level of the variable compared to the first level is $\beta_1 = 0.2$ and $\beta_2 = -0.1$ respectively. The prevalence of the three levels of the covariate among the sample is 0.6, 0.3, and 0.1 respectively. As the levels of the covariate are not balanced, we need to calculate the intercept term in the effect coding system. The overall goal is to calculate the mean of conditional probabilities in two steps with close form equations. The first step is to calculate the the conditional proabilities given the format of information you have; the second step is to take average of the the conditional proabilities over the number of levels, i.e. $\delta_0 = \frac{\sum_{j=1}^{p} \pi_j}{p}$. The first step is trivial if the conditional probability is given. Nevertheless, if the ratios or difference is given, we can calculate the conditional probabilities by first finding the reference level probability by solving

$$\pi = \sum_{j=1}^{p} \frac{n_j}{n} \pi_j$$

where $\pi_j = g^{-1}(g(\pi_1)/\beta_{j-1})$ for ratios and $\pi_j = g^{-1}(g(\pi_1) - \beta_{j-1})$ for differences. Given these numbers it is very easy to derive the coefficients via a linear system according to Table 1. With the given, we can simulate the outcome via a bernoulli distribution

$$Y \sim Bernoulli(g^{-1}(\delta_0 + \delta_1 X_1 + \delta_2 X_2)),$$

where $X_1$ and $X_2$ follows the effect coding suggested in Table 1. This data generating mechanism can be easily generalizable to the covariates with more levels, multiple covariates or statistical interaction of cavoriates by conceptualizing the 2-way contigency table, and treat each cell as one level of the covariate table. This would allow the simulations to address mediators, or colliders. It is also very easy to address continuous covariates by including a centered version of the covariates $(X - E(X))$. The implementation of the data generating process with an effect coding scheme is provided in the supporting information.

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.