Homework 7

Nora Quick

2021-11-09

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 <- rnorm(n = n_1, mean = 2.5, sd = 4)
  sample_2 <- rnorm(n = n_2, mean = 0.5, sd = 2)
  # 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] TRUE
```

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
rep <- replicate(50000, sim_ci(n_1 = 5, n_2 = 5))
m <- mean(rep)
m
```

```
## [1] 0.8653
```

c) Now repeat (b) for sample sizes that are both 10, 25 and 50.

```r
ten <- replicate(10, sim_ci(n_1 = 5, n_2 = 5))
m1 <- mean(ten)
m1
```

```
## [1] 0.8
```

```r
tfive <- replicate(25, sim_ci(n_1 = 5, n_2 = 5))
m2 <- mean(tfive)
m2
```

```
## [1] 0.84
```

```r
fifty <- replicate(50, sim_ci(n_1 = 5, n_2 = 5))
m3 <- mean(fifty)
m3
```

```
## [1] 0.88
```

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 sample size increases the proportion that are TRUE (the proportion of 95% CI that contain a true value) decreases which raises suspicion that it is not valid. This is happening because the more we test it the more FALSEs we are getting meaning it is not very robust.

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.

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

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.

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`.

a) The assumption that there is a normal distribution of ages is invalid. We don't know the spread of ages based on the graphs we're given and age does matter for height.
b) I expect robustness from this data because it seems normally distributed, there is an equal number of men and women.
c) I would watch the distribution of ages to make sure that is normally distributed as well and gain a larger sample size to reduce as much variabce as possible.

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`.

a) The assumption that the population spread is equal is invalid. There needs to be an equal amount from both tests for the two sample test to be valid.
b) Yes, I think there will still be robustness due to the sample sizes being relatively substantial and a relatively normal distribution.
c) I would choose the same number of people and people who shop at the same stores/restaurants/etc. Higher income people have more access to more healthy food and more food in general (typically) so keeping within the same shoping area would help determine a better calorie intake.

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`.

a) The assumption that the data is normally distributed is violated. The proof behind this is the graphs provided in the pdf.
b) We should not expect any robustness due to the lack of normally distributed data. There are also some large outliers that can distrupt the robustness.
c) Instead of residents with comparable training and qualifications I would do people with the same job. Training and qualifications don't always matter to the job you are in. Someone could get an entry level job with the same qualifications as a senior person depending on jobs available, applied to, etc. There could also be women who have training and qualifications who choose to become mothers which effect their lifelyhood (same with men if they so chose).