Lab2

2024-09-06

Stat 5000Lab #2
Fall 2024 Due Tue Sep 10th Name: Sam Olson

Directions: Complete the exercises below. When you are finished, turn in any required files online in Canvas, then check-in with the Lab TA for dismissal.

Introduction to t-Tests in R

Refer to the `fuel_economy.csv` data file posted in Canvas. This data set has information about an observational study of automobiles driven in Canada, including the following two columns:

`Cylinders:` category variable with two levels - 4 or 6

`Consumption:` numeric response variable with the fuel consumption in miles per gallon (mpg)

Researchers are interested in exploring whether there is a difference in the average fuel consumption of vehicles with engines built using differing numbers of cylinders. The code to conduct a two-sample t-test in R is explained below. The full R program is provided in the file `fuel_economy_Lab2.R` posted on Canvas.

- First, load in the data using the *Import Dataset* tool in R Studio. Be sure to change the variable type on the Cylinders column to "factor" and enter"4, 6" as the levels.

```
library(readr)
fuel <- read_csv("fuel_economy.csv",
                 col_types=cols(Cylinders=col_factor(levels=c("4", "6"))))
View(fuel)
```

- Next, compute the corresponding summary statistics within in group.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
sum_stats = fuel |>
    group_by(Cylinders) |>
    summarize(
        Y_n = n(),
        Y_mean = mean(Consumption.mpg),
        Y_sd = sd(Consumption.mpg)
    )
sum_stats
```

```
## # A tibble: 2 x 4
##   Cylinders   Y_n Y_mean  Y_sd
##   <fct>     <int>  <dbl> <dbl>
## 1 4          2111   34.1  6.67
## 2 6          1041   24.5  2.55
```

- Then, use the `t.test()` function to conduct a test for the difference in mean fuel consumption between 4 and 6 cylinder vehicles. Indicate the response variable name before the ~ and the category variable name after, use the `data` option to provide the name of the dataset, and use the `var.equal` option set to "TRUE" to indicate the population variances are assumed equal.

```r
HT = t.test(Consumption.mpg~Cylinders, data=fuel, var.equal=TRUE)
HT
```

```
##
##  Two Sample t-test
##
## data:  Consumption.mpg by Cylinders
## t = 44.664, df = 3150, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
## 95 percent confidence interval:
##  9.148005 9.988067
## sample estimates:
## mean in group 4 mean in group 6
##        34.08100        24.51297
```

You can see what pieces of information are stored in the `HT` variable using the `names()` function. You can access these pieces of information using the `$` operator, e.g.

```r
names(HT)
```

```
##  [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
##  [6] "null.value"  "stderr"      "alternative" "method"      "data.name"
```

```r
HT$null.value
```

```
## difference in means between group 4 and group 6
##                                                0
```

Assignment

1. State the hypotheses for the two-sided test.

$H_0 : \mu_{\text{4Cylinder}} = \mu_{\text{6Cylinder}}$, with $H_A : \mu_{\text{4Cylinder}} \neq \mu_{\text{6Cylinder}}$, descriptively stated as: Null Hypothesis is the mean consumption (mpg) of 4 Cylinder cars is equal to the mean consumption of 6 Cylinder cars with alternate hypothesis that the mean consumption of 4 Cylinder cars is **not equal** to the mean consumption of 6 Cylinder cars.

2. From the output, find/compute the difference in the two sample means.

```r
HT = t.test(Consumption.mpg~Cylinders, data=fuel, var.equal=TRUE)
HT
```

```
##
##  Two Sample t-test
##
## data:  Consumption.mpg by Cylinders
## t = 44.664, df = 3150, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
## 95 percent confidence interval:
##  9.148005 9.988067
## sample estimates:
## mean in group 4 mean in group 6
##        34.08100        24.51297
```

```r
sampleMean1 <- 34.08100
sampleMean2 <- 24.51297
difference <- sampleMean1 - sampleMean2
difference
```

```
## [1] 9.56803
```

The difference in the two sample means is ≈ 9.568 mpg (4 Cylinder - 6 Cylinder).

3. From the output, find/compute the estimate of the pooled standard deviation.

Formula:

$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}$

```r
library(dplyr)
data1 <- fuel %>%
  filter(fuel$Cylinders == 4)

data2 <- fuel %>%
  filter(fuel$Cylinders == 6)

#Step 2: Finding standard deviation
s1 <- sd(data1$Consumption.mpg)
s2 <- sd(data2$Consumption.mpg)

#Step 3: Finding sample size
n1 <- length(data1$Consumption.mpg)
n2 <- length(data2$Consumption.mpg)

#Step 4: Calculate pooled standard deviation
# pooled <- sqrt( ((n1-1)*(s1^2) + (n2-1)*(s2^2)) / (n1+n1-2) )
# pooled

numerator <- (n1-1)*(s1^2) + (n2-1)*(s2^2)
denom <- n1 + n2 - 2
pooled <- sqrt( numerator / denom )
pooled
```

```
## [1] 5.65644
```

```r
# sd(fuel$Consumption.mpg)
```

From the above this yields an estimated pooled standard deviation of ≈ 5.65644 mpg.

4. From the output, find/compute the test statistic for the hypothesis test.

```
library(dplyr)
m1 <- mean(data1$Consumption.mpg)
m2 <- mean(data2$Consumption.mpg)

manualTStatistic <- (m1 - m2) / (pooled * sqrt((1/n1) + (1/n2)))
(m1 - m2) / (5.65644 * sqrt((1/n1) + (1/n2)))
```

```
## [1] 44.66383
```

```
HT$statistic
```

```
##        t
## 44.66384
```

The test statistic for this hypothesis test is ≈ 44.66384

5. From the output, find/compute the degrees of freedom for the test.

```
HT$parameter
```

```
##   df
## 3150
```

$\text{df} = n_1 + n_2 - 2 = 2111 + 1041 - 2 = 3150$

6. From the output, find/compute the p-value for the two-sided hypothesis test.

```
HT
```

```
##
##  Two Sample t-test
##
## data:  Consumption.mpg by Cylinders
## t = 44.664, df = 3150, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
## 95 percent confidence interval:
##  9.148005 9.988067
## sample estimates:
## mean in group 4 mean in group 6
##        34.08100        24.51297
```

```
HT$p.value
```

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

$p - value < 2.2e - 16$ or ≈ 0 (so small R rounds it to 0).

7. Interpret the results of the two-sided test in the context of the research question.

The low p-value means we estimate with very low probability the observed results given the null hypothesis is true (no difference in average fuel consumption between the Cylinder types). This provides support against null hypothesis and possible support of the alternative hypothesis; overall, we have limited evidence to support there being a difference in average fuel consumption (mpg) between 4 and 6 cylinder vehicles (two-tailed alternative hypothesis).

Note: I use "limited evidence" not as a distinction of the (very) low p-value observed, so much as to emphasize this is one p-value we've calculated and present as evidence.

8. By default, R conducts the two-sided hypothesis test. You can change this by adding the parameter "`alternative=greater`" or "`alternative=less`" inside the `t.test()` function. Provide a research question corresponding to either the "greater" or"less" one-sided test.

```
altHT <- t.test(Consumption.mpg~Cylinders, data=fuel, var.equal=TRUE, alternative = "greater")
altHT
```

```
##
##  Two Sample t-test
##
## data:  Consumption.mpg by Cylinders
## t = 44.664, df = 3150, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 4 and group 6 is greater than 0
## 95 percent confidence interval:
##  9.215566      Inf
## sample estimates:
## mean in group 4 mean in group 6
##        34.08100        24.51297
```

For input "alternative = "greater"":

Research Question: Do 4 Cylinder vehicles have greater mpg than 6 Cylinder vehicles, as in are 4 Cylinder vehicles **more fuel efficient** than 6 Cylinder vehicles?

By adding this to the input we change the alternative hypothesis to a one-tailed hypothesis, specifically an alternate hypothesis that the average mpg of 4 cylinder vehicles is higher (greater than) the average mpg of 6 cylinder vehicles. (Though not explicitly called for, it is worth noting that this hypothesis provides a similar p-value and interpretation to the above, but particularly in support of the conclusion that 4 cylinder vehicles have greater mpg, or on average more efficient fuel consumption.)

Total: 25 points # **correct: %:**