Sampling Variability

Chapter 4, Lab 1

OpenIntro Biostatistics

Topics

- Point estimates
- Sampling distribution of the sample mean
- Simulation

A natural way to estimate features of a population, such as a population mean, is to use the corresponding summary statistic calculated from a sample drawn from the population; a sample mean is a point estimate of a population mean. If a different sample is drawn, the new sample mean would likely be different as a result of sampling variability. This lab explores the relationship between point estimates and population parameters through simulation.

The material in this lab corresponds to Section 4.1 of *OpenIntro Biostatistics*.

Background information

This lab uses data from the Youth Risk Behavioral Surveillance System (YRBSS), a yearly survey conducted by the US Centers for Disease Control to measure health-related activity in high-school aged youth. The dataset `yrbss` contains responses from the 13,572 participants in 2013 for a subset of the variables included in the complete survey data.

Variables in `yrbss` include:

- `age`: age in years
- `gender`: gender of participant, recorded as either `female` or `male`
- `grade`: grade in high school (9-12)
- `height`: height, in meters (1 m = 3.28 ft)
- `weight`: weight, in kilograms (1 kg = 2.2 lbs)

The CDC used the response from the 13,572 students to estimate the health behaviors of the target population: the 21.2 million high school aged students in the United States in 2013.

The goal in this lab is to observe the effect of sampling by treating the 13,572 individuals in `yrbss` as a target population and drawing random samples. How do point estimates of mean weight, $\overline{x}_{weight}$, calculated from random samples compare to the population parameter, μ_{weight}?

Taking one sample

1. Run the following code to take a random sample of 10 individuals from yrbss and store the subset as yrbss.sample.

```
#load the data
library(oibiostat)
data("yrbss")

#set parameters
sample.size = 10

#obtain random sample of row numbers
set.seed(5011)
sample.rows = sample(1:nrow(yrbss), sample.size)

#create yrbss.sample
yrbss.sample = yrbss[sample.rows, ]
```

   a) Which rows of yrbss were sampled from?

The following ten rows were sampled from yrbss; the data from the respondents stored in these ten rows are stored as yrbss.sample.

```
sample.rows
```

```
##  [1]  7516 10731  3541 11356 11849  6411 10405 11864  3664  2881
```

   b) How many individuals of each gender have been sampled?

There are 8 female students and 2 male students in yrbss.sample.

```
table(yrbss.sample$gender)
```

```
##
## female   male
##      8      2
```

   c) What is the mean age of the sampled students?

The mean age of the sampled students is 16.5 years.

```
mean(yrbss.sample$age)
```

```
## [1] 16.5
```

   d) Calculate $\overline{x}_{weight}$ and s_{weight}, the mean and standard deviation of weight in the sample.

The mean weight in the sample is 68.13 kg and the standard deviation of weight in the sample is 12.03 kg.

```
mean(yrbss.sample$weight)
```

```
## [1] 68.131
```

```
sd(yrbss.sample$weight)
```

```
## [1] 12.03221
```

e) Calculate μ_{weight}, the mean weight in the yrbss population.

The mean weight in the yrbss population is 67.91 kg.

```
mean(yrbss$weight, na.rm = TRUE)
```

```
## [1] 67.9065
```

2. Take a new random sample of size 10 from yrbss, changing the seed to be the four digits representing your birth month and day (e.g., 1028 for October 28).

```
#load the data
library(oibiostat)
data("yrbss")

#set parameters
sample.size = 10

#obtain random sample of row numbers
set.seed(1028)
sample.rows = sample(1:nrow(yrbss), sample.size)

#create yrbss.sample
yrbss.sample = yrbss[sample.rows, ]
```

a) Use the summary() command on weight to check for any missing values, which are recorded as NA. Are there any missing values in your sample?

There are five missing values in this sample.

```
summary(yrbss.sample$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   65.77   68.49   68.95   71.03   74.84   77.11       5
```

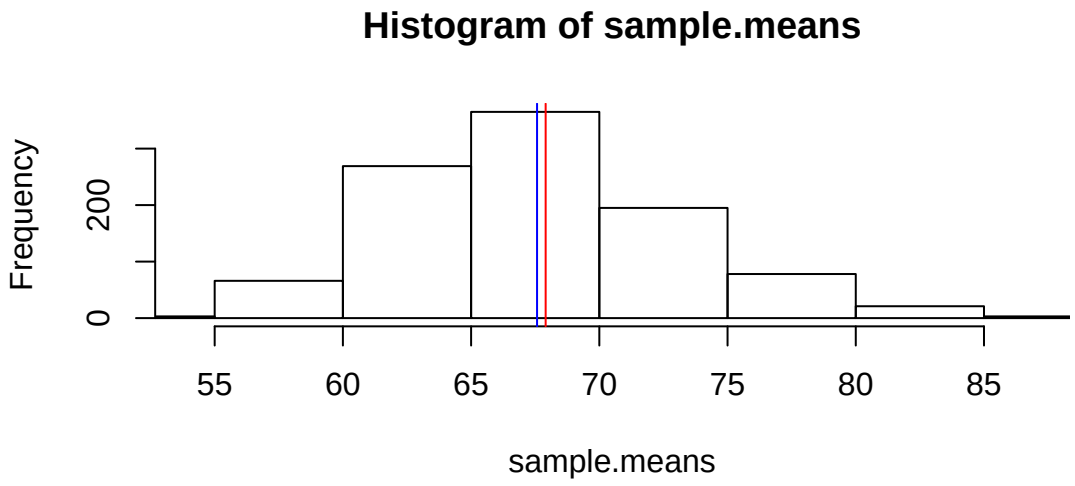
b) What is $\bar{x}_{weight}$, as calculated from your sample? Does it differ from $\bar{x}_{weight}$ as calculated in part d) of the previous question? How do these point estimates compare to the population mean μ_{weight}?

From this sample, $\bar{x}_{weight}$ is 71.03 kg. It is larger than both the sample mean calculated in part d) of the previous question (68.13 kg) and the population parameter μ_{weight} (67.91 kg).

Taking many samples

3. Run the following code to take 1,000 random samples of size 10 from yrbss. For each sample, the code calculates mean weight for participants in the sample and stores the value in the vector sample.means.

```r
#set parameters
sample.size = 10
replicates = 1000

#set seed for pseudo-random sampling
set.seed(5011)

#create empty vector to store results
sample.means = vector("numeric", replicates)

#calculate sample means
for(k in 1:replicates){

  sample.rows = sample(1:nrow(yrbss), sample.size)
  sample.means[k] = mean(yrbss$weight[sample.rows], na.rm = TRUE)

}

#create histogram of sample means
hist(sample.means, xlim = c(54, 87)) #xlim keeps the axis scale fixed

#draw a blue line at the mean of sample means
abline(v = mean(sample.means), col = "blue")

#draw a red line at the population mean weight in yrbss
abline(v = mean(yrbss$weight, na.rm = TRUE), col = "red")
```

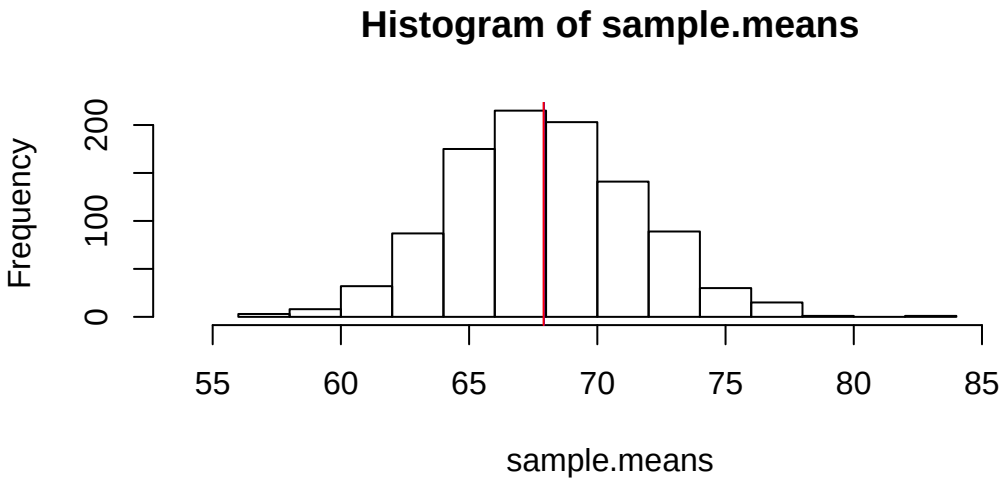


a) Describe the distribution of sample means.

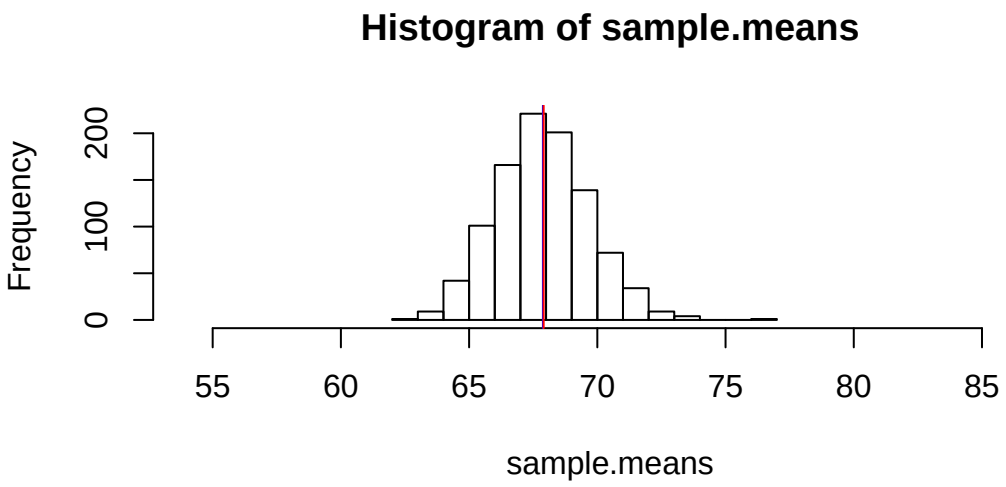
The distribution of sample means is somewhat symmetric, with some right skewing. The mean of sample means is about 67 kg, which is close to the population mean weight from `yrbss`.[1]

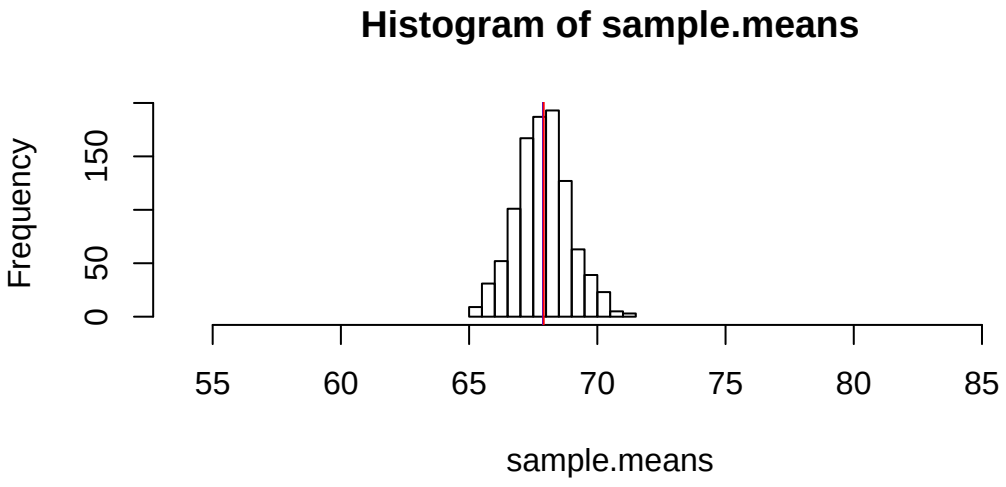
b) Explore the effect of larger sample sizes by re-running the code for sample sizes of 25, 100, and 300. Describe how the distribution of sample means changes as sample size increases.

As the sample size increases, the spread of the distribution decreases. With a sample size of 10, the sample means range from about 55 kg to 85 kg. Once sample size is 300, the sample means are all between 65 kg and 71 kg. It is difficult to discern from the plots as printed in the PDF, but zooming in on the plots in the *RStudio* figure pane reveals that as sample size increases, the mean of sample means (blue line) approaches the population mean weight (red line).



[1] The histogram shows a visual approximation of the theoretical sampling distribution of $\overline{X}$ with $n = 10$.





c) Recall that the goal of inference is to learn about characteristics of a target population through the information obtained by taking one sample. What is the advantage of a larger sample size?

Increasing sample size causes the distribution of $\overline{X}$ to be clustered more tightly around the population mean μ, allowing for more accurate estimates of μ from a single sample. When sample size is large, it is more likely that any one particular sample will have a mean close to the population mean.

d) From what you have observed in this exercise about the relationship between a point estimate for the mean $\overline{x}$ and the population mean (μ), evaluate the following statement:

"Since the mean weight of the 13,572 sampled high school students in the 2013 YRBSS survey is 67.91 kg, it is possible to definitively conclude that the mean weight of the 21.2 million high school students in the US in 2013 is also 67.91 kg."

It is not possible to conclude that the point estimate from the 13,572 sampled students is precisely equal to the population mean weight across the target population of 21.2 million students. As observed in the histograms, there is variation from sample to sample; it is quite possible that the mean of any one particular sample is not equal to the population mean weight.

A more defensible conclusion is to say that the sample mean of 67.91 kg is a reasonable estimate of the actual population mean weight.