Homework 7

Rui Gao

2021-11-14

Instructions

Use the R Markdown version of this file to complete and submit your homework. Make sure you change the author in the header to your own name.

R Simulation

This question asks you to examine the robustness of t-based **confidence intervals** to small sample sizes and non-Normal distributions. You should complete the lab before attempting this question as it walks you through the code you'll need to complete these tasks.

Consider two populations:

1. Uniform(0, 5), which has a mean of 2.5
2. Uniform(0, 1), which has a mean of 0.5

a) **Edit the function `sim_ci()`** from lab to:

- draw a sample of size `n_1` from a population 1, and
- draw a sample of size `n_2` from a population 2, then

perform a t-test with `t.test()`, extract the 95% confidence interval, and return `TRUE` if the interval contains the true difference in means, and `FALSE` if it does not.

```
sim_ci <- function(n_1, n_2){
  delta <- 0
  # 1. Generate data
  sample_1 <- runif(n = n_1, min = 0, max = 5)
  sample_2 <- runif(n = n_2, min = 0, max = 1)
  # 2. Run `t.test()`
  test_result <- t.test(x = sample_1, y = sample_2)
  # 3. Extract CI
  ci <- test_result$conf.int
  # 4. Check if delta is in CI
  lower <- ci[1]
  upper <- ci[2]
  lower < delta & upper > delta
}
```

```r
# Verify the function returns TRUE or FALSE
sim_ci(n_1 = 5, n_2 = 5)
```

```
## [1] FALSE
```

b) Use your function, along with `replicate()` and `mean()` to find the proportion of t-based 95% confidence intervals in 50,000 simulations that contain the true difference in means, when both samples have a sample size of 5.

```r
ci_covers <- replicate(50000, sim_ci(n_1 = 5, n_2 = 5))
mean(ci_covers)
```

```
## [1] 0.40152
```

c) Now repeat (b) for sample sizes that are both 10, 25 and 50. ci_covers <- replicate(50000, sim_ci(n_1 = 5, n_2 = 5))

```r
ci_covers_c_10 <- replicate(50000, sim_ci(n_1 = 10, n_2 = 10))
mean(ci_covers_c_10)
```

```
## [1] 0.01202
```

```r
ci_covers_c_25 <- replicate(50000, sim_ci(n_1 = 25, n_2 = 25))
mean(ci_covers_c_25)
```

```
## [1] 0
```

```r
ci_covers_c_50 <- replicate(50000, sim_ci(n_1 = 50, n_2 = 50))
mean(ci_covers_c_50)
```

```
## [1] 0
```

d) Based on the simulation results, summarize the robustness of the t-based confidence interval when the normality assumption does not hold (in two sentences): describe what is happening to the coverage of the confidence intervals as the sample size increases, and explain why this occurs.
As the number of size increase, the proportion is closer to 0%. The robustness is not obvious here. Because the range for two sample is different.

Conceptual Questions

Each of the following scenarios describe a study that violates one of the assumptions of the proposed analysis.

For **each** scenario:

a) Describe which assumption is most likely violated, and the evidence you have for the violation.

1. Cluster effects, most of them should be Anglo-Saxon. Proximity effects, they should have similar ages.

2. I don't know for the second scenario.

3. Proximity effects, the residents in Oregon should be affected by similar environmental elements.

b) Comment on whether we should expect any robustness from the procedure against the violation.

1. No, the height of you is very personal. It could change a lot.Although the height could be a requirement when you looking for your partner.

2. Yes, the quantity and quality of food you take are highly relative with your income.

3. Be honestly, I don't know. But I want to say "Yes",due to "the income of men and women is unfair" is a classic topic for many years.

c) Regardless of the robustness, make a suggestion for how the study or analysis could be improved to diminish (or remove entirely) the effect of the violation of the assumption.

1. Get more data.How about all couples in the British? I think the government should own this data.

2. Use more complicated methods. I don't know, so I leave this method here.

3. Use a transformation for help. Comparing with the above two, the gap of number in this case is too large.

Histograms for the data in all three cases are provided separately in `homework-07-histograms.pdf`.

1. The Great Britain Office of Population Census and Surveys collected data on a random sample of 170 married, opposite sex, couples in Britain, recording the age (in years) and heights (in cm) of the husbands and wives.

They conduct a two-sample t-test to compare the mean height of husbands to the mean height of wives. Histograms of the heights are provided in `homework-07-histograms.pdf`.

2. In a study on the differences in diet between high and low income households, 50 low income and 50 high income households are randomly selected. Every adult in each household records their calorific intake for one week, and this is summarized to a daily average for each person. In total there are 110 adults in the high income households and 96 adults in the low income houses. The mean *average daily caloric intake* is compared between adults living in low and high income households using a two sample t-test.

Histograms of the average calorie intakes in the study are provided in `homework-07-histograms.pdf`.

3. In an effort to quantify gender inequality in income, the State of Oregon collects a random sample of 2000 residents with comparable qualifications and years of experience (in practice this is really hard to do, but for the purpose of this problem assume it was done well). They compare the mean income of females to the mean income of males using a two-sample t-test.

Histograms of the incomes are provided in `homework-07-histograms.pdf`.