STAT 641: BOOTSTRAPPING METHODS

Jiyoun Myung

Department of Statistics and Biostatistics
California State University, East Bay

Spring 2021, Day 2

Chapter 1: Introduction

Motivating Example

Data: Aspirin Study (New York Times of January 27, 1987)

A study was done to see if small aspirin doses would prevent heart attacks in healthy middle-aged men. The data for the aspirin study were collected in a particularly efficient way: by a controlled, randomized, double-blind study.

	heart attacks*	subjects
asprin group	104	11037
placebo group	189	11034

*: heart attacks (fatal plus non-fatal)

Aspirin data

	heart attacks*	subjects
asprin group	104	11037
placebo group	189	11034

What strikes the eye here is the lower rate of heart attacks in the aspirin group. The ratio of the two rates is

$$\hat{\theta} = \frac{104/11037}{189/11034} = 0.55.$$

It suggests that the aspirin-takers only have 55% as many heart attacks as placebo-takers.

*: heart attacks (fatal plus non-fatal)

Aspirin data

The aspirin study tracked strokes as well as heart attacks, with the following results:

	strokes	subjects
asprin group	119	11037
placebo group	98	11034

For strokes, the ratio of rates is

$$\hat{\theta} = \frac{119/11037}{98/11034} = 1.21.$$

It now looks like taking aspirin is actually harmful. However the interval for the true stroke ratio θ turns out to be $0.93 < \theta < 1.59$ with 95% confidence. This includes the neutral value $\theta = 1$, at which aspirin would be no better or worse than placebo. *In the language of statistical hypothesis testing, aspirin was found to be significantly beneficial for preventing heart attacks, but not significantly harmful for causing strokes.*

Aspirin data: bootstrap replicates

Obtaining 1000 bootstrap replicates $\hat{\theta}^*$

- Step 1: Create two populations: the first of 119 ones and $11037 - 119 = 10918$ zeros, and the second consisting of 98 ones and $11034 - 98 = 10936$ zeros.
- Step 2: Draw with replacement a sample of 11037 items from the first population, and a sample of 11034 items from the second population. Each of these is called a **bootstrap sample.**
- Step 3: Derive the bootstrap replicate of $\hat{\theta}$:

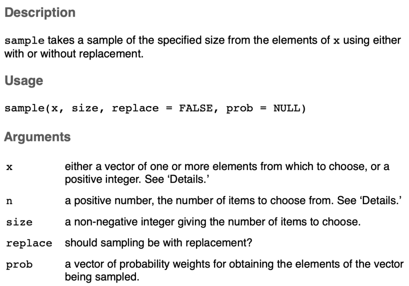
$$\hat{\theta}^* = \frac{\text{proportion of ones in first bootstrap sample}}{\text{proportion of ones in the second bootstrap sample}}$$

- Step 4: Repeat this process (Step 1 - Step 3) a large number of times, say 1000 times, and obtain 1000 bootstrap replicates $\hat{\theta}^*$.

Aspirin data: bootstrap replicates

Let's use the sample() function in the basic R packages that lies at the heart of resampling.

```
?sample
```



Description

`sample` takes a sample of the specified size from the elements of `x` using either with or without replacement.

Usage

```
sample(x, size, replace = FALSE, prob = NULL)
```

Arguments

`x` either a vector of one or more elements from which to choose, or a positive integer. See 'Details.'

`n` a positive number, the number of items to choose from. See 'Details.'

`size` a non-negative integer giving the number of items to choose.

`replace` should sampling be with replacement?

`prob` a vector of probability weights for obtaining the elements of the vector being sampled.

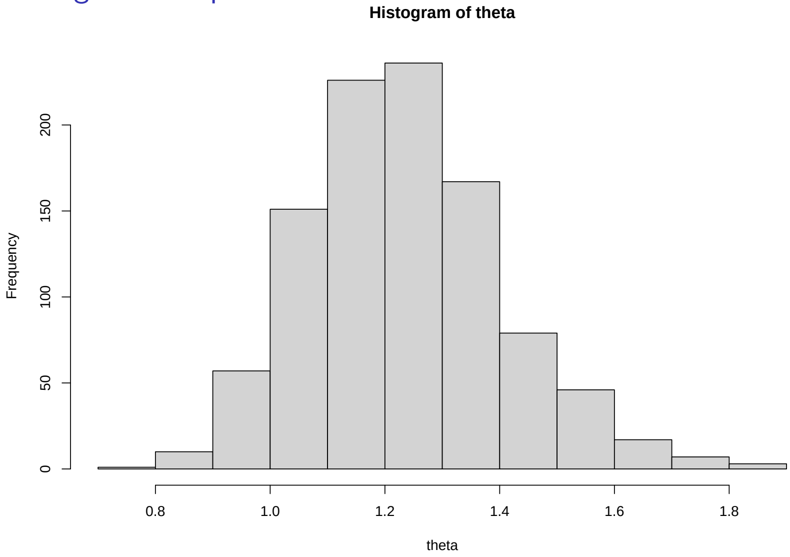
Aspirin data: bootstrap replicates

R Example code:

```r
aspirin <- c(rep(0, 10918), rep(1, 119))
n_a <- length(aspirin)
placebo <- c(rep(0, 10936), rep(1, 98))
n_p <- length(placebo)
theta <- c()
for(i in 1:1000){
  s_a <- sample(aspirin, n_a, replace = TRUE)
  s_p <- sample(placebo, n_p, replace = TRUE)
  theta[i] <- (sum(s_a) / n_a) / (sum(s_p) / n_p)
}
```

Aspirin data: bootstrap replicates

Histogram of replicates:



Aspirin data: bootstrap replicates

Histogram of replicates:

```r
mean(theta)
```

```
## [1] 1.235235
```

```r
quantile(theta, c(0.025, 0.975))
```

```
##      2.5%     97.5% 
## 0.9414128 1.6077280
```

Some of the notation used in the book

- Lower case bold letters refer to data vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.
- Upper case bold letter $\boldsymbol{X}$ refers to a random variable.
- Upper case plain letter X refers to a matrix.
- A superscript "*" indicates a bootstrap random variable:

  $\mathbf{x}^*$ indicates a bootstrap data set generated from a data set $\mathbf{x}$.
- Parameters are denoted by Greek letters such as θ.
- A hat on a letter indicates an estimate, such as $\hat{\theta}$.
- The letters F and G refer to populations.
- The notation $F \rightarrow (x_1, x_2, \ldots, x_n)$ indicates an independent and identically distributed sample drawn from F.

  Or, $x_i \overset{\text{i.i.d.}}{\sim} F$ for $i = 1, 2, \ldots, n$.

Chapter 2: The accuracy of a sample mean

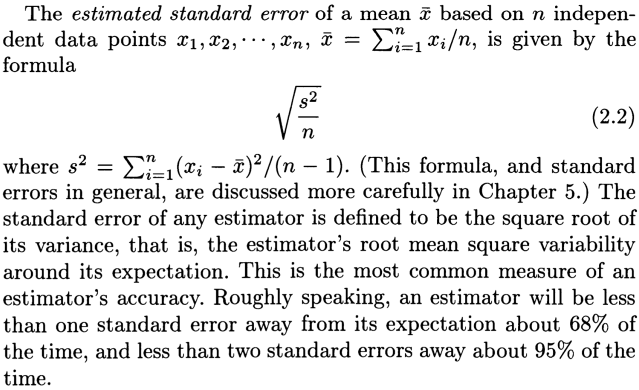
Central Limit Theorem

Suppose a population parameter θ is the target of a study; say for example, θ is the household mean income of a chosen community. A random sample of size n yields the data $(x_1, x_2, \ldots, x_n)$.The corresponding sample statistic computed from this data set is $\hat{\theta}$.

For most sample statistics, the sampling distribution of $\hat{\theta}$ for large n ($n \geq 30$ is generally accepted as large sample size), is bell shaped with center θ and standard deviation $s/\sqrt{n}$, where s depends on the population and the statistics θ.

YouTube: CLT

The estimated standard error of a sample mean



The *estimated standard error* of a mean $\bar{x}$ based on n independent data points $x_1, x_2, \cdots, x_n$, $\bar{x} = \sum_{i=1}^{n} x_i/n$, is given by the formula

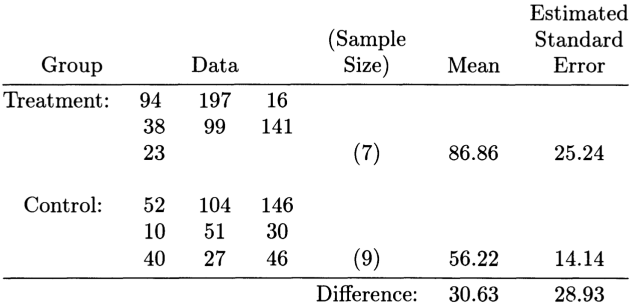
$$\sqrt{\frac{s^2}{n}} \tag{2.2}$$

where $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$. (This formula, and standard errors in general, are discussed more carefully in Chapter 5.) The standard error of any estimator is defined to be the square root of its variance, that is, the estimator's root mean square variability around its expectation. This is the most common measure of an estimator's accuracy. Roughly speaking, an estimator will be less than one standard error away from its expectation about 68% of the time, and less than two standard errors away about 95% of the time.

Mouse Data

Mouse Data: Did treatment prolong surivival?

The table shows the results of a small experiment, in which 7 out of 16 mice were randomly selected to receive a new medical treatment, while the remaining 9 were assigned to the non-treatment (control) group. The treatment was intended to prolong survival after a test surgery. The table shows the survival time following surgery, in days, for all 16 mice.



Group	Data			(Sample Size)	Mean	Estimated Standard Error
Treatment:	94	197	16			
	38	99	141			
	23			(7)	86.86	25.24
Control:	52	104	146			
	10	51	30			
	40	27	46	(9)	56.22	14.14
				Difference:	30.63	28.93

Mouse Data

In the mouse data example, we are interested in the question whether the new treatment lead to an increase in survival time. For this, we might consider the studentized test statistics

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\widehat{\mathsf{se}}\left(\overline{X}\right)^2 + \widehat{\mathsf{se}}\left(\overline{Y}\right)^2}}.$$

The observed value of T is $30.63/28.93 = 1.05$, which indicates that the effect of the new treatment on survival is not significant.

Bootstrap Estimate For Standard Error

Invendted by Efron in 1979.

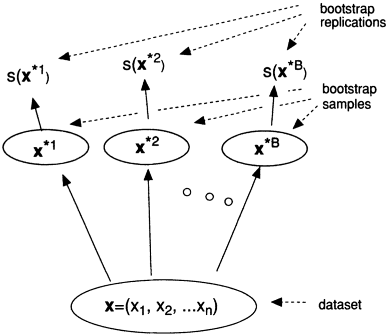


Figure 2.1. *Schematic of the bootstrap process for estimating the standard error of a statistic $s(\mathbf{x})$. B bootstrap samples are generated from the original data set. Each bootstrap sample has n elements, generated by sampling with replacement n times from the original data set. Bootstrap replicates $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \ldots s(\mathbf{x}^{*B})$ are obtained by calculating the value of the statistic $s(\mathbf{x})$ on each bootstrap sample. Finally, the standard deviation of the values $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \ldots s(\mathbf{x}^{*B})$ is our estimate of the standard error of $s(\mathbf{x})$.*

Bootstrap Estimate For Standard Error

- The bootstrap algorithm begins by generating a large number of independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$, each of size n.
- Typical values for B, the number of bootstrap samples, range from 50 to 200 for standard error estimation.
- A crude recommendation for the size B could be $B = n^2$. If that's is too large, you could use $B = n \log_e n = n \ln n$.
- The bootstrap estimate of standard error is the standard deviation of the bootstrap replications,

$$\widehat{\mathsf{se}}_{\mathsf{boot}} = \left\{ \sum_{b=1}^{B} \left[s(\mathbf{x}^{*b}) - s(\cdot)) \right]^2 / (B-1) \right\}^{\frac{1}{2}},$$

where $s(\cdot) = \sum_{b=1}^{B} s(\mathbf{x}^{*b})/B$.

- *It is easy to write a bootstrap program that works for any computable statistic.*

Bootstrap Central Limit Theorem

- Let $\hat{\theta}_B$ stand for a random quantity which represents the same statistic computed on a bootstrap sample drawn out of $x_1, x_2, \ldots, x_n$.
- Bickel and Freedman (1981) and Singh (1981) provided large sample answers for most of the commonly used statistics. As $n \to \infty$, the sampling distribution of $\hat{\theta}_B$ is also bell shaped with θ as the center and the same standard deviation $(s/\sqrt{n})$ in limit.
- The bootstrap distribution of $\hat{\theta}_B - \hat{\theta}$ approximates (fairly well) the sampling distribution of $\hat{\theta} - \theta$.
- For a proof of bootstrap CLT for the mean, check Singh (1981).

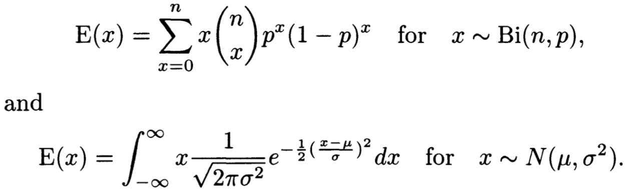
Chapter 3 & 4: More definitions

Random Sample

- A random sample of size n is defined as a collection of n units $x_1, x_2, \ldots, x_n$ selected at random from $\mathcal{X}$. The symbol $\mathcal{X}$ denote the population of measurements of $X_1, X_2, \ldots, X_N$.
- Each unit in the population thus has probability $1/N$ of being included in the sample.
- Random samples can be selected with or without replacement. In STAT 641, we are going to assume sampling with replacement.

Probability Distribution Function

Binominal Distribution and Normal Distribution



$$\mathrm{E}(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^x \quad \text{for} \quad x \sim \mathrm{Bi}(n,p),$$

and

$$\mathrm{E}(x) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx \quad \text{for} \quad x \sim N(\mu, \sigma^2).$$

- If $X \sim \mathsf{Bi}(n,p)$,

$$E(X) = \sum_{x=0}^{n} x_n C_x p^x (1-p)^{n-x}.$$

- If $X \sim N(\mu, \sigma^2)$,

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Indicator Function

Let A be a subset of the sample space, and take $r = I_{\{x \in A\}}$ where $I_{\{x \in A\}}$ is the indicator function

$$I_{\{x \in A\}} = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Then $E(r)$ equals $\text{Prob}\{x \in A\}$, or equivalently

$$E\left(I_{\{x \in A\}}\right) = \text{Prob}\{x \in A\}.$$

For example if $x \sim N(\mu, \sigma^2)$ with the pdf f, then

$$\begin{aligned} E(r) &= \int_{-\infty}^{\infty} I_{\{x \in A\}} f(x) dx \\ &= \int_A f(x) dx = \text{Prob}\{x \in A\}. \end{aligned}$$

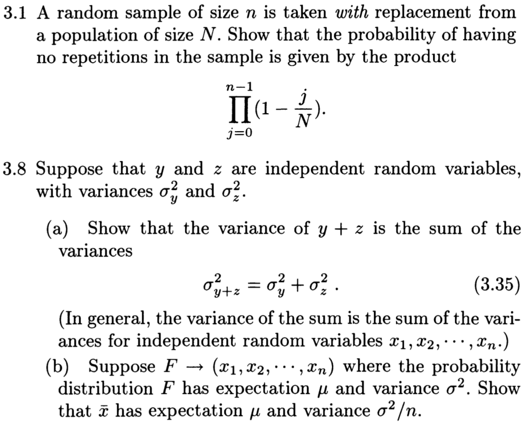
Problem

Let's form several subgroups to discuss!

Suppose three mice who are littermates have weights $82, 107,$ and $93g.$

1. What is the means weight of the mice?
2. How many possible bootstrap samples of this sample are there?
3. List all of the possible bootstrap samples as triples.
4. Compute the mean of each bootstrap sample.
5. Compute the mean of the resample means. How does this compare with the original sample mean?
6. What are the high and low values of the resample means?

More Problems from Chapter 3



3.1 A random sample of size n is taken *with* replacement from a population of size N. Show that the probability of having no repetitions in the sample is given by the product

$$\prod_{j=0}^{n-1}(1-\frac{j}{N}).$$

3.8 Suppose that y and z are independent random variables, with variances σ_y^2 and σ_z^2.

(a) Show that the variance of $y+z$ is the sum of the variances

$$\sigma_{y+z}^2=\sigma_y^2+\sigma_z^2\ . \qquad (3.35)$$

(In general, the variance of the sum is the sum of the variances for independent random variables $x_1, x_2, \cdots, x_n$.)

(b) Suppose $F \rightarrow (x_1, x_2, \cdots, x_n)$ where the probability distribution F has expectation μ and variance σ^2. Show that $\bar{x}$ has expectation μ and variance σ^2/n.