Computers analysis

George Mount

Does the presence of a CD-ROM affect sales price?

Introduction

CD-ROMs seem to be the hot new thing, etc., etc. . . .

Let's call in all of our packages and get started:

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
##
## Attaching package: 'psych'
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

Data

The source of this data is the *Journal of Applied Econometrics*. It came to us in an Excel file:

```
computers <- read_excel("../../datasets/computers.xlsx")
computers
```

```
## # A tibble: 6,259 x 11
##       id price speed    hd   ram screen cd    multi premium   ads trend
##    <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <chr> <chr> <chr>   <dbl> <dbl>
##  1     1  1499    25    80     4     14 no    no    yes        94     1
##  2     2  1795    33    85     2     14 no    no    yes        94     1
##  3     3  1595    25   170     4     15 no    no    yes        94     1
##  4     4  1849    25   170     8     14 no    no    no         94     1
##  5     5  3295    33   340    16     14 no    no    yes        94     1
##  6     6  3695    66   340    16     14 no    no    yes        94     1
##  7     7  1720    25   170     4     14 yes   no    yes        94     1
##  8     8  1995    50    85     2     14 no    no    yes        94     1
##  9     9  2225    50   210     8     14 no    no    yes        94     1
## 10    10  2575    50   210     4     15 no    no    yes        94     1
## # ... with 6,249 more rows
```

There are 6259 rows and 11 columns.

The dataset has a mean of 2219.58.

```
describe(computers$price)
```

```
##    vars    n    mean    sd median trimmed    mad min  max range skew kurtosis
## X1    1 6259 2219.58 580.8   2144 2182.58 593.04 949 5399  4450 0.71     0.73
##      se
## X1 7.34
```

Methods

We will conduct an independent samples t-test at the 95% confidence level. Our hypothesis: **there is no difference in sales price of computers with and without a CD rom**.

First we'll visually inspect the distribution of price by CD to confirm the CLT is likely to apply.

Results

Visualization

It looks good:

```
describe(computers$price)
```

```
##    vars    n    mean    sd median trimmed    mad min  max range skew kurtosis
## X1    1 6259 2219.58 580.8   2144 2182.58 593.04 949 5399  4450 0.71     0.73
##      se
## X1 7.34
```

T-test

The results of the t-test are as follows:

```
cd_test <- t.test(price ~ cd, data = computers)
tidy(cd_test)
```

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic  p.value parameter conf.low conf.high
##      <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>    <dbl>     <dbl>
## 1    -230.     2113.     2343.     -16.1 5.24e-57     6257.    -258.     -202.
## # ... with 2 more variables: method <chr>, alternative <chr>
```

Conclusion

CD-ROMs seem to be the hot new thing. Are there other things influencing it?

Maybe premium computers tend to have CD Roms, and *that's* what is really affecting the price.

```
head(computers)
```

```
## # A tibble: 6 x 11
##      id price speed    hd   ram screen cd    multi premium   ads trend
##   <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <chr> <chr> <chr>   <dbl> <dbl>
## 1     1  1499    25    80     4     14 no    no    yes        94     1
## 2     2  1795    33    85     2     14 no    no    yes        94     1
## 3     3  1595    25   170     4     15 no    no    yes        94     1
## 4     4  1849    25   170     8     14 no    no    no         94     1
## 5     5  3295    33   340    16     14 no    no    yes        94     1
```

```
## 6     6  3695    66   340    16     14 no    no    yes        94     1
```