Section 10: Sampling and Resampling

Ricky Truong (rickytruong@college.harvard.edu),
Emily Xing (exing@college.harvard.edu)

1 Introduction

1.1 Logistics

Welcome! The goal of our weekly section is to review content from lecture, with emphasis on big-picture intuition, mathematical proof, and hands-on practice.

Specifically, we aim for our sections to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

1.2 Office Hours

- Mondays, 7:30 - 9:30 PM in Adams D-Hall (Ricky).
- Fridays, 11 AM - 12 PM in Maxwell-Dworkin 2nd Floor (Emily).
- Saturdays, 10:30 – 11:30 AM in Cabot D-Hall (Emily).

2 Big Picture

We begin with a discussion of *sampling*, where we *sample* n units from a larger *population* of N units. Often, we conduct either a *simple random sampling* or a *stratified random sample*. Regardless of what *sampling design* we employ, we can use the *Horvitz-Thompson estimator* to estimate population mean. We then move to *resampling*, where we conduct a *simple random sample* on our *observed* data $\vec{y}$. This allows us to extend inference through *bootstrap confidence intervals* and *permutation tests*.

Importantly, this week requires a big shift in perspective. Like before, Y is a *random* data point in our sample, but we regard the population values as *fixed*. E.g., consider N slips of paper, each with a number on it. These numbers don't change. However, if we sample only 1 slip, it is random which number we collect (i.e., what Y will crystallize to), with the randomness coming from the sampling and not the numbers themselves.

Especially for this week, it is helpful to denote what an expectation is with respect to (i.e., what is random).[1] As a recurring theme of sampling, we denote I as the population ID of a sampled unit, so we often take expectation with respect to the randomness in I. Related, a crucial identity is $Y_j = y_{I_j}$, which expresses the jth data point in terms of I and the population value.

3 Sampling: Design-Based Inference

Idea. "If you don't believe in random sampling, the next time you have a blood test, tell

[1] E.g., suppose $X \mid p \sim \text{Bern}(p)$ and $p \sim \text{Unif}(0,1)$. Then $\mathbb{E}[X] = \mathbb{E}_p[\mathbb{E}_X[X \mid p]] = \mathbb{E}_p[p] = 0.5$ by Adam's Law, where the first expectation is with respect to the randomness in X and the second one is with respect to the randomness in p.

the doctor to take it all." - Arthur Nielsen

In *design-based inference*, we collect a *sample* of size n from a specific, finite *population* of size N, where the population values of interest are *fixed* numbers: $y_1, \ldots, y_N$. In this case, the *randomness* comes from sampling, not from any modeled distribution. We need to infer about the *estimand* because we don't have all the information needed (i.e., our sample size n is less than the population size N). We must specify the *sampling design* (i.e., how sampling is done). Often in the course, we'll be working with an *equal probability sample*, which has nice properties for inference.

Definition 1 (Finite sample estimand)**.** A function of $y_1, \ldots, y_N$.

- E.g., population mean is $\mu = \frac{1}{N}\sum_{j=1}^{N} y_j$.
- E.g., population variance is $\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(y_j - \mu)^2$.
- E.g., population CDF is $F(y) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{y_j \leq y\}$.

Concept Checker 1. For population variance, why is it N in the denominator instead of $N-1$?

Solution

The $n-1$ in the denominator of the sample variance/OLS estimator of variance—$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$—accounts for a missing degree of freedom, allowing the estimator to be unbiased for σ^2. In contrast, population variance is an estimand. Theoretically, if we had data on the entire population, we're not missing any degrees of freedom as we don't have to estimate anything!

Definition 2 (Census)**.** In a *census*, we observe the entire population: $y_1, \ldots, y_N$.

- We rarely get to conduct a census, so we often collect a random sample from the population.

Definition 3 (Sampling design)**.** Any time we collect a random sample of size n from a population of size N, the *sampling design* is the joint PMF of $I_1, \ldots, I_n$, where I_j is the population ID of the jth unit sampled: $P(I_1 = i_1, \ldots, I_n = i_n) = P(\vec{I}_{1:n} = \vec{i}_{1:n})\ \forall\ \vec{i}_{1:n} \in \{1, \ldots, N\}^n$.

- The sampling design determines the properties of our data $\vec{Y} = (Y_1, \ldots, Y_n) = (y_{I_1}, \ldots, y_{I_n})$.
- Again, we regard $y_1, \ldots, y_N$ as fixed, but notice $Y_j = y_{I_j}$, so we can describe the randomness from the sampling with the random variables $I_1, \ldots, I_n$.
- E.g., consider a population of $N = 3$ students with the following SAT scores: $\{1200, 1300, 1000\}$. Notice $y_1 = 1200$, $y_2 = 1300$, $y_3 = 1000$. These scores are fixed. We randomly sample $n = 2$ scores: $\{1300, 1000\}$. Since the first person we sampled was the second person in the population, $I_1 = 2$, so $Y_1 = y_{I_1} = y_2 = 1300$.

Definition 4 (Equal probability sample)**.** A *sampling design* such that the marginal PMF satisfies $P(I_j = k) = \frac{1}{N}\ \forall\ j \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, N\}$.

- I.e., the unconditional probability the jth unit sampled is the kth unit in the population is equally likely for any j and k—there's no reason why, e.g., the first unit sampled is more likely to be the first unit in the population.

• All *equal probability samples* have $\mathbb{E}_I[Y_j] = \mu$, $\text{Var}_I[Y_j] = \sigma^2$, $\mathbb{E}_I[\bar{Y}] = \mu$, and $\mathbb{E}_I[\hat{F}(y)] = F(y)\ \forall\ j \in \{1, \ldots, n\}$, where expectation is with respect to the sampling design (i.e., with respect to the randomness in I).

Concept Checker 2. Recall $\mathbb{E}[g(X)] = \sum_{\text{supp}(X)} g(x)P(X = x)$ for X discrete by LOTUS. Use this to show all equal probability samples have $\mathbb{E}_I[Y_j] = \mu$.

Solution

First, I_j is a discrete random variable with support $\{1, \ldots, N\}$. Additionally, in an equal probability sample, $P(I_j = k) = \frac{1}{N}$. Finally, $Y_j = y_{I_j}$ is a function of I_j.

$$\begin{aligned}
\mathbb{E}_I[Y_j] &= \mathbb{E}_I[y_{I_j}] && \text{by substituting} \\
&= \sum_{k=1}^{N} y_k P(I_j = k) && \text{by LOTUS} \\
&= \sum_{k=1}^{N} y_k \frac{1}{N} && \text{by equal probability sample} \\
&= \frac{1}{N} \sum_{k=1}^{N} y_k && \text{by simplifying} \\
&= \mu && \text{by definition}
\end{aligned}$$

4 Sampling: Simple Random Sampling

Idea. A common *sampling design* is the *simple random sample* (SRS), which can be done with or without *replacement*. *SRS with replacement* and *SRS without replacement* are both *equal probability samples*, so we not only enjoy the previous results (e.g., the expected value of $\bar{Y}$), but we can also gain more information on our sample statistics (e.g., the variance of $\bar{Y}$).

Definition 5 (Discrete Uniform distribution)**.** The discrete analogue of the (Continuous) Uniform distribution. If $X \sim \text{DUnif}(1, N)$, then X is a discrete random variable where $P(X = x) = \frac{1}{N}\ \forall\ x \in \{1, \ldots, N\}$.

Definition 6 (SRS with replacement)**.** An equal probability sample where we draw the n population IDs as $I_j \overset{\text{i.i.d.}}{\sim} \text{DUnif}(1, N)$ and set $Y_j = y_{I_j}\ \forall\ j \in \{1, \ldots, n\}$.

• The sampling design is given by

$$\begin{aligned}
P(I_1 = i_1, \ldots, I_n = i_n) &= \prod_{j=1}^{n} P(I_j = i_j) && \text{by i.i.d.} \\
&= \prod_{j=1}^{n} \frac{1}{N} && \text{by Discrete Uniform} \\
&= \frac{1}{N^n} && \text{by simplifying}
\end{aligned}$$

• All *SRS with replacement* have $\mathbb{E}_{\text{with}}[S^2] = \sigma^2$, $\text{Var}_{\text{with}}[\bar{Y}] = \frac{\sigma^2}{n}$, and $\text{Var}_{\text{with}}[\hat{F}(y)] = \frac{F(y)(1-F(y))}{n}$. Like before, expectation is with respect to the randomness in I, but we sometimes write $\mathbb{E}_{\text{with}}$ as a reminder the sampling is with replacement.

Definition 7 (SRS without replacement)**.** An equal probability sample where we draw the n population IDs such that all $\frac{N!}{(N-n)!}$ permutations are equally likely and set $Y_j = y_{I_j}$ $\forall$ $j \in \{1, \ldots, n\}$.

• The sampling design is given by

$$P(I_1 = i_1, \ldots, I_n = i_n) = \frac{1}{N!/(N-n)!} \quad \text{by naive probability}$$

• Notice this is the same setup as the *Birthday Problem*, where there are $N(N-1)(N-2)\ldots(N-n+1) = \frac{N!}{(N-n)!}$ permutations.[2]

• All *SRS without replacement* have $\text{Cov}_{\text{w/o}}[Y_j, Y_k] = \frac{-\sigma^2}{N-1}$ and $\text{Var}_{\text{w/o}}[\bar{Y}] = \frac{\sigma^2}{n}(\frac{N-n}{N-1})$ $\forall$ $j, k \in \{1, \ldots, n\}$ with $j \neq k$, where $\frac{N-n}{N-1}$ is the *finite population correction.* Like before, expectation is with respect to the randomness in I, but we sometimes write $\mathbb{E}_{\text{w/o}}$ as a reminder the sampling is without replacement.

Proof. Notice $Y_1 + \ldots + Y_N = y_1 + \ldots + y_N$. The left side is expressed in terms of our random data while the right side is expressed in terms of the fixed population values, but they're equal in a census. Since the right side is a constant, its variance is 0.

$$\begin{aligned}
Y_1 + \ldots + Y_N = y_1 + \ldots + y_N &\implies \text{Var}_{\text{w/o}}[Y_1 + \ldots + Y_N] = \text{Var}_{\text{w/o}}[y_1 + \ldots + y_N] && \text{by algebra} \\
&\implies \text{Var}_{\text{w/o}}[Y_1 + \ldots + Y_N] = 0 && \text{by substituting} \\
&\implies \sum_{j=1}^{N} \text{Var}_{\text{w/o}}[Y_j] + 2\sum_{j<k} \text{Cov}_{\text{w/o}}[Y_j, Y_k] = 0 && \text{by bilinearity} \\
&\implies \sum_{j=1}^{N} \sigma^2 + 2\sum_{j<k} \text{Cov}_{\text{w/o}}[Y_j, Y_k] = 0 && \text{by eq. prob. samp.} \\
&\implies N\sigma^2 + 2\binom{N}{2}\text{Cov}_{\text{w/o}}[Y_j, Y_k] = 0 && \text{by simplifying} \\
&\implies N\sigma^2 + N(N-1)\text{Cov}_{\text{w/o}}[Y_j, Y_k] = 0 && \text{by simplifying} \\
&\implies \text{Cov}_{\text{w/o}}[Y_j, Y_k] = \frac{-\sigma^2}{N-1} && \text{by algebra}
\end{aligned}$$

□

Concept Checker 3. Use the previous result (i.e., $\text{Cov}_{\text{w/o}}[Y_j, Y_k] = \frac{-\sigma^2}{N-1}$) to show all SRS without replacement have $\text{Var}_{\text{w/o}}[\bar{Y}] = \frac{\sigma^2}{n}(\frac{N-n}{N-1})$ $\forall$ $j, k \in \{1, \ldots, n\}$ with $j \neq k$.

[2]The number of samples of size n from a population of size N—without replacement and where order matters—is given by $\frac{N!}{(N-n)!}$.

Solution

$$
\begin{aligned}
\mathrm{Var}_{\mathrm{w/o}}[\bar{Y}] &= \mathrm{Var}_{\mathrm{w/o}}\left[\frac{1}{n}\sum_{j=1}^{n} Y_j\right] && \text{by substituting} \\
&= \frac{1}{n^2}\left(\sum_{j=1}^{n}\mathrm{Var}_{\mathrm{w/o}}[Y_j] + 2\sum_{j<k}\mathrm{Cov}_{\mathrm{w/o}}[Y_j, Y_k]\right) && \text{by bilinearity} \\
&= \frac{1}{n^2}\left(\sum_{j=1}^{n}\sigma^2 + 2\sum_{j<k}\mathrm{Cov}_{\mathrm{w/o}}[Y_j, Y_k]\right) && \text{by equal probability sample} \\
&= \frac{1}{n^2}\left(\sum_{j=1}^{n}\sigma^2 + 2\sum_{j<k}\frac{-\sigma^2}{N-1}\right) && \text{by substituting} \\
&= \frac{1}{n^2}\left(n\sigma^2 + 2\binom{n}{2}\frac{-\sigma^2}{N-1}\right) && \text{by simplifying} \\
&= \frac{1}{n^2}\left(n\sigma^2 + n(n-1)\frac{-\sigma^2}{N-1}\right) && \text{by simplifying} \\
&= \frac{\sigma^2}{n}\left(1 - \frac{(n-1)}{N-1}\right) && \text{by simplifying} \\
&= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) && \text{by simplifying}
\end{aligned}
$$

Notice this looks very similar to the $\mathrm{Var}[\bar{Y}] = \frac{\sigma^2}{n}$ result when the data are i.i.d. (with the difference being the finite population correction).

5 Sampling: Stratified Sampling

Idea. Often, it is more feasible and desirable to sample in groups.[3] Thankfully, we can perform *stratified sampling*, where the population is decomposed into groups called *strata*. In each stratum exist *stratum quantities*, which we can relate to *population quantities*. Related, we can pool *stratum-specific estimators* (e.g., sample mean within a stratum) to obtain *stratified estimators* for *population quantities* (e.g., a stratified estimator for population mean).

Definition 8 (Strata)**.** Partition the population IDs—$I_1, \ldots, I_N$—into *strata*, each of size $N_\ell \; \forall \; \ell \in \{1, \ldots, L\}$, where L is the number of strata. Assume $N_\ell \geq 1$—i.e., no stratum has fewer than 1 unit—and $\sum_{\ell=1}^{L} N_\ell = N$—i.e., all stratum sizes add up to the population size. Within the ℓth stratum, we denote the fixed population values as $\{y_{1,\ell}, \ldots, y_{N_\ell,\ell}\}$. Within the ℓth stratum, we define the stratum quantities as the following:

- *Stratum mean* is $\mu_\ell = \frac{1}{N_\ell}\sum_{j=1}^{N_\ell} y_{j,\ell}$, which we can relate to *population mean*: $\mu =$

[3] E.g., we may want to ensure a small but extreme group like the ultra-wealthy are represented when sampling incomes. In an SRS, there's a chance our sample won't capture one of those extreme values, leading to a drastically different estimate and thus a highly variable estimator. We can make our estimator more *efficient* by taking into account *variability within groups* and *variability between groups* (like in Eve's Law)!

$\sum_{\ell=1}^{L} \frac{N_\ell}{N} \mu_\ell$.

• *Stratum variance* is $\sigma_\ell^2 = \frac{1}{N_\ell} \sum_{j=1}^{N_\ell} (y_{j,\ell} - \mu_\ell)^2$, which we can relate to *population variance*: $\sigma^2 = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \sigma_\ell^2 + \sum_{\ell=1}^{L} \frac{N_\ell}{N} (\mu_\ell - \mu)^2$.

• *Stratum CDF* is $F_\ell(y) = \frac{1}{N_\ell} \sum_{j=1}^{N_\ell} \mathbb{1}\{y_{j,\ell} \leq y\}$, which we can relate to *population CDF*: $F(y) = \sum_{\ell=1}^{L} \frac{N_\ell}{N} F_\ell(y)$.

Concept Checker 4. Show $\sum_{\ell=1}^{L} \frac{N_\ell}{N} \mu_\ell = \mu$. Additionally, what role does the $\frac{N_\ell}{N}$ term play? E.g., if $N = 100$, $N_1 = 2$, and $N_2 = 98$ with $L = 2$, what should we take into consideration?

Solution

$$\begin{aligned}
\sum_{\ell=1}^{L} \frac{N_\ell}{N} \mu_\ell &= \sum_{\ell=1}^{L} \frac{N_\ell}{N} \left(\frac{1}{N_\ell} \sum_{j=1}^{N_\ell} y_{j,\ell} \right) && \text{by substituting} \\
&= \frac{1}{N} \sum_{\ell=1}^{L} \sum_{j=1}^{N_\ell} y_{j,\ell} && \text{by simplifying} \\
&= \frac{1}{N} \sum_{j=1}^{N} y_j && \text{by rewriting} \\
&= \mu && \text{by definition}
\end{aligned}$$

For the third step, recall our definition of strata (along with the assumptions used). For some intuition, the $\frac{N_\ell}{N}$ term weighs large strata more. In our example, stratum 2 is much larger than stratum 1, so we want to weigh μ_2 more than μ_1.

Definition 9 (Stratified sampling design)**.** A sampling design is a *stratified sampling design* if the sampling is done independently across strata. We define stratum-specific estimators as the following:

• *Stratum sample mean* is $\bar{Y}_\ell = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} Y_{j,\ell}$, which we can pool to obtain an estimator for *population mean*: $\hat{\mu}_{\text{strat}} = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \bar{Y}_\ell$.

• *Stratum empirical CDF* is $\hat{F}_\ell(y) = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbb{1}\{Y_{j,\ell} \leq y\}$, which we can pool to obtain an estimator for *population CDF*: $\hat{F}_{\text{strat}}(y) = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \hat{F}_\ell(y)$.

• All *stratified sampling designs* have $\mathbb{E}_I[\hat{\mu}_{\text{strat}}] = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \mathbb{E}_I[\bar{Y}_\ell]$ and $\text{Var}_I[\hat{\mu}_{\text{strat}}] = \sum_{\ell=1}^{L} \left(\frac{N_\ell}{N}\right)^2 \text{Var}_I[\bar{Y}_\ell]$, where expectation is with respect to the randomness in I.[4]

Definition 10 (Equal probability stratified sampling design)**.** A sampling design is an *equal probability stratified sampling design* if $P(I_{j,\ell} = k) = \frac{1}{N_\ell}$ $\forall$ $j \in \{1, \ldots, n_\ell\}$, $\ell \in \{1, \ldots, L\}$, and $k \in \{1, \ldots, N_\ell\}$, where $I_{j,\ell}$ is the population ID of the jth unit in the ℓth stratum.

• I.e., the unconditional probability the jth unit sampled is the kth unit in the population within the ℓth stratum is equally likely for any j, ℓ, and k—there's no reason why, e.g.,

[4]For some intuition, the sampling is done independently across strata, so the covariance terms from the bilinearity of variance disappear.

the first unit sampled is more likely to be the first unit in the population within the first stratum.

• All *equal probability stratified sampling designs* have $\mathbb{E}_I[\bar{Y}_\ell] = \mu_\ell$, so $\mathbb{E}_I[\hat{\mu}_{\text{strat}}] = \mu$. Additionally, $\text{Var}_{\text{with}}[\bar{Y}_\ell] = \left(\frac{1}{n_\ell}\right)\sigma_\ell^2$ for *SRS with replacement* and $\text{Var}_{\text{w/o}}[\bar{Y}_\ell] = \left(\frac{1}{n_\ell}\right)\left(\frac{N_\ell - n_\ell}{N_\ell - 1}\right)\sigma_\ell^2$ for *SRS without replacement*.

6 Sampling: Horvitz-Thompson Estimator

Idea. We conclude our discussion of sampling with a famous estimator used in the context of survey sampling: the *Horvitz-Thompson estimator*.

Definition 11 (Horvitz-Thompson estimator)**.** Let the sampling design be given by $P(I_1 = i_1, \ldots, I_n = i_n)$—i.e., any sampling design, with or without replacement. Let $C_j = \mathbb{1}\{I_1 = j\} + \ldots + \mathbb{1}\{I_n = j\}$—i.e., the number of times we sample the jth unit in the population. Let $\pi_j = P(C_j \geq 1)$—i.e., the probability we sample the jth unit in the population. Assume N and $\pi_j > 0$ are known $\forall\ j \in \{1, \ldots, N\}$. The *Horvitz-Thompson estimator* is given by $\hat{\mu}_{\text{HT}} = \frac{1}{N}\sum_{j=1}^{N} \frac{\mathbb{1}\{C_j \geq 1\}}{\pi_j} y_j$.

• The Horvitz-Thompson estimator is unbiased for μ (i.e., $\mathbb{E}_I[\hat{\mu}_{\text{HT}}] = \mu$).

Proof.

$$\begin{aligned}
\mathbb{E}_I[\hat{\mu}_{\text{HT}}] &= \mathbb{E}_I\left[\frac{1}{N}\sum_{j=1}^{N}\frac{\mathbb{1}\{C_j \geq 1\}}{\pi_j} y_j\right] && \text{by substituting} \\
&= \frac{1}{N}\sum_{j=1}^{N}\frac{y_j}{\pi_j}\mathbb{E}_I\left[\mathbb{1}\{C_j \geq 1\}\right] && \text{by linearity} \\
&= \frac{1}{N}\sum_{j=1}^{N}\frac{y_j}{\pi_j} P(C_j \geq 1) && \text{by fundamental bridge} \\
&= \frac{1}{N}\sum_{j=1}^{N}\frac{y_j}{\pi_j}\pi_j && \text{by substituting} \\
&= \frac{1}{N}\sum_{j=1}^{N} y_j && \text{by simplifying} \\
&= \mu && \text{by definition}
\end{aligned}$$

Though unbiased, this estimator isn't necessarily "good" as it can have very high variance! □

Concept Checker 5. The sum in the Horvitz-Thompson estimator is indexed by N. Why is it still a valid estimator?

Solution

The indicator—$\mathbb{1}\{C_j \geq 1\}$—"kills" the summands for units in the population we didn't sample. E.g., if unit k in the population wasn't sampled (even though there was a non-zero probability they could've been), then $\mathbb{1}\{C_j \geq 1\} = 0$, so we don't need y_k!

7 Resampling: Bootstrapping

Idea. The *bootstrap* is an incredibly powerful idea in statistics based on *resampling* our data. Through this deceptively simple action, we can conduct inference to learn more about our estimator (specifically, by approximating *standard error* and constructing *bootstrap confidence intervals*). There are two worlds: *real world* (generated by F) and *bootstrap world* (generated by $\hat{F}$). Importantly, $\hat{F}(y) \xrightarrow{p} F(y)$ by LLN, connecting the two. The bootstrap sounds magical, but it relies on a sufficiently large n and the assumption of i.i.d. data. It is also *computationally expensive*, so in practice, we often estimate bootstrap-world quantities via *Monte Carlo estimation*.

7.1 Fundamentals

Definition 12 (Bootstrap)**.** Let $\vec{y} = (y_1, \ldots, y_n)$ be the *observed* dataset, assumed to be i.i.d. from some unknown CDF F. We create a synthetic dataset $\vec{Y}^* = (Y_1^*, \ldots, Y_n^*)$ by performing an *SRS with replacement* from $\vec{y}$. Equivalently, $Y_j^* \overset{\text{i.i.d.}}{\sim} \hat{F}$, where $\hat{F}$ is the ECDF of $\vec{y}$—i.e., $\hat{F}(y) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{y_j \leq y\}$. We call $\vec{Y}^*$ a *bootstrap sample*.

• *Bootstrapping* involves generating some large number B of independent bootstrap samples, each of size n, and using the bootstrap samples for inferential tasks.

Definition 13 (Real world)**.** F generates $\vec{Y}$, which generates $\hat{\theta}$. We regard $\hat{\theta}$ as an estimate for θ.

Definition 14 (Bootstrap world)**.** $\hat{F}$ generates $\vec{Y}^*$, which generates $\hat{\theta}^*$. We regard $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ as estimates for $\hat{\theta}$, where $\hat{\theta}_j^*$ is a statistic calculated from the jth bootstrap sample $\vec{Y}_j^*$. We define the following bootstrap-world quantities, where we use the *bootstrap expectation* (i.e., sampling with replacement from $\hat{F}$, conditional on the observed data $\vec{y}$):

• $\text{Bias}_{\text{boot}}[\hat{\theta}^*] = \mathbb{E}_{\text{boot}}[\hat{\theta}^*] - \hat{\theta}$.

• $\text{SE}_{\text{boot}}[\hat{\theta}^*] = \sqrt{\mathbb{E}_{\text{boot}}[(\hat{\theta}^* - \mathbb{E}_{\text{boot}}[\hat{\theta}^*])^2]}$.

• $F_{\text{boot}}(\theta) = P_{\text{boot}}(\hat{\theta}^* \leq \theta)$.

Concept Checker 6. Suppose we generate one bootstrap sample $\vec{Y}^* = (Y_1^*, \ldots, Y_n^*)$. Show the bootstrap expectation of an arbitrary point in the bootstrap sample is the observed sample mean (i.e., $\mathbb{E}_{\text{boot}}[Y_j^*] = \bar{y}$).

Solution

First, Y_j^*, an arbitrary resampled point, is a random variable that can crystallize into any number within $\{y_1, \ldots, y_n\}$. Let I_j be the sample ID of the jth unit resampled such that $Y_j^* = y_{I_j}$. Notice $I_j \overset{\text{i.i.d.}}{\sim} \text{DUnif}(1, n)$ by SRS with replacement.

$$\begin{aligned}
\mathbb{E}_{\text{boot}}[Y_j^*] &= \mathbb{E}_{\text{boot}}[y_{I_j}] && \text{by substituting} \\
&= \sum_{k=1}^{n} y_k P(I_j = k) && \text{by LOTUS} \\
&= \sum_{k=1}^{n} y_k \frac{1}{n} && \text{by Discrete Uniform} \\
&= \frac{1}{n} \sum_{k=1}^{n} y_k && \text{by simplifying} \\
&= \bar{y} && \text{by definition}
\end{aligned}$$

Definition 15 (Connection between worlds)**.** Often, it is difficult to learn about an estimator $\hat{\theta}$ mathematically, so it is natural to try a simulation-based approach. A fundamental challenge is we only observe one dataset (and thus only one observed value of $\hat{\theta}$). Ideally, we could generate data $\{Y_1^{[1]}, \ldots, Y_n^{[1]}\}, \ldots, \{Y_1^{[B]}, \ldots, Y_n^{[B]}\}$, each $Y_j^{[b]} \overset{\text{i.i.d.}}{\sim} F \ \forall \ b \in \{1, \ldots, B\}$ and $j \in \{1, \ldots, n\}$. From there, we'd compute replications $\hat{\theta}_1, \ldots, \hat{\theta}_B$, and for B large, the standard deviation of these replications would be close our desired $\text{SE}[\hat{\theta}]$. However, F is unknown! Instead, we use $\hat{F}$, which is known, to compute bootstrap replications $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. Since $\hat{F}(y) \xrightarrow{p} F(y)$ by LLN, $\hat{F}$ is likely to approximate F well if n is sufficiently large.

Definition 16 (Computational cost of bootstrap)**.** Suppose we want the bootstrap expectation of some statistic $\hat{\theta}^* = T(\vec{Y}^*)$. It can be shown $\mathbb{E}_{\text{boot}}[T(\vec{Y}^*)] = \frac{1}{n^n} \sum_{(i_1, \ldots, i_n) \in \{1, \ldots, n\}^n} T(y_{i_1}, \ldots, y_{i_n})$, where the n^n becomes very computationally expensive for even moderately large n.[5]. Thus, in practice, we use *Monte Carlo estimation* (i.e., we use only a subset of all possible bootstrap samples).

Definition 17 (Monte Carlo estimation)**.** *Monte Carlo estimation* works since the bootstrap replications $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ are i.i.d. draws from the bootstrap world, so each estimator converges in probability to its respective bootstrap-world quantity as $B \to \infty$ by LLN. We use the bootstrap replications to calculate the following estimators:

- $\hat{\mathbb{E}}_{\text{boot}}[\hat{\theta}^*] = \bar{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^* \xrightarrow{p} \mathbb{E}_{\text{boot}}[\hat{\theta}^*]$.
- $\widehat{\text{Bias}}_{\text{boot}}[\hat{\theta}^*] = \bar{\theta}^* - \hat{\theta} \xrightarrow{p} \text{Bias}_{\text{boot}}[\hat{\theta}^*]$.
- $\widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*] = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b^* - \bar{\theta}^*)^2} \xrightarrow{p} \text{SE}_{\text{boot}}[\hat{\theta}^*]$.
- $\hat{F}_{\text{boot}}(\theta) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{\hat{\theta}_b^* \leq \theta\} \xrightarrow{p} F_{\text{boot}}(\theta)$.

[5] E.g., for $n = 30$, $30^{30} \approx 2 \times 10^{44}$.

Example 1 (Standard error)**.** We want $\text{SE}[\hat{\theta}] = \sqrt{\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]}$. We can approximate this with $\text{SE}_{\text{boot}}[\hat{\theta}^*] = \sqrt{\mathbb{E}_{\text{boot}}[(\hat{\theta}^* - \mathbb{E}_{\text{boot}}[\hat{\theta}^*])^2]}$. But this is often computationally expensive, so in practice, we estimate it with $\widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*] = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}^*)^2}$. Crucially, there are two different sources of error.

• The difference between $\text{SE}_{\text{boot}}[\hat{\theta}^*]$ and $\text{SE}[\hat{\theta}]$ is caused by the use of $\hat{F}$ over F. This error falls as n increases.

• The difference between $\widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*]$ and $\text{SE}_{\text{boot}}[\hat{\theta}^*]$ is caused by the use of Monte Carlo simulation. This error falls as B increases, which is under our control.

7.2 Bootstrap Confidence Intervals

Definition 18 (Setup of bootstrap confidence intervals)**.** Let θ be the estimand and $\hat{\theta}$ be an estimator for θ. Suppose we want to use the bootstrap to get an approximate 95% confidence interval (we choose $\alpha = 0.05$ to simplify notation, but this can be generalized to any α). We have three different procedures.[6]

Definition 19 (Normal interval with bootstrap standard error)**.** $\hat{\theta} \pm 1.96 \times \widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*]$.

• For some intuition, this is analogous to the confidence interval $\hat{\theta} \pm 1.96 \times \widehat{\text{SE}}[\hat{\theta}]$ but uses the bootstrap estimate of the standard error.

• We avoid the need to do math to approximate standard error, but for this to work well, we want the distribution of $\hat{\theta}$ to be approximately Normal with mean θ.

Definition 20 (Percentile method)**.** $\left[\hat{\theta}^*_{(\lceil 0.025B \rceil)}, \hat{\theta}^*_{(\lceil 0.975B \rceil)}\right]$.

• For some intuition, this is analogous to how Bayesian credible intervals are estimated by simulation.

• This is easy to compute and doesn't require $\hat{\theta}$ to be approximately Normal, but for this to work well, we don't want $\hat{\theta}$ to be biased or have a skewed distribution.

Definition 21 (Bootstrap t interval)**.** $\left[\hat{\theta} - \hat{Q}^*(0.975)\widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*], \hat{\theta} - \hat{Q}^*(0.025)\widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*]\right]$, where $\hat{Q}^*(p) = T^*_{(\lceil Bp \rceil)}$ is the sample p-quantile of the bootstrap replications $T_1^*, \ldots, T_B^*$ of $T = \frac{\hat{\theta} - \theta}{\widehat{\text{SE}}[\hat{\theta}]}$.[7]

• For some intuition, this is analogous to the classical method of creating a pivot that has a t distribution (though despite the name, this procedure doesn't use a t distribution—or require any parametric assumptions)!

[6] The first two are "quick and dirty" while the last one, though more computationally extensive, tends to have better performance.

[7] $T_j^* = \frac{\hat{\theta}_j^* - \hat{\theta}}{\text{SE}_{\text{boot}}[\hat{\theta}_j^*]}$, but like before, $\text{SE}_{\text{boot}}[\hat{\theta}_j^*]$ may be computationally expensive, so we replace it with the Monte Carlo estimate, which may involve a second "level" of bootstrapping.

8 Resampling: Permutation Tests

Idea. In addition to *bootstrap confidence intervals*, resampling can be extended to *hypothesis tests*. Specifically, we can run a *permutation test* to compare the data from two groups of interest. If the two groups came from the same underlying data-generating process, then it shouldn't matter if we shuffle values between the two groups when calculating a statistic. Like before, we make *no parametric assumptions*, and computational effort replaces mathematical effort.

Definition 22 (Permutation test)**.** Suppose there are two groups. We observe $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F_X$ for group 0 and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_Y$ for group 1 and assume $\vec{X} \perp\!\!\!\perp \vec{Y}$. We can use a *permutation test* for the hypotheses $H_0 : F_X = F_Y$ vs. $H_A : F_X \neq F_Y$.

• **Strategy**: Let T be a test statistic, chosen such that large values of T are evidence against H_0.[8] Compute the observed value t_0 of T. Generate a large number B of random permutations of the data, each an SRS without replacement of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$.[9] For each permutation, compute the test statistic and call these $t_1, \ldots, t_B$. The p-value is $P_0(T \geq t_0) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{t_b \geq t_0\}$.[10]

9 Practice Problems

Problem 1. Let $y_1, \ldots, y_N$ be the SAT scores in a population of N students. Apparently, Harvard wants to see "at least a 1500 to have a chance at being considered," so an applicant is interested in the estimand $\theta = \frac{1}{\sum_{j=1}^{N} \mathbb{1}\{y_j \geq 1500\}} \sum_{j=1}^{N} y_j \mathbb{1}\{y_j \geq 1500\}$ (i.e., the average SAT score for those who scored at least a 1500).[11] They manage to collect a simple random sample with replacement, obtaining the dataset $Y_1, \ldots, Y_n$. They consider estimating θ by replacing y_j with Y_j and replacing N with n, resulting in the estimator $\hat{\theta} = \frac{1}{\sum_{j=1}^{n} \mathbb{1}\{Y_j \geq 1500\}} \sum_{j=1}^{n} Y_j \mathbb{1}\{Y_j \geq 1500\}$. Assume $P(\sum_{j=1}^{n} \mathbb{1}\{Y_j \geq 1500\} = 0) = P(\sum_{j=1}^{N} \mathbb{1}\{y_j \geq 1500\} = 0) = 0$.

(a) Is $\hat{\theta}$ a consistent estimator for θ?

(b) Since it isn't always reasonable to assume $P(\sum_{j=1}^{n} \mathbb{1}\{Y_j \geq 1500\} = 0) = 0$, the applicant considers an alternative estimator: $\tilde{\theta} = \frac{1}{n}\sum_{j=1}^{n} Y_j \mathbb{1}\{Y_j \geq 1500\}$.[12] Is $\tilde{\theta}$ an unbiased, positively biased, or negatively biased estimator for θ? If it is biased, when does $\text{Bias}_{\text{with}}[\tilde{\theta}] = 0$?

Solution

Let's clean up the notation. Let $X_j = \mathbb{1}\{Y_j \geq 1500\}$, $x_j = \mathbb{1}\{y_j \geq 1500\}$, $Z_j =$

[8] For example, one common choice is $T = |\bar{X} - \bar{Y}|$.

[9] This scrambles which data points belong to which groups.

[10] P_0 denotes probability under the permutation distribution of T (i.e., the distribution under random shuffles of the data rather than under repeated samples).

[11] https://www.prepscholar.com/sat/s/colleges/Harvard-sat-scores-GPA

[12] Arguably, this is also a questionable estimator. If $\sum_{j=1}^{n} \mathbb{1}\{Y_j \geq 1500\} = 0$, then $\tilde{\theta} = 0$, which is a pretty bad estimate! But at least it's a mathematically well-defined quantity.

$Y_j \mathbb{1}\{Y_j \geq 1500\}$, and $z_j = y_j \mathbb{1}\{y_j \geq 1500\}$. Thus, $\hat{\theta} = \frac{1}{\sum_{j=1}^n X_j} \sum_{j=1}^n Z_j$ and $\theta = \frac{1}{\sum_{j=1}^N x_j} \sum_{j=1}^N z_j$. Importantly, $\hat{\theta} = \frac{1}{\sum_{j=1}^n X_j} \sum_{j=1}^n Z_j = \frac{1}{\frac{1}{n}\sum_{j=1}^n X_j} \left(\frac{1}{n} \sum_{j=1}^n Z_j\right) = \frac{\bar{Z}_n}{\bar{X}_n}$ by algebra.

First, let's see if $\hat{\theta}$ is consistent. Notice $\frac{1}{n} \sum_{j=1}^n X_j \xrightarrow{p} \mathbb{E}_{\text{with}}[X_j]$ by LLN (we're sampling with replacement, so we can allow $n \to \infty$).

$$\begin{aligned}
\mathbb{E}_{\text{with}}[X_j] &= \mathbb{E}_{\text{with}}[x_{I_j}] && \text{by substituting} \\
&= \sum_{k=1}^N x_k P(I_j = k) && \text{by LOTUS} \\
&= \sum_{k=1}^N x_k \frac{1}{N} && \text{by Discrete Uniform} \\
&= \frac{1}{N} \sum_{k=1}^N x_k && \text{by simplifying}
\end{aligned}$$

Similarly, $\frac{1}{n} \sum_{j=1}^n Z_j \xrightarrow{p} \mathbb{E}_{\text{with}}[Z_j]$ by LLN.

$$\mathbb{E}_{\text{with}}[Z_j] = \frac{1}{N} \sum_{k=1}^N z_k \quad \text{by the reasoning above}$$

Thus, $\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{j=1}^n X_j} \left(\frac{1}{n} \sum_{j=1}^n Z_j\right) \xrightarrow{p} \frac{1}{\frac{1}{N}\sum_{k=1}^N x_k} \left(\frac{1}{N} \sum_{k=1}^N z_k\right) = \frac{1}{\sum_{j=1}^N x_j} \sum_{j=1}^N z_j = \theta$ by Theorem 3.5.7. We conclude $\hat{\theta}$ is a consistent estimator for θ.

Next, let's investigate the bias of $\tilde{\theta}$ by first finding $\mathbb{E}_{\text{with}}[\tilde{\theta}]$.

$$
\begin{aligned}
\mathbb{E}_{\text{with}}[\tilde{\theta}] &= \mathbb{E}_{\text{with}}\left[\frac{1}{n}\sum_{j=1}^{n} Y_j \mathbb{1}\{Y_j \geq 1500\}\right] && \text{by substituting} \\
&= \mathbb{E}_{\text{with}}\left[\frac{1}{n}\sum_{j=1}^{n} Z_j\right] && \text{by substituting} \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_{\text{with}}\left[Z_j\right] && \text{by linearity} \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{N}\sum_{k=1}^{N} z_k\right) && \text{by substituting} \\
&= \frac{1}{N}\sum_{k=1}^{N} z_k && \text{by simplifying} \\
&= \frac{1}{N}\left(\theta \sum_{k=1}^{N} \mathbb{1}\{y_k \geq 1500\}\right) && \text{by substituting} \\
\text{Bias}_{\text{with}}[\tilde{\theta}] &= \mathbb{E}_{\text{with}}[\tilde{\theta}] - \theta && \text{by definition} \\
&= \frac{1}{N}\left(\theta \sum_{k=1}^{N} \mathbb{1}\{y_k \geq 1500\}\right) - \theta && \text{by substituting} \\
&= \theta\left(\frac{1}{N}\sum_{k=1}^{N} \mathbb{1}\{y_k \geq 1500\} - 1\right) && \text{by simplifying}
\end{aligned}
$$

Thus, $\tilde{\theta}$ is negatively biased for θ. Notice $\text{Bias}_{\text{with}}[\tilde{\theta}] = 0$ when $\sum_{k=1}^{N} \mathbb{1}\{y_k \geq 1500\} = N$ (i.e., when all N students in the population score at least a 1500), which makes sense since in that scenario, $Y_j \mathbb{1}\{Y_j \geq 1500\} = Y_j$ and $y_j \mathbb{1}\{y_j \geq 1500\} = y_j \ \forall\ j$, so we're working with the usual sample mean in sampling with replacement!

Problem 2. A company has been accused of gender discrimination, allegedly tending to pay higher salaries to men than to women. Suppose $X_1, \ldots, X_n$ are i.i.d. draws from a theoretical distribution of salaries for men at the company and $Y_1, \ldots, Y_m$ are i.i.d. draws from a theoretical distribution of salaries for women at the same company. The salaries $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are independent but not necessarily i.i.d. Let $T = \bar{X}_n - \bar{Y}_m$. Intuitively, we would like to know whether T has such a large positive value it would be implausible to observe that value (or an even larger value) if the two theoretical distributions were equal. However, n and m are small. The CLT is not applicable with such small sample sizes, so it is not reasonable to assume $\bar{X}_n$ and $\bar{Y}_m$ are approximately Normal.[13]

(a) Explain how we can use a permutation test to answer our scientific question of interest.

(b) Suppose $n = m = 2$ and the data are $X_1 = 8, X_2 = 4, Y_1 = 6, Y_2 = 2$, measured in tens of thousands of dollars. Find the p-value for the permutation test.

[13]Inspired by Problem 6 in "Stat 111 Final Exam, Spring 2025" by Joseph K. Blitzstein and Neil Shephard.

Solution

We are interested in finding a one-sided p-value (this corresponds with the given test statistic—$T = \bar{X}_n - \bar{Y}_m$—and scientific question of interest—whether men are paid higher than women at this company.

Let α be the (pre-determined) value for the nominal size of this test. First, we can compute t_0, the observed value of T from the data at hand. Then, we can generate replications $t_1, \ldots, t_B$ of the test statistic through resampling, where B is a large number (e.g., $B = 10^5$). We obtain the bth replication as follows:

• Randomly permute which salaries are in the men's group and which are in the women's group by choosing a completely random set of n salaries and assigning them to the men's group (and then assigning the remaining m salaries to the women's group).

• Let t_b be the new difference in sample means between the salaries in the men's group and the salaries in the women's group.

We can compare t_0 with the replications (which are made under the null hypothesis). The (approximate) p-value is the proportion of replications that are at least as extreme as the observed value—i.e., $P_0(T \geq t_0) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{t_b \geq t_0\}$. We reject H_0 if $P_0(T \geq t_0) < \alpha$.

Now, if the data are $X_1 = 8, X_2 = 4, Y_1 = 6, Y_2 = 2$ with $n = m = 2$, then we can follow the procedure above to find the (approximate) p-value. First, $t_0 = \frac{1}{2}(8+4) - \frac{1}{2}(6+2) = 2$.

Next, notice if we choose the $n = 2$ salaries in the men's group, then it is deterministic which salaries are in the women's group. There are $\binom{4}{2} = 6$ possible combinations of salaries in the men's group (where order doesn't matter since $\{8, 4\}$ and $\{4, 8\}$ both produce $\bar{X}_n = 6$ and, deterministically, $\bar{Y}_m = 4$ for $T = 2$): $\{8, 4\}$, $\{8, 6\}$, $\{8, 2\}$, $\{4, 6\}$, $\{4, 2\}$, $\{6, 2\}$. Corresponding to that order, the replications $t_1, \ldots, t_6$ of $T = \bar{X}_n - \bar{Y}_m$ would be 2, 4, 0, 0, -4, -2. Since $P_0(T \geq t_0) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{t_b \geq t_0\}$ (i.e., the proportion of these 6 cases where the difference in sample means is at least as large as $t_0 = 2$), the (approximate) p-value is $P_0(T \geq t_0) \approx \frac{1}{3}$.