Confidence Intervals

Chapter 4, Lab 2

OpenIntro Biostatistics

Topics

- Interval estimates
- Simulation

The previous lab discussed calculating point estimates from samples drawn from a population. This lab introduces the calculation of interval estimates, i.e. **confidence intervals**. While a point estimate consists of a single value, a confidence interval gives a plausible range of values for a parameter.

The material in this lab corresponds to Section 4.2 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 interval estimates of mean weight, $(\overline{x}_{weight} - m, \overline{x}_{weight} + m)$, calculated from random samples compare to the population parameter, μ_{weight}?

Calculating confidence intervals

Run the following code to take a random sample of 30 individuals from `yrbss` and store the subset as `yrbss.sample`. The code includes a section that removes any rows from `yrbss` where there are no data recorded for `weight`. The version of `yrbss` without missing weight values is stored as `yrbss.complete`.

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

#remove rows with missing values
yrbss.complete = yrbss[complete.cases(yrbss$weight), ]

#set parameters
sample.size = 30

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

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

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

1. A confidence interval is calculated from four quantities: the sample mean $\overline{x}$, the sample standard deviation s, the sample size n, and the critical z-value $z^\star$.

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

   b) For a 95% confidence interval, $z^\star$ is the point on a standard normal distribution that has area 0.975 to the left (and area 0.025 to the right). Calculate the value of $z^\star$ for a 95% confidence interval.

   c) Calculate a 95% confidence interval based on the sampled weights. The quantity $(z^\star \times \frac{s}{\sqrt{n}})$ is known as the margin of error, m.

$$\overline{x} \pm z^\star \times \frac{s}{\sqrt{n}} \rightarrow \left(\overline{x} - z^\star \frac{s}{\sqrt{n}},\ \overline{x} + z^\star \frac{s}{\sqrt{n}}\right)$$

   d) The standard deviation of weight in the sample is 18.26 kg. Suppose that the standard deviation in the sample were 20 kg or 25 kg, but that $\overline{x}$ and n remain constant. Re-run the calculation from part c) and describe the effect of larger (sample) standard deviation on the confidence interval.

2. In general, for a confidence interval of $(1-\alpha)(100)\%$, $z^\star$ is the point on a standard normal distribution that has area $1-(\alpha/2)$ to the left (and area $\alpha/2$ to the right). For a 95% confidence interval, $\alpha = 0.05$; $z^\star$ is the point on a standard normal distribution with area $1-(0.05/2) = 0.975$ to the left.

a) Calculate a 90% confidence interval for mean weight based on the sample data.

b) Calculate a 99% confidence interval for mean weight based on the sample data.

c) Compare the 95% confidence interval calculated in the previous question to the 90% and 99% confidence intervals. Describe the relationship between confidence level and width of the interval.

d) Which of the intervals calculated (90%, 95%, 99%) do you find to be the most informative as an estimate of the mean weight of high school age students in the US? Explain your answer.

3. The `t.test()` command can be used to calculate confidence intervals. For example, the command to calculate a 95% confidence interval for `height` in `yrbss.complete` is

```
t.test(yrbss.complete$height, conf.level = 0.95)$conf.int
```

a) Calculate a 95% confidence interval for mean weight using `t.test()`.

The answer will differ slightly from the one in Question 1 because R calculates confidence intervals using a critical value from the t distribution rather than from the standard normal distribution; the t distribution will be introduced in Unit 5.

b) Examine the effect of larger sample sizes on the confidence interval by re-running the code for sample sizes of 50, 100, and 300. Describe your observations.

The interpretation of "confidence"

4. The method discussed for computing an x% confidence interval will produce an interval that x times out of 100 (on average) contains the population mean.

a) Consider the individuals in `yrbss.complete` as the target population. Calculate the population mean weight, μ_{weight}.

b) Does the 95% interval calculated in part b) of Question 3 for $n = 100$ contain μ_{weight}?

5. Run the following code to take 1,000 random samples of size 100 from `yrbss.complete`. For each sample, R calculates mean weight for participants in the sample and stores the value in the vector `sample.means`. The margin of error m is calculated according to the defined confidence level and stored in the vector `m`. The logical variable `contains.mu` records `TRUE` if a confidence interval contains μ_{weight} and `FALSE` otherwise.

```
#set parameters
sample.size = 100
conf.level = 0.95
replicates = 1000

#set seed for pseudo random sampling
set.seed(2017)

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

```r
#calculate sample means and margins of error
for(k in 1:replicates){

  sample.rows = sample(nrow(yrbss.complete), sample.size)

  z.star = qnorm(1 - (1 - conf.level)/2)

  sample.means[k] = mean(yrbss.complete$weight[sample.rows])
  m[k] = z.star * (sd(yrbss.complete$weight[sample.rows]) / sqrt(sample.size))

}

#define upper and lower bounds of confidence interval
ci.lb = sample.means - m
ci.ub = sample.means + m

#does the confidence interval contain mu?
mu = mean(yrbss.complete$weight)
contains.mu = (ci.lb < mu) & (ci.ub > mu)
table(contains.mu)
```

a) How many intervals contain the population mean μ_{weight}?

b) Re-run the simulation with confidence levels of 0.90 and 0.99. What happens to the number of intervals that contain μ_{weight}?

c) What is a disadvantage associated with using an interval that gives a more precise estimate of the parameter; e.g., a 90% interval rather than a 95% interval?

d) From what you have observed in this lab about the relationship between an interval estimate ($\bar{x} \pm m$) and the population mean (μ), evaluate the following statement:

"The 95% confidence interval as calculated from the 13,572 sampled high school students in the 2013 YRBSS survey is (67.61, 68.20) kg. It is possible to definitively conclude that this interval contains the mean weight of the 21.2 million high school students in the US in 2013."