Chapter 4 Workshop

Table of contents

Dataset Prestige

We will again be using a well-known dataset called Prestige from the car R package. This dataset deals with prestige ratings of Canadian Occupations. The Prestige dataset has 102 rows and 6 columns. The observations are occupations.

This data frame contains the following columns:

- **education** - Average education of occupational incumbents, years, in 1971.
- **income** - Average income of incumbents, dollars, in 1971.
- **women** - Percentage of incumbents who are women.
- **prestige** - Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.
- **census** - Canadian Census occupational code.
- **type** - Type of occupation. A factor with levels: bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar. (includes four missing values).

Exercise 4.1

Perform a one-sample t-test to test the hypothesis that the true mean prestige is exactly 50.

```r
library(tidyverse)
library(car)
data(Prestige)

# Alternative hyp: greater or less than 50
t.test(Prestige$prestige, mu=50)

# Alternative hyp: greater than 50
t.test(Prestige$prestige, mu=50, alternative="greater")
```

Exercise 4.2

Test whether the true mean prestige score for professionals is 50% more than the true mean prestige score for white collar occupations.

```
prof.data <- Prestige |>
  filter(type=="prof") |>
  pull(prestige)

wc.data <- Prestige |>
  filter(type=="wc") |>
  pull(prestige)

t.test(prof.data,
       wc.data,
       mu = 0.5 * mean(wc.data),
       alternative = 'greater')
```

Exercise 4.3

Explore the skewness in the income variable using a boxplot, mids-vs-spread plot. Compute the D-Statistics. Obtain a suitable power transformation to correct the skewness. Compute the 95% confidence interval for the true mean Top measurement using the raw and transformed data.

```
Prestige |>
  ggplot() +
  aes(income) +
  geom_boxplot()
```

```
# or
boxplot(Prestige$income, horizontal = TRUE)
```

```
# D-Stat codes under a few shrinking transformations

D1 = function(x) {
(mean(x) - median(x)) / sd(x)
}
```

```
D2 = function(x) {
(mean(x) - median(x)) / (fivenum(x)[4] - fivenum(x)[2])
}

D3 = function(x) {
((fivenum(x)[4] + fivenum(x)[2]) / 2+-median(x)) / (fivenum(x)[4] - fivenum(x)[2])
}

x = Prestige$income

VMat <- cbind(
  Vreci = -1 / x,
  V = x,
  VSq = sqrt(x),
  VLog = log(x)
  )

apply(VMat, 2, D1)
apply(VMat, 2, D2)
apply(VMat, 2, D3)

# or obtain the D-stats individually
D1(sqrt(x)); D2(sqrt(x)); D3(sqrt(x))
```

```
library(lindia)

gg_boxcox(lm(x ~ 1))
```

```
# or
require(MASS)
b <- boxcox(x ~ 1)
title("Log-likelihood curve of boxcox parameter")
k <- b$x[which.max(b$y)]
mtext(paste("optimum power=", formatC(k)))
```

```
t.test(x)
t.test(log(x))
```

More R code examples are here