Submitted to the Annals of the Department of Statistics and Operations Research

THE BOOTSTRAP FOR CAUSAL INFERENCE WITH AN APPLICATION TO MARRIAGE MARKET INCENTIVES

By Gregor Steiner

University of Vienna

The causal bootstrap is a modification of the classical bootstrap that accounts for uncertainty in the assignment of treatment. This allows for inference on causal parameters. This paper outlines the procedure for two-sided inference on the Average Treatment Effect. This is applied to the problem of marriage market incentives. I find evidence that women signal less professional ambition when their preferences are observed by potential romantic partners.

1. Introduction. The bootstrap was originally proposed by Efron (1982) and is frequently used in situations where standard parametric inference is difficult or even impossible. Imbens and Menzel (2021) suggest a modified version of the bootstrap for causal inference problems.

In causal inference, there are generally two sources of uncertainty: sampling uncertainty and design uncertainty. Sampling uncertainty stems from the stochastic nature of the sampling process, while design uncertainty refers to the stochastic nature of the treatment assignment. The latter is somewhat unique to situations where the interest lies on the causal effect of some treatment and is not present in classical statistics. That is why the classical bootstrap needs to be modified in order to account for both sources of uncertainty.

In this paper, constructing confidence intervals for the Average Treatment Effect (ATE) based on Imbens and Menzel (2021) is extensively discussed. As an illustrative example, I then apply the method to marriage market incentives: Men tend to favor less ambitious romantic partners. This creates a trade-off for women: Actions that lead to professional success may lead to failure in the marriage market. Utilizing data by Bursztyn, Fujiwara and Pallais (2017), I test whether women are more cautious about signaling professional ambition when potential partners observe their preferences.

Section 2 introduces the setting and some notation, section 3 explains the causal bootstrap, section 4 applies it to the marriage market incentives problem, and section 5 concludes.

2. Setting. This paper focuses on situations of the following form: A sample of n units is either exposed to a treatment (treatment group) or not (control group), and we are interested in the causal effect of said treatment for the entire population of $N \geq n$ units. That is, we observe an outcome Y_i and a binary variable $D_i \in \{0, 1\}$ indicating treatment for $i = 1, \ldots, n$. In such situations we face two types

MSC 2010 subject classifications: 62D05, 62D20, 62G09
Keywords and phrases: causal inference, bootstrap

of uncertainty: sampling and design uncertainty. Sampling uncertainty stems from the fact that we only observe $n \leq N$ units, while design uncertainty refers to the stochastic nature of treatment assignment.

Using Rubin's potential outcome framework (Rubin, 2005), we denote the outcome as $Y(1)$ if treated and as $Y(0)$ otherwise. The actual outcome Y_i is simply

$$Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0) = \begin{cases} Y_i(1), & D_i = 1 \\ Y_i(0), & D_i = 0 \end{cases}.$$

For some unit i the unit-causal effect is $Y_i(1) - Y_i(0)$, but only one of these is actually observed. Therefore, the unit-causal effect cannot be identified. Holland (1986) calls this the fundamental problem of causal inference. However, we can estimate the Average Treatment Effect (ATE), that is

$$\tau := ATE = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - Y_i(0).$$

Let n_0, n_1 denote the number of untreated and treated units respectively in a sample of $n = n_0 + n_1$ observations and let $R_i \in \{0, 1\}$ be an indicator whether the i-th population unit is included in the sample. Then the standard difference in means estimator for τ is

$$\widehat{\tau} = \frac{1}{n_1}\sum_{i=1}^{N} R_i D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{N} R_i(1 - D_i)Y_i = \bar{Y}_1 - \bar{Y}_0.$$

Let

$$S_D^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i(D) - \bar{Y}(D))^2,$$

where $\bar{Y}(D) = \frac{1}{N}\sum_{i=1}^{N} Y_i(D)$ for $D \in \{0, 1\}$, and

$$S_{01}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i(1) - Y_i(0) - \tau).$$

Then the true variance of $\widehat{\tau}$ is

$$\mathbb{V}ar(\widehat{\tau}) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{01}^2}{N}.$$

Traditional estimators ignored the S_{01}^2 term, which leads to overestimation of the true variance. Aronow, Green and Lee (2014) proposed an estimator based on the lower bound:

$$\widehat{\mathbb{V}ar}_{AGL} = \frac{\widehat{S}_0^2}{n_0} + \frac{\widehat{S}_1^2}{n_1} - \frac{\widehat{S}_{01}^2}{N},$$

where $\widehat{S}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{N} R_i D_i Y_i$, $\widehat{S}_0^2 = \frac{1}{n_0 - 1}\sum_{i=1}^{N} R_i(1 - D_i)Y_i$, and $\widehat{S}_{01}^2$ estimates the lower bound of S_{01}^2 (see Aronow, Green and Lee, 2014, section 3). Therefore, $\widehat{\mathbb{V}ar}_{AGL}$ estimates the upper bound of $\mathbb{V}ar(\widehat{\tau})$.

3. The causal bootstrap. This section introduces the causal bootstrap proposed by Imbens and Menzel (2021) and closely follows their sections 2, 3, and 5.

3.1. *Preliminaries and assumptions.* The aim of the causal bootstrap is to account for the stochastic nature of the treatment assignment. This takes us back to the potential outcome problem. For each observation we only observe one potential outcome. A natural approach is to impute the missing potential outcome. To do this, it is useful to start with the population joint distribution function of the pair of potential outcomes,

$$F_{01}^{p}(y_0, y_1) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}.$$

Unfortunately, there is no consistent estimator for F_{01}^{p}, but we can express it as a function of the marginal distribution functions,

$$F_{01}^{p}(y_0, y_1) = C(F_0^{p}(y_0), F_1^{p}(y_1)),$$

where $C : [0,1]^2 \to [0,1]$ is non-decreasing in both arguments. Such a copula C is guaranteed to exist by Sklar's theorem, but it need not be unique. Using C we can obtain an estimate $\widehat{F}_{01} := C(\widehat{F}_0, \widehat{F}_1)$. However, since C need not be unique, we want to choose it conservatively. If the first four moments of $F_0(y_0)$ and $F_1(y_1)$ are bounded, the isotone coupling,

$$C^{iso}(u, v) := \min(u, v),$$

is the least favorable coupling in the sense that it attains an upper bound on the asymptotic variance of $\widehat{\tau}$ (Imbens and Menzel, 2021, sections 2.4 and 2.5). The isotone coupling allows us to impute the missing potential outcomes as

$$Y_i(0) = \begin{cases} Y_i, & D_i = 0 \\ \widehat{F}_0^{-1}(\widehat{F}_1(Y_i)), & D_i = 1 \end{cases} \tag{3.1}$$

$$Y_i(1) = \begin{cases} \widehat{F}_1^{-1}(\widehat{F}_0(Y_i)), & D_i = 0 \\ Y_i, & D_i = 1 \end{cases}. \tag{3.2}$$

However, this is only valid for two-sided inference on the ATE. For other problems the appropriate least favorite coupling will likely be a different one.

Furthermore, we need a few assumptions:

1. The observed units are sampled at random from the population, that is $(Y_i(0), Y_i(1))$ are independent of R_i.
2. Treatment assignment is randomized, that is $(Y_i(0), Y_i(1))$ are independent of D_i[1].

[1]The approach can be extended to observational studies, if the outcome is independent of treatment assignment conditional on observable covariates (unconfoundedness). For details see Imbens and Menzel (2021) section 6.

3. The outcome $Y_i(D_i)$ for unit i is not affected by other units' treatment status, that is the Stable Unit Treatment Value Assumption (SUTVA) holds.

After establishing these preliminaries, we can finally outline the bootstrap procedure below.

3.2. *The bootstrap procedure.* Given a sample generated in accordance with the assumptions above, we are interested in the distribution of the t-ratio

$$T = \frac{\widehat{\tau} - \tau}{\widehat{\sigma}/\sqrt{n}},$$

where $\widehat{\tau}$ is the estimator of the ATE and $\widehat{\sigma} = \widehat{\sigma}(\widehat{F}_0, \widehat{F}_1)$ is the upper bound standard deviation.

The proposed algorithm consists of four main steps:

1. *Generating the population:* To generate an artificial population of size N, the n sampled observations are replicated according to the following procedure:
   (a) Split the sample in treated and untreated, i.e. $(Y_j^0)_{j=1}^{n_0}$ and $(Y_j^1)_{j=1}^{n_1}$ and order them in an increasing fashion.
   (b) Let $N_0 = \lceil \frac{n_0}{n} N \rceil$, $N_1 = N - N_0$ and include

$$M_j^0 := \lceil \frac{j}{n_0} N_0 \rceil - \lceil \frac{j-1}{n_0} N_0 \rceil$$

   copies of Y_j^0 with $D_j = 0$ for $j = 1, \ldots, n_0$, and

$$M_j^1 := \lceil \frac{j}{n_1} N_1 \rceil - \lceil \frac{j-1}{n_1} N_1 \rceil$$

   copies of Y_j^1 with $D_j = 1$ for $j = 1, \ldots, n_1$.

   Thus, we end up with a population $(Y_j, D_j)_{j=1}^N$, where the pairs of values are drawn from the sample according to the rule above.
2. *Imputing missing potential values:* The next step is imputing the missing potential outcomes. That is, for treated units we want an estimate of $Y_i(0)$ and for untreated units of $Y_i(1)$. As already discussed above, we simply use 3.1 and 3.2, where $\widehat{F}_0$ and $\widehat{F}_1$ are the empirical distribution functions estimated from the untreated and treated units respectively.
3. *Resampling*: For each bootstrap replication $b = 1, \ldots, B$, n pairs of $(Y_{ib}^*(0), Y_{ib}^*(1))$ are randomly drawn from the generated population. In the drawn bootstrap sample, treatment is randomly assigned with probability $p := \mathbb{P}(D_i = 1 | R_i = 1)$, that is the probability of treatment in the actual sample. Then, we can compute the bootstrap sample estimators $\widehat{\tau}_b^*$ and $\widehat{\sigma}_b^*$, and record the t-ratio

$$T_b^* = \frac{\widehat{\tau}_b^* - \widehat{\tau}}{\widehat{\sigma}_b^*/\sqrt{n}}.$$

4. *Confidence intervals:* Let $\widehat{G}(t) = \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}\{T_b^* \leq t\}$ be the empirical distribution function of the t-ratios obtained through resampling. Confidence intervals can then be constructed as

$$CI(1-\alpha) = \left[\widehat{\tau} - \frac{\widehat{\sigma}}{\sqrt{n}}\widehat{G}^{-1}(1-\alpha); \widehat{\tau} - \frac{\widehat{\sigma}}{\sqrt{n}}\widehat{G}^{-1}(\alpha)\right], \tag{3.3}$$

where $\widehat{G}^{-1}$ denotes the quantile function of the empirical distribution $\widehat{G}$.

3.3. *Asymptotics.* This section provides a characterization of the asymptotic behavior of the bootstrap procedure.

THEOREM 3.1. *Assume all the assumption from above hold, then $\widehat{\tau}$ and $\widehat{\sigma}$ are consistent for τ and σ.*

PROOF. By Glivenko-Cantelli, $\widehat{F}_0(y_0)$ and $\widehat{F}_1(y_1)$ converge almost surely to the true distribution functions, $F_0(y_0)$ and $F_1(y_1)$. This implies convergence in probability and therefore consistency.

Recall that τ and σ are functions of the marginal distribution functions, so they can be written as $\tau(F_0(y_0), F_1(y_1))$ and $\sigma(F_0(y_0), F_1(y_1))$. Under the assumption that the first four moments of $F_0(y_0)$ and $F_1(y_1)$ are bounded, they are even continous functions of $F_0(y_0)$ and $F_1(y_1)$ (see Imbens and Menzel, 2021, section 5.1). Thus, by the continous mapping theorem, the estimators converge in probability to the true parameters (or parameter bounds),

$$\widehat{\tau}(\widehat{F}_0(y_0), \widehat{F}_1(y_1)) \xrightarrow{p} \tau(F_0(y_0), F_1(y_1))$$

and

$$\widehat{\sigma}(\widehat{F}_0(y_0), \widehat{F}_1(y_1)) \xrightarrow{p} \sigma(F_0(y_0), F_1(y_1)),$$

that is they are consistent. □

THEOREM 3.2. *Assume all the assumptions from above hold, then*

$$\frac{\widehat{\tau} - \tau}{\widehat{\sigma}/\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(F_{01}^p)}{\sigma^2(F_0^p, F_1^p)}\right),$$

where $\sigma^2(F_{01}) = \lim_{n\to\infty} n\mathbb{V}ar_{F_{01}}(\widehat{\tau})$ and $\xrightarrow{d}$ denotes convergence in distribution.

The proof of this result is rather involved, the interested reader is referred to Imbens and Menzel (2021), Appendix C.

The next step is establishing a central limit theorem (CLT) for the bootstrap analogues. Let $\widehat{F}_0^*$ and $\widehat{F}_1^*$ denote the empirical distribution functions for a bootstrap replication for the untreated and treated observations respectively. Furthermore, let $\widehat{\tau}^* = \tau(\widehat{F}_0^*, \widehat{F}_1^*)$ and $\widehat{\sigma}^* = \sigma(\widehat{F}_0^*, \widehat{F}_1^*)$ denote the estimators' bootstrap analogues.

THEOREM 3.3. *Assume all the assumptions from above hold, then*

$$\frac{\widehat{\tau}^* - \widehat{\tau}}{\widehat{\sigma}^*/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

that is the bootstrapped t-ratio converges in distribution to a standard normal.

This result can be proven in a similar manner as 3.3 (see Imbens and Menzel, 2021, section 5 and Appendix C).

Finally, we can use these results to show that the confidence intervals obtained by the bootstrap procedure are indeed valid, that is the asymptotic coverage probability is at least as large as the nominal coverage probability.

THEOREM 3.4. *Under the assumptions from above, the confidence interval obtained by 3.3 is asymptotically valid, that is,*

$$\lim_{n\to\infty} \inf_{F_{01}^p} \mathbb{P}_{F_{01}^p}(\tau(F_{01}^p) \in CI(1-\alpha)) \geq 1-\alpha.$$

This result follows from the previous results and the definition of the variance bound (see Imbens and Menzel, 2021, Corollary 5.1). Thus, the bootstrap procedure is indeed asymptotically valid.

4. Application to marriage market incentives. In this section I apply the method outlined above to experimental data by Bursztyn, Fujiwara and Pallais (2017). Experimental data is convenient since the randomization assumption can be guaranteed to hold. This is rarely the case for observational data.

Men tend to favor romantic partners who are less ambitious than themselves. This creates a trade-off for women in the labor market: Actions that lead to professional success may lead to misfortune in the marriage market. Thus, single women may be more cautious about signaling professional ambition than their married counterparts. Bursztyn, Fujiwara and Pallais (2017) test this hypothesis in an experiment.

4.1. *Experimental design.* They study students in an elite MBA program. Their primary experiment took place during a career center on the first day of the MBA program, where students were asked to fill out a questionnaire that would be used to place them in a summer internship. Summer internships are incredibly important for MBA students: A substantial share of students ends up working for the firm where they did their summer internship. Therefore, this should be a high stakes event for students. Their answers influence what jobs the career center views as a good match for them.

Students randomly received one of two versions of the questionnaire: A public or a private one. The public version said that 'your' answers would be discussed in class. The private version, on the other hand, told the students that anonymized versions would be discussed in class. Students that received the public version are thus led to believe that their answers might be observable by their classmates. However, the

students did not know that there were two versions of the questionnaire or that they were part of a research project.

The heart of the questionnaire is a series of questions about their desired compensation, preferred hours of work per week, and willingness to travel for work. The women that received the public version (treatment) face a trade-off: Providing answers that present them favorably to the labor market may make them less attractive to their potential romantic partners. Thus, one would expect that the answers by treated single-women differ from those by their married or male classmates. At the same time, this effect should be less-pronounced or non-existent among the untreated.

4.2. *Data.* Bursztyn, Fujiwara and Pallais (2017) make their data available for replication purposes. Table 1 shows summary statistics for the variables of interest. In total, the sample includes 354 students, 346 of which provided information on their sex and marital status. Out of those 346, 60 students are female and single.

We refer to the public version of the questionnaire as treatment (i.e. $D_i = 1$ if student i received the public version), while the students who received the private version act as the control group. Out of the 60 single women, 29 received the treatment.

TABLE 1
Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
treatment	354						
... Private	176	49.7%					
... Public	178	50.3%					
male	354	0.681	0.467	0	0	1	1
single	346	0.474	0.5	0	0	1	1
class	346						
... Female & Non-single	51	14.7%					
... Female & Single	60	17.3%					
... Male & Non-single	131	37.9%					
... Male & Single	104	30.1%					
compensation	353	136.494	33.856	75	112.5	162.5	250
hours	353	51.66	8.779	40	45.5	55.5	80

To draw conclusions about the effect of observability on professional ambition, we focus on desired compensation and hours of work per week. The respective questions on the questionnaire are *What is your desired level of compensation?* and *How many hours per week are you willing to work on a regular basis?* Many students provided their answers as a range, e.g. 175-200 (in thousand USD) or 50-55 hours. Such answers are transformed by taking the mean of the range.

Table 2 shows the mean values by group and treatment. To be consistent with the hypothesis that women display less professional ambition when potential romantic partners can observe their preferences, the difference in means should be greater among single females. This seems to be the case for desired compensation. For willingness to work long hours, however, the difference between single and non-single women is rather small.

TABLE 2
Mean values for desired compensation and hours by type and treatment

	Compensation			Hours		
	Private	Public	Difference	Private	Public	Difference
Female & Non-single	134.72	132.29	-2.43	52.33	48.60	-3.73
Female & Single	130.65	113.79	-16.85	51.88	48.33	-3.54
Male & Non-single	140.67	133.73	-6.94	51.06	53.88	2.82
Male & Single	144.69	145.31	0.62	51.80	52.16	0.36

4.3. *Bootstrapping confidence intervals.* The next step is using the bootstrap procedure to construct confidence intervals for the ATE based on the difference in means estimator. In particular, I want to focus on desired compensation and hours of work among single women, since a significant effect for this group would confirm the hypothesis discussed by Bursztyn, Fujiwara and Pallais (2017). To the best of my knowledge, no R implementation for the causal bootstrap by Imbens and Menzel (2021) exists. That is why I write my own function for the bootstrap procedure. The code is available on my Github.

It is important to consider what population we generalize to, that is what N is in this case. It seems reasonable to take all single females in the US who are close in age to the students studied by Bursztyn, Fujiwara and Pallais (2017). According to the 2010 census, the female population in the age range 20-29 is about 21 million[2]. Unfortunately, information on marital status is not available. Using the proportion of singles in our sample, yields about 9.95 million single women in the population. Population sizes for the other three classes can be obtained analogously.

Table 3 shows the estimated 95% confidence intervals for all four classes. For desired compensation among single females, the upper bound is well below zero. This indicates a significant average effect of the treatment[3]. On average, the students demand between USD 2,600 and USD 26,000 less in compensation if they received the public version of the questionnaire. This makes the magnitude of the effect economically very relevant. However, for the desired number of weekly work hours, the confidence interval includes zero. There seems to be no significant effect for the desired hours of weekly work. Surprisingly, the upper bound of the non-single females is below zero, indicating a significant ATE for that group. In the male classes, there are no significant effects.

TABLE 3
95% Confidence Intervals for the ATE

	Compensation		Hours	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
Female & Non-single	-12.97	12.45	-5.28	-0.28
Female & Single	-25.77	-2.63	-4.28	2.72
Male & Non-single	-16.59	4.03	-4.24	1.47
Male & Single	-11.95	10.28	-4.22	1.62

[2]Population size by sex for selected age groups is available at the Census website.

[3]I refer to the effect as significant if the confidence interval does not include zero. This means that an inverted test used to obtain the confidence interval would reject the null hypothesis of a zero effect.

Thus, I can only partially confirm the results obtained by Bursztyn, Fujiwara and Pallais (2017). They also obtain a significant effect for hours of work. However, they use standard errors from regressing hours on treatment. This is conceptually different from what is done here. While the causal bootstrap also accounts for design uncertainty, their approach only accounts for sampling uncertainty.

However, some doubts remain. First of all, the sample for the group of interest (female & single) contains only 60 observations. This is a rather small sample. Also, it is doubtful whether elite MBA students are really representative of the general population. In many dimensions they likely are not. At the very least, they are not a random sample from the population, which violates the random sampling assumption. Thus, these results should be taken with a grain of salt.

5. Conclusion. The causal bootstrap by Imbens and Menzel (2021) is a powerful alternative to the classical bootstrap for situations where one is interested in inference on causal parameters. It provides asymptotically valid confidence intervals under randomized treatment assignment. This paper focuses on two-sided inference for the ATE and presents the bootstrap procedure for such settings.

As an illustrative example, the procedure is then applied to experimental data by Bursztyn, Fujiwara and Pallais (2017). I find some evidence for the hypothesis that single women display less professional ambition when their preferences are observed by their peers. In particular, they decrease their desired compensation significantly when they think their answers may be observable. Unlike Bursztyn, Fujiwara and Pallais (2017) I find no significant effect for desired hours of weekly work.

References.

Aronow, P. M., Green, D. P. and Lee, D. K. (2014). Sharp bounds on the variance in randomized experiments. *The Annals of Statistics* **42** 850–871.

Bursztyn, L., Fujiwara, T. and Pallais, A. (2017). 'Acting Wife': Marriage Market Incentives and Labor Market Investments. *American Economic Review* **107** 3288-3319.

Efron, B. (1982). *The Bootstrap* In *The Jackknife, the Bootstrap and Other Resampling Plans* 27-36.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* **81** 945–960.

Imbens, G. and Menzel, K. (2021). A causal bootstrap. *The Annals of Statistics* **49** 1460 – 1488.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association* **100** 322-331.