Bootstrap sampling is comparable to continued sampling of a population of a known distribution.

Parker Hicks, HMGGP PhD Student

Abstract

In many cases, sampling an entire population to elucidate the true value for a given statistic is unfeasible. Bootstrap sampling offers a solution by estimating the distribution of a statistic based on a single sample. This study tests the comparability between bootstrap-sampled distributions and those obtained through continuous sampling of populations with known distributions. Here, we show that these two methods are comparable, further affirming that bootstrap sampling provides comparable results to continued sampling from a population with a known underlying distribution. This validation underscores the utility of bootstrap techniques in approximating true population statistics, offering a practical and reliable alternative in scenarios where complete population sampling is unattainable.

Publicly-available and reproducible code for these simulations is available at https://github.com/phicks22/BIOS-6611-midterm.

Keywords: boostrap, population sampling, PRS

Introduction

Drawing meaningful statistical inferences about populations is a fundamental challenge in all research disciplines. Traditional methods for this task often rely on stringent assumptions about the underlying population distribution. However, bootstrap sampling provides a technique that allows for the estimation of the true distribution of a statistic of a population without the need for assumptions about the underlying population. This makes bootstrap sampling valuable in situations where the population characteristics are not well understood. Nonetheless, bootstrap sampling is only an estimation of the true sample distribution. Thus, in this study, we test to see if the bootstrap sampling method for estimating the sample distribution of a population is comparable to *actually* performing continuous sampling from a population where the true distribution is known.

Motivating Problem

In the statistical genetics field, polygenic risk scores (PRS) are estimates of an individual's relative risk of developing a common disease, given their genetic make-up of genomic regions that have been associated with the disease from genome wide association studies (GWAS). Since there are common genetic differences between ancestral groups, PRS distributions are now commonly separated by genetic ancestry. Individuals that lie within the top 15th percentiles of their ancestral distributions are genetically most at risk for a given disease. Furthermore, populations of African ancestry typically exhibit lower PRS distributions compared to European populations given the lack of African representation in GWAS studies [1]. Consequently, there is typically a difference in the top 15th percentiles of PRS scores between the two ancestral populations.

Additionally, researchers occasionally employ bootstrap sampling techniques to estimate PRS distributions and statistics across ancestries [2]. Thus, we test if boostrapping is an accurate estimator for the mean difference in the upper 15th percentiles of PRS distributions, where our simulation parameters were specifically designed to replicate the differences of PRS distributions of African and European populations.

Simulation Set-up

PRS score distributions tend to follow a normal distribution where mean and standard deviations vary between ancestral groups for a given disease. Furthermore, populations of African ancestry typically exhibit lower PRS distributions compared to European populations given the lack of African representation in GWAS studies [2]. Thus, we simulate normal distributions for populations of European ancestry as $N(\mu = 0, \sigma^2 = 2)$ and $N(\mu = -1, \sigma^2 = 2)$ for populations of African ancestry (Figure 1).

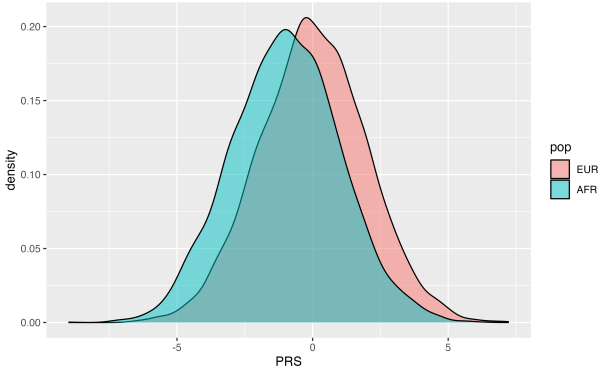

Figure 1: Common PRS distributions between

African and European populations.

We compared the statistic distribution of the mean differences of PRS scores from the upper 15th percentile of simulated European and African populations using two methods. **1)** Iteratively generating normal distributions and computing the mean difference at each iteration which is meant to represent repeated sampling of a population and **2)** boostrapping the mean difference statistic distribution from a single PRS distribution sample. Both methods were run at 10^4 iterations. Given the lack of representation of African populations in GWAS studies, we set our sample sizes at $n = 10,000$ for European populations. and $n = 5,000$ for African populations.

Results

Overall, we show that the boostrap sampling and continuous sampling methods are comparable. Summary statistics for both statistical distributions with the mean (μ), standard error (SE), and variance (σ^2) for both methods are provided in Table 1. Bias for the continued sampling method is not provided since this sample distribution represents the distribution of the true population.

The bootstrap mean, with a 95% confidence interval of 0.96 ± (0.885, 1.042), demonstrated a low bias of -0.036, thus reinforcing its reliability in estimating population parameters and robust estimate of the true population mean. There was a slight difference of 0.04 in the centers of both distributions. However, we acknowledge that the center of a bootstrap distribution does not accurately reflect the true center of a sampling distribution, thus this discrepancy should not be considered in our comparison.

Table 1: Distribution summary statistics

Statistic	**Cont. Samp.**	**Boostrap**
μ	1.0	0.96
SE	0.058	0.041
σ^2	0.003	0.003
$Bias$	-	-0.036

We also observed a difference in the skews of the two sampling distributions. Our continuous method exhibited symmetry, while the bootstrap method was slightly left-skewed. However, both distributions exhibited comparable patterns of spread, showing similarity in their variance estimates (Table 1)(Figures 2 & 3).

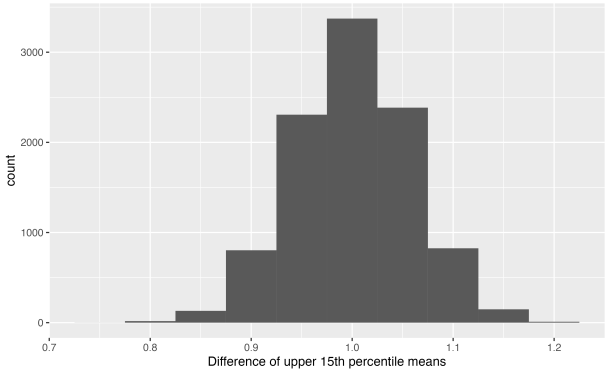

Figure 2: Histogram of mean differences from continued sampling.

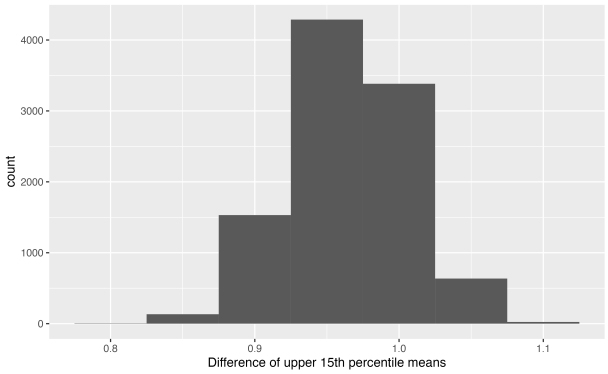

Figure 3: Histogram of mean differences from bootstrap sampling.

Conclusion

Despite the disparities of center and skew, the congruence in spread, relative similarity of the centers, and low bias of the bootstrap estimator signifies a comparable estimation capability between the two methods. Overall, we show that the sampling distribution for the mean difference between the upper 15th percentiles of two normal distributions generated by bootstrap sampling is comparable to computing the sampling distribution through continuous sampling of the true population.

References

1. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet. 2020;107: 788–789.
2. Song, S., Jiang, W., Hou, L., & Zhao, H. (2020). Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. PLoS computational biology, 16(2), e1007565.