Inferential Statistics

ECO 6416

2022-09-27

Contents

Here are all the packages needed to get started.

```
library(readxl)



sessionInfo()
```

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19044)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] readxl_1.4.1
##
## loaded via a namespace (and not attached):
##  [1] digest_0.6.29    cellranger_1.1.0 magrittr_2.0.3   evaluate_0.15
##  [5] rlang_1.0.3      stringi_1.7.8    cli_3.3.0        rstudioapi_0.14
##  [9] rmarkdown_2.14   tools_4.2.1      stringr_1.4.0    xfun_0.31
## [13] yaml_2.3.5       fastmap_1.1.0    compiler_4.2.1   htmltools_0.5.2
## [17] knitr_1.39
```

1 Univariate

If we want to compare the mean of our sample and make a statement about the population, we can run a very simple t-test. We are going to use the motor trend dataset within R.

```
t.test(mtcars$mpg)
```

```
##
##  One Sample t-test
##
## data:  mtcars$mpg
## t = 18.857, df = 31, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  17.91768 22.26357
## sample estimates:
## mean of x
##  20.09062
```

This test shows that there is sufficient evidence to reject the null hypothesis and state that the average miles per gallon on a vehicle is different than zero.

Not very surprising. If the average car had 0 miles per gallon, it wouldn't be a useful car.

Let's try something different. Suppose the null hypothesis is that the population mean is:

$$H_0 : \mu = 22$$

and the alternative is:

$$H_A : \mu \neq 22$$

```
t.test(mtcars$mpg, mu = 22)
```

```
##
##  One Sample t-test
##
## data:  mtcars$mpg
## t = -1.7921, df = 31, p-value = 0.08288
## alternative hypothesis: true mean is not equal to 22
## 95 percent confidence interval:
##  17.91768 22.26357
## sample estimates:
## mean of x
##  20.09062
```

At the 95% confidence level, there is not sufficient evidence to claim that the mean miles per gallon is different than 22 mpg (fail to reject).

Let's see how we can manipulate it so that we get statistical significance. Let's test this alternative hypothesis:

$$H_A : \mu < 22$$

```
t.test(mtcars$mpg, mu = 22, alternative = "less")
```

```
##
##  One Sample t-test
```

```
##
## data:  mtcars$mpg
## t = -1.7921, df = 31, p-value = 0.04144
## alternative hypothesis: true mean is less than 22
## 95 percent confidence interval:
##      -Inf 21.89707
## sample estimates:
## mean of x
##  20.09062
```

since we did a one-tail test, we cut the p-value in half. Now we get statistical significance. You can see how easy it is to manipulate results.

2 Regression - First Class Mail Volume

This one comes from previous exercises about first class mail volume and some potential factors from it.

```
mail <- read_excel("../Data/First Class Mail.xlsx")
```

Let's just dive right into the meat and potatoes. You can reference all the proper steps to do before this analysis in the previous content.

2.1 Hypothesis Testing

```
model <- lm(FirstClVol ~ Time + PopUSA + Price, data = mail)

regsummary <- summary(model) # I do this so that I can pull values later on

regsummary
```

```
##
## Call:
## lm(formula = FirstClVol ~ Time + PopUSA + Price, data = mail)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.721  -9.379  -2.381  13.465  29.161
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 430.2370   137.2978   3.134 0.002747 **
## Time          8.3667     2.2221   3.765 0.000401 ***
## PopUSA       -2.2099     0.7812  -2.829 0.006472 **
## Price        -2.1365     1.3565  -1.575 0.120894
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.34 on 56 degrees of freedom
## Multiple R-squared:  0.5002, Adjusted R-squared:  0.4734
## F-statistic: 18.68 on 3 and 56 DF,  p-value: 1.605e-08
```

2.2 Overall Model Significance

Based on the results above, we can conclude at at least one of the coefficients are different than zero. You can simply look at the p-value in the above output, or find it this way:

```
regsummary$fstatistic # To see what this results in
```

```
##    value    numdf    dendf
## 18.67956  3.00000 56.00000
```

```
pf(regsummary$fstatistic[1], regsummary$fstatistic[2], regsummary$fstatistic[3], lower.tail = FALSE)
```

```
##       value
## 1.60548e-08
```

2.3 Individual t-test Results

You can refer to this and see that Time and Population test statistically significant. This means that we have sufficient evidence in that the slopes of the coefficients are different than zero.

What is interesting about this dataset is that we actually see a negative relationship between population and mail volume. Although this is statistically significant, we will not throw out our underlying theory. In this case, we will have to discuss this phenomenon later.

There is insufficient evidence to conclude that the slope coefficient of price is different than zero in a two tailed test.

We might be able to see if it tests significant on a one-tailed test. We can justify a one tailed test because we have theory to support the claim that it should be negative (law of demand).

```
regsummary$coefficients["Price",4]/2 #[,4] is the p-value
```

```
## [1] 0.06044709
```

Still insufficient evidence at the 95% level, but at a confidence level of below $\approx 94\%$ we would have statistically significant results.

2.4 Marginal Change Analysis

Simply hand calculate it using the output.

2.4.1 Bonus: Way to Calculate this automatically

There are more complicated ways of getting marginal analysis done automatically. I am playing with building a function that does it automatically, so it is in beta form. It is assuming you have already built the summary object like I just did. You'll also have to check and make sure you can even do this type of analysis first. If it is not statistically significant, you cant do this analysis.

```
thing2 <- function(summaryobject, independent, cv, change){
  meanchange <- summaryobject$coefficients[independent,1]*change
  moe <- summaryobject$coefficients[independent,2]*cv*change

  print(paste0("The expected change due to a change in ",
               round(change,2),
               " is ",
               round(meanchange,2),
               " give or take ",
               round(moe,2),
               "." ))
  print(paste0("Put another way: As low as ",
               round(meanchange - moe,2),
               " and as high as ",
               round(meanchange + moe,2),
```

```
                "." ))
  }


thing2(regsummary,"Time", 2, 4)
```

```
## [1] "The expected change due to a change in 4 is 33.47 give or take 17.78."
## [1] "Put another way: As low as 15.69 and as high as 51.24."
```