

# Problem Set 3

Cem Kozanoglu

October 14, 2025

Collaborators: Wooyong Park, Roberto Gonzalez-Tellez, Hanniel Ho and Aileen Wu.

## 1 Problem 1

All simulations were performed in Python, the code is available as a python notebook with the results printed out. I used 10,000 bootstrap iterations for each estimate.

### 1.1 Part A: Plain Vanilla bootstrap

For the vanilla bootstrap, I resampled the dataset with replacement 10,000 times, calculated the difference in means for each sample, and then calculated the 5th and 95th percentiles of the resulting distribution to get a 90% confidence interval. The results were as follows:

Variable	Point Estimate	90% Confidence Interval
Earnings Year 1	1.136	(0.915, 1.357)
Earnings Year 4	1.232	(0.826, 1.639)

Table 1: Classical Bootstrap Results

We can reject the null hypothesis with a significance level of 90% that the training program has no effect on earnings at both year 1 and year 4, as the confidence intervals do not include zero.

## 1.2 Part B: Bayesian bootstrap

For the Bayesian bootstrap, I assigned weights to each observation drawn from a Dirichlet distribution with parameters all equal to 1 (which is equivalent to drawing from a uniform distribution over the simplex). I then calculated the weighted difference in means for each sample, and calculated the 5th and 95th percentiles of the resulting distribution to get a 90% confidence interval. The results were as follows:

Variable	Point Estimate	90% Confidence Interval
Earnings Year 1	1.136	(0.917, 1.356)
Earnings Year 4	1.232	(0.816, 1.648)

Table 2: Bayesian Bootstrap Results

We can also reject the null hypothesis here with a significance level of 90% that the training program has no effect on earnings at both year 1 and year 4, as the confidence intervals do not include zero.

## 1.3 Part C: t-statistic bootstrap

For the t-statistic bootstrap, I calculated the standard error of the difference in means for each bootstrap sample, and then calculated the t-statistic for each sample. I then found the 5th and 95th percentiles of the resulting t-statistic distribution to get a 90% confidence interval, according to formula:

$$(\hat{\tau} - t_{0.95} \cdot SE(\hat{\tau}), \hat{\tau} - t_{0.05} \cdot SE(\hat{\tau}))$$

The results were as follows:

Variable	Point Estimate	90% Confidence Interval
Earnings Year 1	1.136	(0.912, 1.355)
Earnings Year 4	1.232	(0.806, 1.626)

Table 3: Bootstrap T-Statistic Results

We can also reject the null hypothesis here with a significance level of 90% that the training program has no effect on earnings at both year 1 and year 4, as the confidence intervals do not include zero.

## 1.4 Part D: Subsampling

For the subsampling without replacement, I randomly selected  $\sqrt{N}$  (in this case 73) of the dataset without replacement 10,000 times, calculated the difference in means for each sample, and then calculated the 5th and 95th percentiles of the resulting distribution to get a 90% confidence interval. The results were as follows:

Variable	Point Estimate	90% Confidence Interval
Earnings Year 1	1.136	(0.911, 1.361)
Earnings Year 4	1.232	(0.811, 1.654)

Table 4: Subsampling without Replacement Results

We can also reject the null hypothesis here with a significance level of 90% that the training program has no effect on earnings at both year 1 and year 4, as the confidence intervals do not include zero.

## 1.5 Part E: Stratified Bayesian bootstrap

Stratifying the dataset by highschool completion can help us improve the variance estimate of our treatment effect because highschool completion is likely correlated with earnings. By combining this with a feasible bootstrap method, we can get a fairly good confidence interval for our point estimate.

In this case I wanted to use a Bayesian Bootstrap approach for the following reasons: (1) Bayesian bootstrap avoids problems that come with typical resampling such as duplicate observations in a sample or the lack of some observations in the sample by making the weights continuous rather than discrete, (2) Bayesian bootstrap is more robust to outliers, (3) Bayesian bootstrap can be easily extended to a stratified sampling

framework, which is what I wanted to do in this case since I wanted to stratify by high-school completion status, and (4) Bayesian methods in general have some nice properties such as incorporating prior information and providing a full posterior distribution of the parameter of interest.

Besides all these statistical benefits, I personally like Bayesian methods within statistics and want to use them whenever I can, so some personal preference may have clouded my otherwise perfectly objective judgement. This is particularly important to note considering that the high sample size means that all the Bootstrap methods we discussed should yield very similar results.

To calculate my stratified Bayesian bootstrap estimates, I separated my dataset into two populations: the highschool population and the no-highschool population. The full process is enumerated below:

1. Split the dataset into two strata based on high school completion status
2. For each stratum, perform Bayesian bootstrap sampling (10,000 iterations)
3. Calculate the treatment effect (difference in means) for each stratum separately
4. Combine the stratum-specific estimates using weighted averages based on sample proportions
5. Calculate the combined variance using the weighted sum of stratum-specific variances
6. Construct confidence intervals using the combined estimates and variances

I calculated the difference in means for each population independently and combined them using the formula below:

$$\hat{\tau} = \frac{n_{hs}}{n} \cdot \hat{\tau}_{hs} + \frac{n_{nohs}}{n} \cdot \hat{\tau}_{nohs}$$

Where  $\hat{\tau}$  is the combined difference in means,  $n_{hs}$  is the number of observations in the highschool population,  $n_{nohs}$  is the number of observations in the no-highschool population,  $n$  is the total number of observations,  $\hat{\tau}_{hs}$  is the difference in means for

the highschool population, and  $\hat{\tau}_{nohs}$  is the difference in means for the no-highschool population.

I similarly calculated the variance of the combined estimate using the formula below:

$$\widehat{Var}(\hat{\tau}) = \left(\frac{n_{hs}}{n}\right)^2 \cdot \widehat{Var}(\hat{\tau}_{hs}) + \left(\frac{n_{nohs}}{n}\right)^2 \cdot \widehat{Var}(\hat{\tau}_{nohs})$$

Where  $\widehat{Var}(\hat{\tau})$  is the variance of the combined estimate,  $\widehat{Var}(\hat{\tau}_{hs})$  is the variance of the difference in means for the highschool population, and  $\widehat{Var}(\hat{\tau}_{nohs})$  is the variance of the difference in means for the no-highschool population.

The results are as follows:

### Earnings Year 1

Stratification	Point Estimate	95% Confidence Interval
High School	1.500	(1.100, 1.900)
No High School	0.737	(0.413, 1.061)
Combined	1.136	(0.876, 1.396)

Table 5: Stratified Bayesian Bootstrap Results - Earnings Year 1

### Earnings Year 4

Stratification	Point Estimate	95% Confidence Interval
High School	1.480	(0.669, 2.291)
No High School	0.960	(0.479, 1.442)
Combined	1.232	(0.750, 1.715)

Table 6: Stratified Bayesian Bootstrap Results - Earnings Year 4

We observe that the stratification actually ended up increasing the variance of our estimates, as the confidence intervals are slightly wider than those from the non-stratified Bayesian bootstrap. This is likely because highschool completion status is not as strongly correlated with earnings as we might have expected, and the benefits from stratification are overwhelmed by the downsides (such as having to estimate two separate variances which adds more overall uncertainty). The intricacies of this are discussed in 2 (a) below.

## 2 Problem 2

### 2.1 Part A

The right estimator depends on the context of the problem, particularly because we are applying this during analysis instead of stratifying pre-experiment. Since the groups are unbalanced, accounting for the stratification could either increase the variance if the correlation between the stratification variable and the outcome is low, or decrease the variance if the correlation is high. If we have reason to believe that the stratification variable is correlated with the outcome, we should use the stratified estimator. If we have no reason to believe that, we should use the simple difference-in-means estimator.

If we take the program above for example, it's likely that younger people will be able to take better advantage of the training's human capital improvements, so age is likely correlated with the outcome. In this case, we should use the stratified estimator to reduce variance and correct for the sample size imbalance.

### 2.2 Part B

We want to show that the stratified estimator is unbiased for the average treatment effect. We will first show that the simple difference-in-means estimator is unbiased, and then extend that logic to the stratified estimator.

Let  $W_i$  be the treatment indicator for unit  $i$ , and let  $Y_i(1)$  and  $Y_i(0)$  be the potential outcomes. The observed outcome is  $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$ . Let  $N_1 = \sum_{i=1}^N W_i$  and  $N_0 = \sum_{i=1}^N (1 - W_i)$  be the fixed number of treated and control units. The difference-in-means estimator is:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^N W_i Y_i - \frac{1}{N_0} \sum_{i=1}^N (1 - W_i) Y_i$$

Its expectation is:

$$\begin{aligned}
E[\hat{\tau}] &= E \left[ \frac{1}{N_1} \sum_{i=1}^N W_i Y_i(1) - \frac{1}{N_0} \sum_{i=1}^N (1 - W_i) Y_i(0) \right] \\
&= \frac{1}{N_1} \sum_{i=1}^N E[W_i] Y_i(1) - \frac{1}{N_0} \sum_{i=1}^N E[1 - W_i] Y_i(0) \\
&= \frac{1}{N_1} \sum_{i=1}^N \frac{N_1}{N} Y_i(1) - \frac{1}{N_0} \sum_{i=1}^N \frac{N_0}{N} Y_i(0) \\
&= \frac{1}{N} \sum_{i=1}^N Y_i(1) - \frac{1}{N} \sum_{i=1}^N Y_i(0) \\
&= \bar{Y}(1) - \bar{Y}(0) = \tau
\end{aligned}$$

Now we extend this logic to the stratified estimator.

Let the population be partitioned into  $G$  strata, where stratum  $g$  has  $N_g$  units. Within each stratum  $g$ ,  $n_{g1}$  units are randomly assigned to treatment and  $n_{g0}$  to control. The stratified estimator  $\hat{\tau}_S$  is defined as:

$$\hat{\tau}_S = \sum_{g=1}^G \frac{N_g}{N} \hat{\tau}_g \quad \text{where} \quad \hat{\tau}_g = \frac{1}{n_{g1}} \sum_{i \in g} W_i Y_i - \frac{1}{n_{g0}} \sum_{i \in g} (1 - W_i) Y_i$$

Following the logic of the simple estimator, the expectation of the within-stratum estimator is the average treatment effect within that stratum:

$$E[\hat{\tau}_g] = \bar{Y}_g(1) - \bar{Y}_g(0) \equiv \tau_g$$

The expectation of the stratified estimator is then:

$$\begin{aligned}
E[\hat{\tau}_S] &= E \left[ \sum_{g=1}^G \frac{N_g}{N} \hat{\tau}_g \right] \\
&= \sum_{g=1}^G \frac{N_g}{N} E[\hat{\tau}_g] \\
&= \sum_{g=1}^G \frac{N_g}{N} \tau_g \\
&= \sum_{g=1}^G \frac{N_g}{N} \left( \frac{1}{N_g} \sum_{i \in g} Y_i(1) - \frac{1}{N_g} \sum_{i \in g} Y_i(0) \right) \\
&= \sum_{g=1}^G \frac{1}{N} \sum_{i \in g} (Y_i(1) - Y_i(0)) \\
&= \frac{1}{N} \sum_{g=1}^G \sum_{i \in g} (Y_i(1) - Y_i(0)) \\
&= \frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0)) \\
&= \bar{Y}(1) - \bar{Y}(0) = \tau
\end{aligned}$$