Exploring sampling variability

Adapted from OpenIntro Biostatistics lab 1

6/22/2023

As we discussed, 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.

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 in years gender of participant, recorded as either female} or male} grade in high school (9-12) height, in meters (1 m = 3.28 ft) 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, $\bar{x}_{weight}$, 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
yrbss = read.csv('yrbss.csv')
#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?

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

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

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

Recall that to omit missing values before a summary statistic is computed, specify na.rm = TRUE.

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

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., for October 28).

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

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}?

Taking many samples

Set eval = TRUE in the code chunks for the results to display in the knitted html

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 .

```
#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  (you can also use `replicate` as we did in the poll exampl
e)
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.

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.

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?

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."