Representative Norm Construction via Post-Stratification Weighting

John T. Kulas[1], Yang Yang[2], and & Mike Morris[3]

[1] Montclair State University

[2] China Select

[3] CPP Inc

Author Note

Correspondence concerning this article should be addressed to John T. Kulas, 250 Dickson Hall, Montclair State University, 1 Normal Ave., Montclair, NJ 07043. E-mail: kulasj@montclair.edu

Abstract

Large-scale testing and assessment operations benefit from opportunities to sample via stratification to construct norms that are representative of target populations (e.g., US workforce, college graduate, elementary school grade equivalent). We propose the application of a procedure most commonly retained in the context of sample remediation (post-stratification weighting) as a method to "build" rather than funnel norms. This approach is leveraged and then evaluated against "population" values via controlled simulation. Norms are traditionally created through stratified sampling of larger distributions of data. Not all vendors or publishers have this luxury - some do simply not have access to large samples of respondents. We propose an alternative strategy that zeros in on the same goal: representative normative samples.

Keywords: assessment norms, post stratification weighting, norm construction, comparative norms

Word count: X

Representative Norm Construction via Post-Stratification Weighting

"Norms" are distributions of psychological or educational assessment scores against which an individual examinee's absolute rating can be contrasted. Common normative metrics include standard scores (typically derivative of a z-score) and percentile ranks - both reflect an examinee's relative standing within the normative distribution (aka "norm"). The constituency of normative groups can be a marketable asset for assessment vendors. Having access to norms that include respondents representing populations of interest might, for example, entice a client to utilize an assessment. Similarly, one vendor's possession of, for example, workforce representative norms may be a deciding factor regarding consultative partnership if another vendor *does not* have access to workforce representative norms.

Assessment "owners" who have access to large representative samples of respondents have the luxury of tailoring their norms to a wide variety of desirable constituencies, and would therefore appear to hold advantage over smaller, boutique, or niche consultancies that may not have as great reach or opportunity to develop large norms bases. The OPQ32r, for example, in total possesses 92 unique static normative distributions culled from more than two-hundred thousand individuals, covering 24 languages and 37 countries/regions, as reported in its Norm Update Technical Documentation (SHL Group Ltd., 2011). These norms are further differentiated by general population, general work population, managerial and professional, senior managers, graduate, and industry specific norms. This norm library is an absolute competitive advantage over other local and international vendors.

Traditionally,[1] norms are developed in this manner - either simply collecting a very

[1] Perhaps more appropriately, "*according to textbooks*". There is at least one mention of the currently explored "building up" procedure that is acknowledged in the published literature (Holt, 1993). It is possible that vendors have been engaging in this creative form of norms creation, but the procedure has not migrated to the published literature.

large number of responses (e.g., GRE, ACT, SAT) or stratified random sampling from such a large number of responses (or, more rarely in the context of psychological assessment - a fully representative population frame) to obtain a normative group of desired constituency. In the current paper, we explore an alternative to these traditional approaches whereby representative norm groups are "built up" from smaller respondent samples. We do this via application of a procedure that is intended to align sample constituencies with population parameters in post-survey administration contexts: post-stratification weighting (see, for example, Kulas et al., 2018; Yang et al., 2017).

Although not considered a typical approach, there is at least one record of something similar being done within the educational assessment domain. Holt (1993) describes one SAT 8th edition reading comprehension subgroup norm as being created via weighting (because of the uniqueness of the desired normative group - age representative deaf and hard-of-hearing students). This 1993 application is the sole example of a similar-minded methodology that we were able to locate within the published literature. The current investigation explicitly proposes and then investigates the viability of a norms-building strategy for those interested in creating representative norms (perhaps without the luxury of expansive data collection).

A Brief Review of the Norms Development Process Implicated within the Measurement Literature

Kinnear and Sahraie (2002) describe the development of norms for an eye test, strategically sampling equal numbers of males and females from ages 5 to 79, although the final normative constituency ($N = 382$) was ultimately comprised of individuals who opted in (e.g., a non-probability sample). The Boston Naming Test (BNT) has many age, education (e.g., Tombaugh & Hubiey, 1997) and gender (e.g., Zec et al., 2007) stratified norms generated by various researchers. However, Hawkins and Bender (2002) review of available BNT norms revealed that only a few of them were adequately representative of

the population and most norms were skewed by highly educated subjects. The norms for the Job Descriptive Index and Job in General measures of job satisfaction have been recently revised via stratified random sampling drawn from an online panel (to represent the US working population on key variables, resulting in the first US overall norms and subgroup norms, Gillespie et al., 2016).

Need more plus more recent - it's possible that the process may need to be taken from textbooks or tech reports, current scale development articles don't seem to describe norms development as rigorously (maybe address in Discussion).

Although seeking representative samples from larger populations comprises the traditional approach to norm-building, alternatives have been sought. Much of this work seems to have peaked (and subsequently perhaps been abandoned) in the 1960's. Lord (1962) may have been one of the first to publicly disseminate an alternative strategy. He describes a procedure first investigated in a 1961 project for the Educational Testing Service (ETS, Lord, 1961) essentially sampling items instead of individuals (e.g., what provides a truer approximation of the population distribution - the entire measure administered to one respondent group or item parcels administered to different respondent groups?). In his first exploration, the "item sampling" approach was in fact deemed superior to the more traditional "respondent sampling" strategy, with Lord (1961) concluding that such a strategy may be enticing to educational assessment professionals because it is less disruptive (administering smaller item sets requires less time commitment). Plumlee (1964) also administered smaller item subsets to multiple samples of respondants. Both procedures administered smaller item subsets to smaller groups of respondents, and then aggregated from these multiple samples and item sets. The summary finding of each of these studies was that the item subgroup estimates provided better (closer to true normative) estimates than did sub-sample estimates. **will need to re-read these - current explanation isn't clear** The current investigation revisits these concepts, but with: 1) controlled simulations, and 2) an alternative sample

adjustment procedure (the investigations in the 1960's did not attempt weighting). Figure 1 communicates our intent: to evaluate the viability of an alternative (C) to the commonly executed (B) norms-development strategy.

Methods

The current exploration utilizes simulations to test the efficacy of the weighting procedure. First, we generate a "population" with known subgroup differences along a measured variable representing a scale score from a Psychological inventory (for context, we refer to this as a summary "engagement" rating per individual). Next, we *randomly sample* subsets from these populations and build (via post-stratification weighting) representative normative groups. As an index of how the procedure performs, we evaluate the normative distributions against the original population distributions, documenting which distributions (weighted versus unweighted) closer approximates the population engagement value.

Procedure

Workforce representation was identified via the Bureau of Labor Statistics' Labor Force Statistics from the Current Population Survey (Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity (Table 11), Statistics, Bureau of Labor, 2018). These are approximately 153,337,000 individuals aged 16 and older. The population percentages were specified at variable margin levels and these values are presented in the "Census Parameter" column of Table 1.

"Populations" of 10,000 individuals were constructed with ethnic distributions of values specified as normal, positively skewed, or negatively skewed (in three different *across-group* patterns; see Table 1 - this was done to evaluate the impact of different distributional forms within differently sized subgroups). Occupation and gender were then randomly assigned at census parameters (for example, the 1,000 other, 1,200 black, and 7,800 white respondents were randomly assigned gender and occupational characteristics for purposes of multi-strata rating [e.g., aka "raking"]). From these 10,000 records, a

random or stratified random[2] sampling was performed at n's of 100. The samplings were then raked using the *anesrake* (Pasek, 2018) package within *R*. Across the six experimental conditions, populations were specified, sampled from, and raked 10,000 times each (e.g., each of the 60,000 total simulations estimated a different simulated population).

We varied the distributional forms of the three ethic groups, with the distribution of scores (1 to 5) representing negative (1) to positive (5) attitudes. In condition 1, for example, "Whites" realized a positively skewed distribution, "blacks" were represented with a negatively skewed distribution, and "asians and others" exhibited a normally distributed range of attitudes (Figure 2 presents one population simulation as an example of a Condition 1 set of subgroup distributions).

To evaluate the "quality" of the random, stratified random, and weighted samples, we collected distributional values at seven different percentile locations: the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. These values were extracted from: 1) each simulated population, 2) each random or stratified random sampling, and 3) each weighted sample. Our criteria for "quality" of normative distribution was discrepancy from the population distributional values at each of the seven percentiles.

The other demographic specifications (occupation and gender) were randomly assigned to the 10,000 ethnic cases (e.g., Table 1 percentages were specified and randomly assigned to cases).

Results

We present the majority of our results visually in Figures 3 through 8. Because of the "busy-ness" of the information within these figures, we present one visual

[2] These terms are a bit misleading in the context of the current simulations - all samplings were random, but "stratified random" was approximated in conditions 1 through 3 (see Table 1) whereas differentially stratified was specified in conditions 4 through 6 (where we were interested in evaluating the efficacy of the weighting procedure with known biased [in terms of representative constituency] samples).

representation of each experimental condition, although every figure conveys the same information: 1) the "population" distribution[3] (1st row), 2) discrepancies between population and unweighted sample distribution values at each of the 7 retained percentiles (2nd row), and 3) discrepancies between population and *weighted* sample distribution values (3rd row). The x-axis scales on these discrepancy graphs are intentionally asymmetric toward negative discrepancies because our most extreme outliers were negative (although not possible to visually see within the histograms). Negative values indicate a "larger" population than sample value. Conditions 1 through 3 demonstrate that, regardless of distributional form across the small or large constituent groups, if sampling can be reasonably characterized as stratified (e.g., if the sample has representative constituencies), there is no need or indeed added value in weighting the sample (this is admittedly not an exceedingly earth-shattering finding, but if the sample already approximates the population, there is in fact no need for weighting).

Conditions 4 through 6, however (Figures 6 through 8) demonstrate that weighted normative distributions do far exceed the quality of randomly sampled distributions if those random samplings result in subgroup proportional inconsistencies (e.g., not representative constituency). The narrow nature of these distributions of discrepancies (regardless of whether the distribution is weighted or unweighted) shows that the samplings result in largely similar values across the 10,000 simulations. The *location* of the distributions, however, highlights the advantage of weighting: the unweighted samples deviate from centering on a value of zero (zero indicates a population-sample match at the percentile of focus). Discrepancy distributions centering on non-zero values represent *bias* in the normative values. In all three conditions where bias was present (at at least one percentile location), performing the weighting corrected this bias.

---

[3] These are labeled "frames" within Figures 4 through 9 as these simulated distributions could be viewed as representations of either true populations or population frames.

Discussion

Post-stratification weighting does appear to hold some promise as a norms-building strategy. Particularly in situations where frames (or even merely larger samples) are not proportionally representative of populations, the weighting procedure effectively reproduces the output of stratified sampling: producing a proportionally representative sample distribution. We attempted to introduce large constituent group differences in distributional forms (within the "populations") as a challenge to the weighting procedure, but, as long as strata are properly identified (along which value distributions may vary), the resulting weighted distribution does do a very good job at representing the population distribution (much better than does the unweighted normative distribution). Future investigations should introduce additional discrepancies (both distributional as well as subgroup proportional) to further test the boundaries of the procedure. Attempts to "construct" representative norms with real-world data are also warrented.

Scale development papers should devote some attention to their norms creation - speculate that much of this is located in technical reports, but those are not as commonly available to researchers.

References

Gillespie, M. A., Balzer, W. K., Brodke, M. H., Garza, M., Gerbec, E. N., Gillespie, J. Z., Gopalkrishnan, P., Lengyel, J. S., Sliter, K. A., Sliter, M. T., et al. (2016). Normative measurement of job satisfaction in the US. *Journal of Managerial Psychology*, *31*(2), 516–536.

Hawkins, K. A., & Bender, S. (2002). Norms and the relationship of boston naming test performance to vocabulary and education: A review. *Aphasiology*, *16*(12), 1143–1153.

Holt, J. A. (1993). Stanford achievement test—8th edition: Reading comprehension subgroup results. *American Annals of the Deaf*, *138*(2), 172–175.

Kinnear, P. R., & Sahraie, A. (2002). New farnsworth-munsell 100 hue test norms of normal observers for each year of age 5–22 and for age decades 30–70. *The British Journal of Ophthalmology*, *86*(12), 1408–1411. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1771429/

Kulas, J. T., Robinson, D. H., Smith, J. A., & Kellar, D. Z. (2018). Post-stratification weighting in organizational surveys: A cross-disciplinary tutorial. *Human Resource Management*.

Lord, F. M. (1961). Estimating norms by item sampling. *ETS Research Bulletin Series*, *1961*(1), i–13.

Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement*, *22*(2), 259–267. https://doi.org/10.1177/001316446202200202

Pasek, J. (2018). *Anesrake: ANES raking implementation*. https://CRAN.R-project.org/package=anesrake

Plumlee, L. B. (1964). Estimating means and standard deviations from partial data—an empirical check on lord's item sampling technique. *Educational and Psychological Measurement*, *24*(3), 623–630.

SHL Group Ltd. (2011). *OPQ32r norm update technical documentation* [Technical Report]. SHL Group Ltd.

Statistics, Bureau of Labor. (2018). *CPS tables*. https://www.bls.gov/cps/tables.htm

Tombaugh, T. N., & Hubiey, A. M. (1997). The 60-item boston naming test: Norms for cognitively intact adults aged 25 to 88 years. *Journal of Clinical and Experimental Neuropsychology*, *19*(6), 922–932.

Yang, Y., Kulas, J. T., & Robinson, D. R. (2017, April). *Nonresponse and sample weighting in organizational surveying*. Society for industrial and organizational psychology.

Zec, R. F., Burkett, N. R., Markwell, S. J., & Larsen, D. L. (2007). Normative data stratified for age, education, and gender on the boston naming test. *The Clinical Neuropsychologist*, *21*(4), 617–637. https://doi.org/10.1080/13854040701339356

Table 1

Variable	Level	Census Parameter	Cond 1	Cond 2	Cond 3	Sampling Frequency	Cond 4	Cond 5	Cond 6
Gender	Male	53%							
	Female	47%							
Ethnicity	White	78%	Pos Skew	Normal	Neg Skew	33%	Pos Skew	Normal	Neg Skew
	Black	12%	Neg Skew	Pos Skew	Normal	33%	Neg Skew	Pos Skew	Normal
	Asian or Other	10%	Normal	Neg Skew	Pos Skew	34%	Normal	Neg Skew	Pos Skew
Industry	Management	40%							
	Sales	22%							
	Service	17%							
	Production	12%							
	Natural Resources	9%							

Note. Cell-level constituencies (e.g., percentage white male managers) were not available via census data.

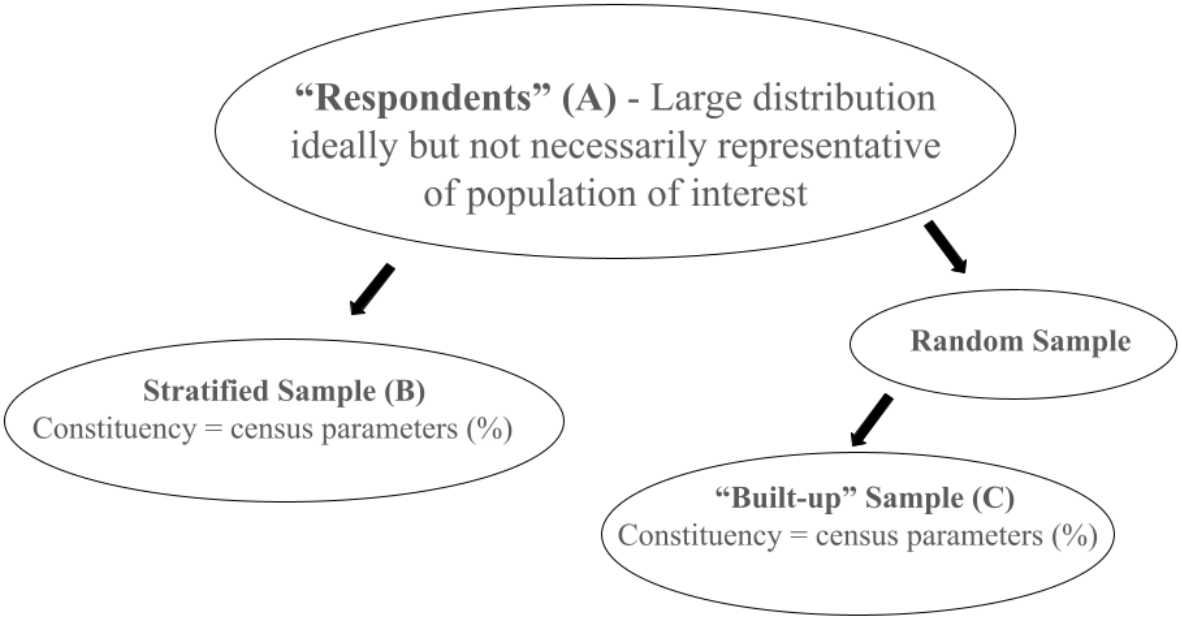

Figure 1

Relationship between Study 1 concepts of population (A), traditional norms (B), and proposed norm building procedure (C).

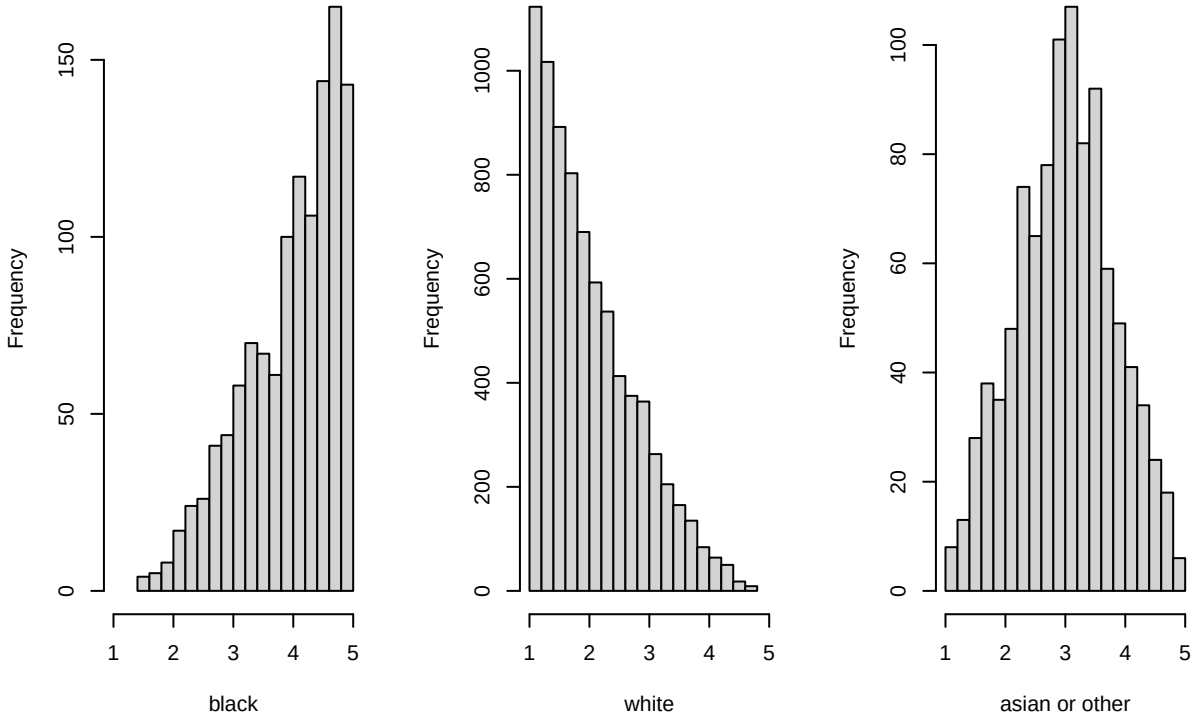

Figure 2

Example distributional forms (population specifications Condition 1).

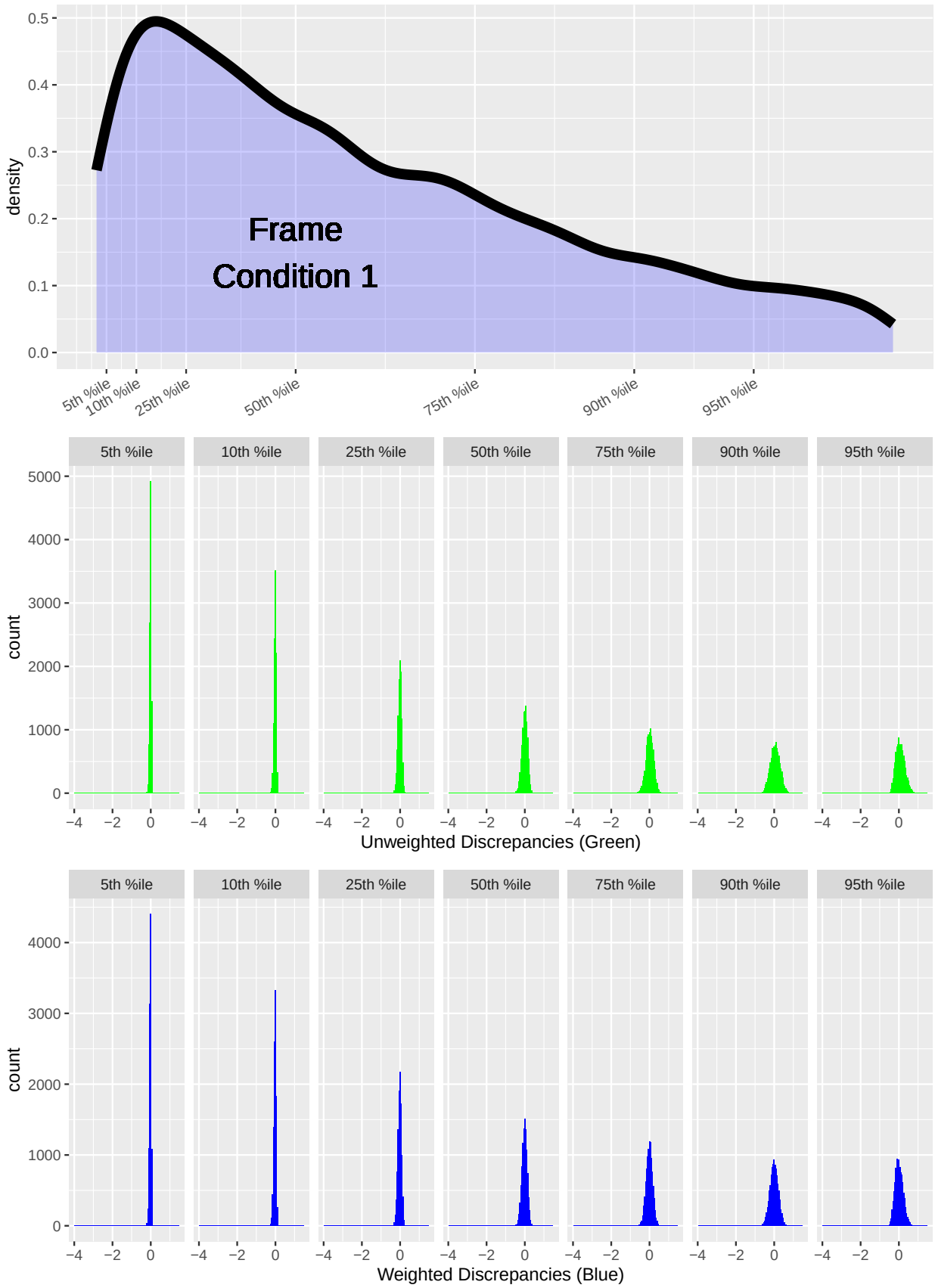

Figure 3

Population percentile locations with unweighted (approximately stratified), and weighted discrepancy distributions (Condition 1).

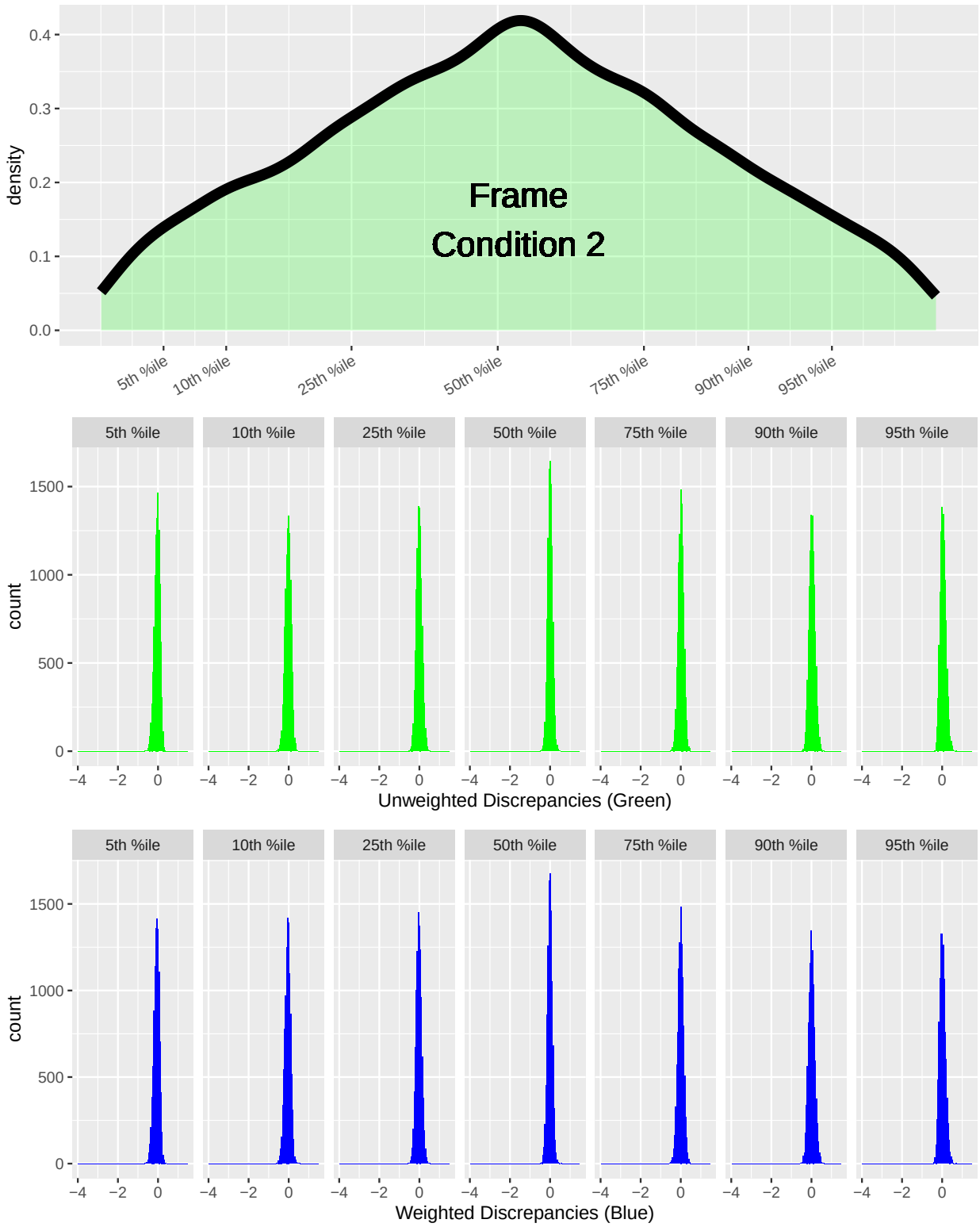

Figure 4

Population percentile locations with unweighted (approximately stratified), and weighted discrepancy distributions (Condition 2).

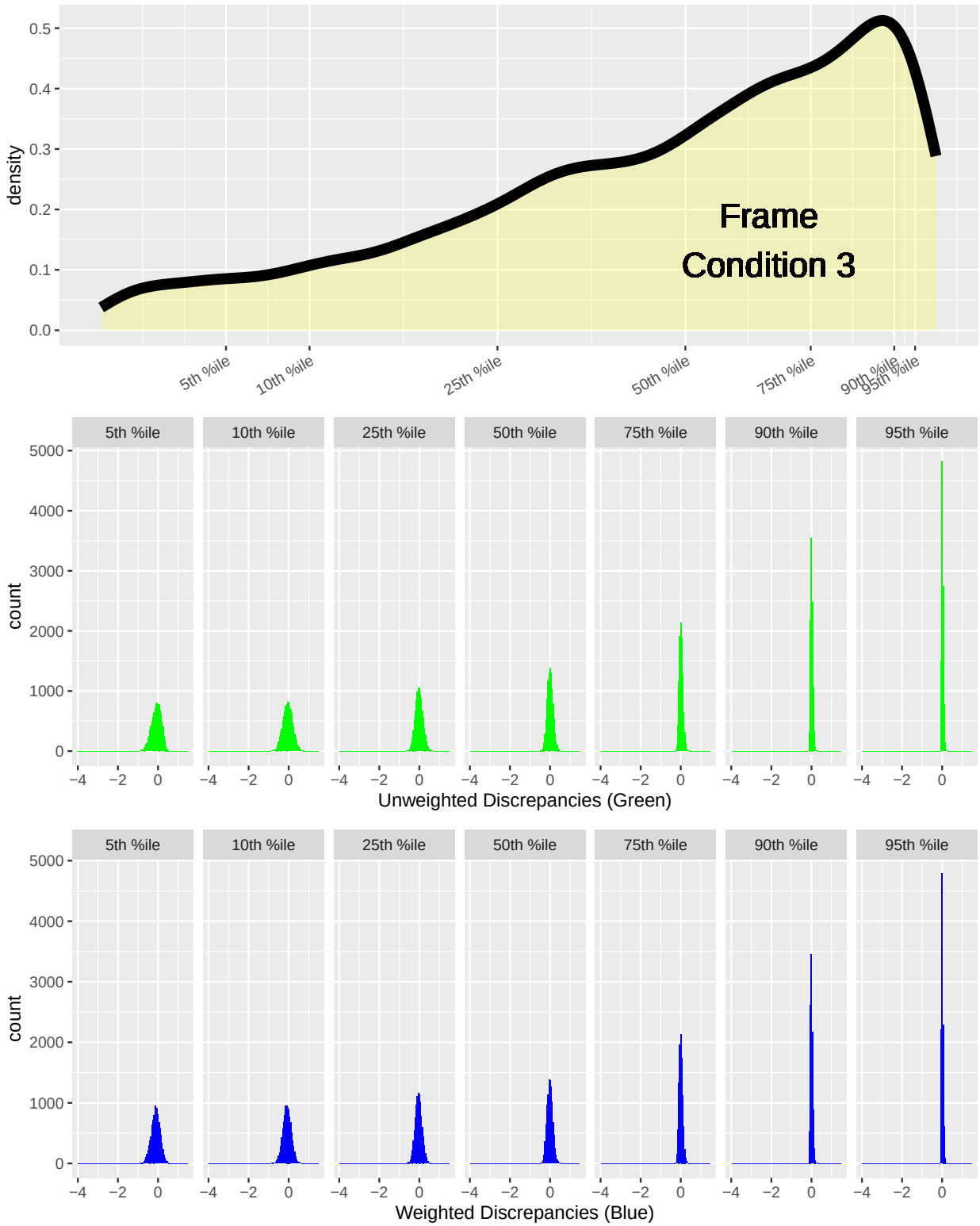

Figure 5

Population percentile locations with unweighted (approximately stratified), and weighted discrepancy distributions (Condition 3).

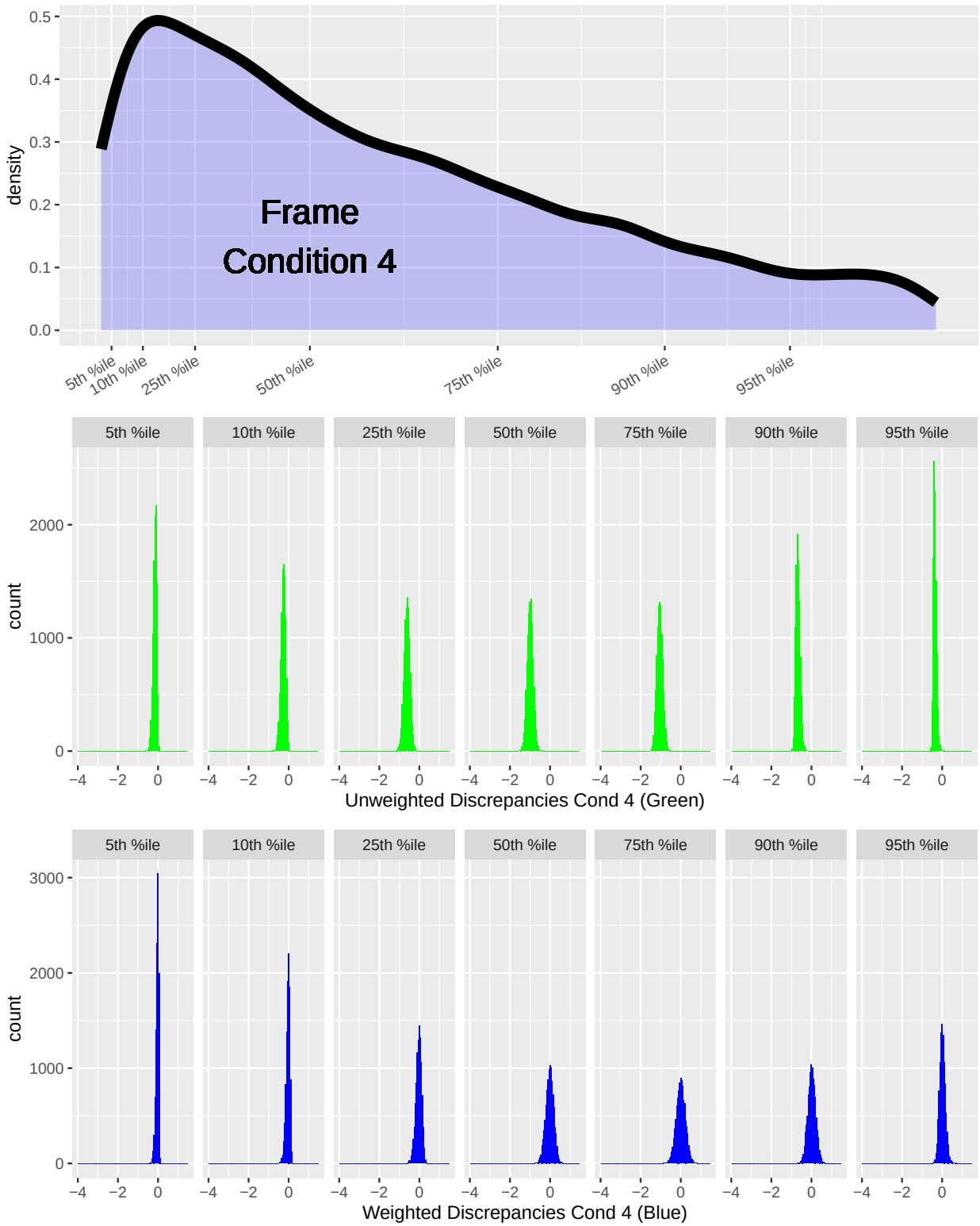

Figure 6

Population percentile locations with unweighted (disproportionate), and weighted discrepancy distributions (Condition 4).

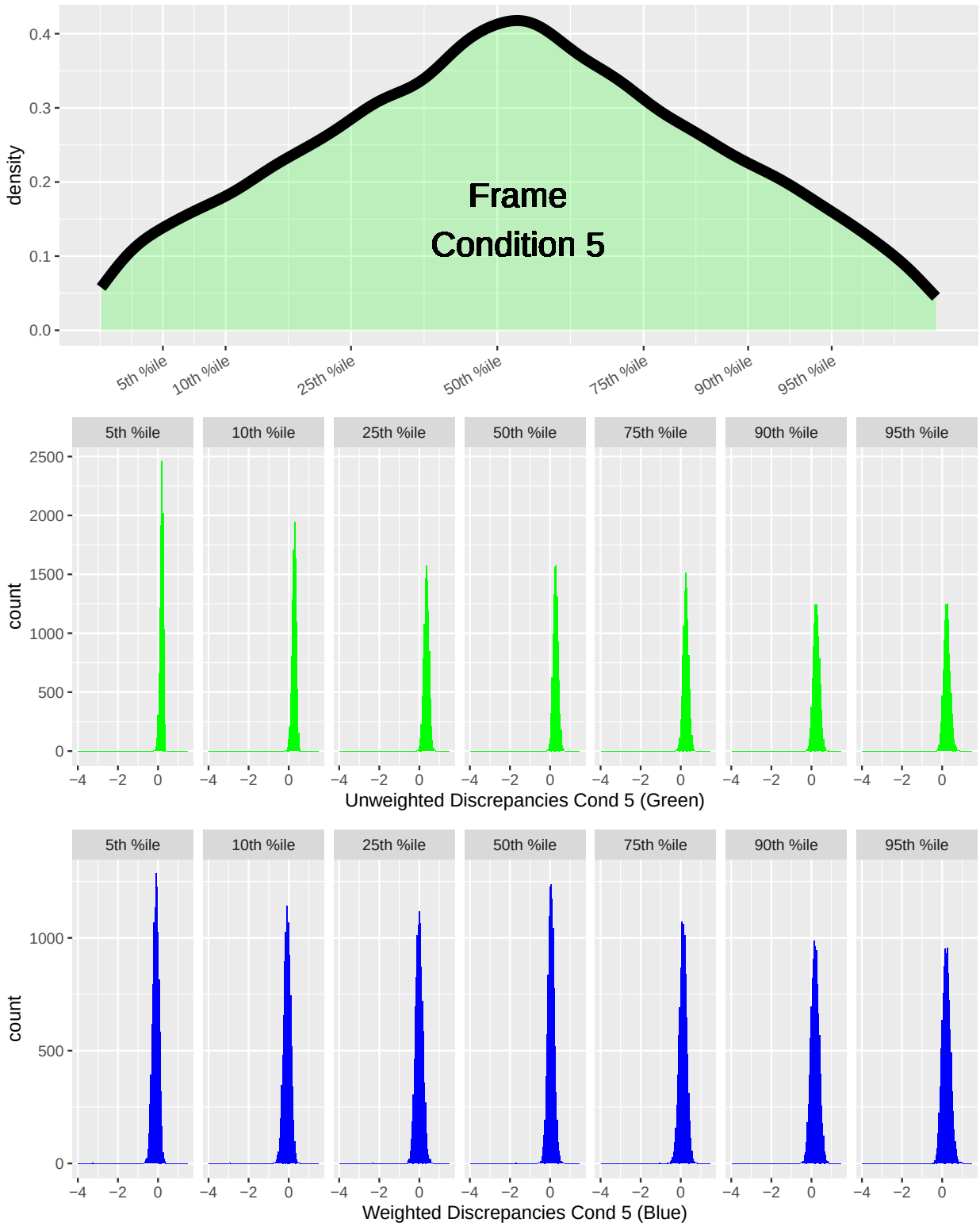

Figure 7

Population percentile locations with unweighted (disproportionate), and weighted discrepancy distributions (Condition 5).

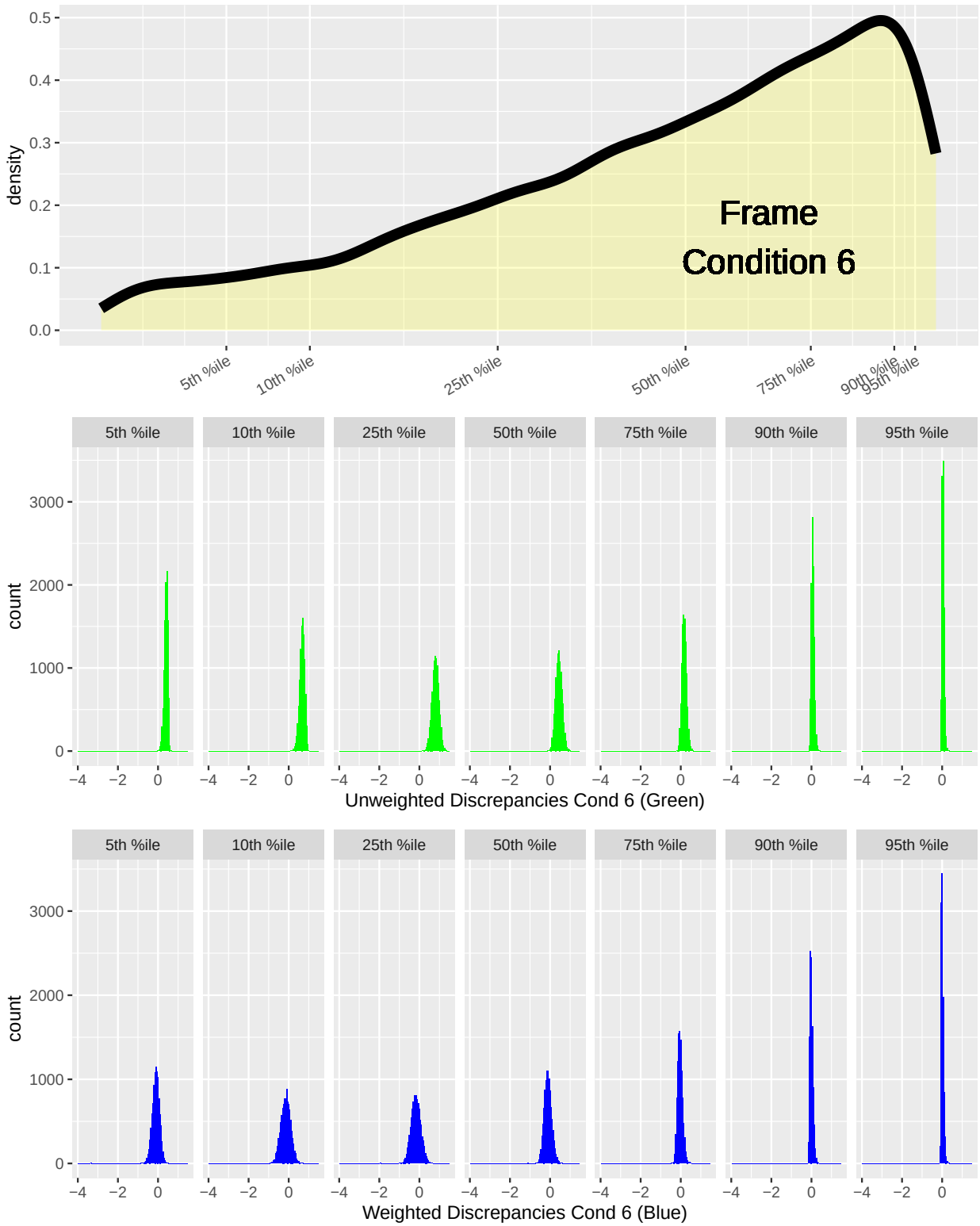

Figure 8

Population percentile locations with unweighted (disproportionate), and weighted discrepancy distributions (Condition 6).