In Class Exercise 1

ECO 6416

2024-06-28

Contents

Here are all the packages needed to get started.

```
library(readxl) # to read in xlsx files
options(scipen = 999) # to remove scientific notation

sessionInfo()
```

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## 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_1.4.2
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.2.1   fastmap_1.1.1    cli_3.6.1        tools_4.2.1
##  [5] htmltools_0.5.5  rstudioapi_0.14  yaml_2.3.7       cellranger_1.1.0
##  [9] rmarkdown_2.21   knitr_1.42       xfun_0.39        digest_0.6.31
## [13] rlang_1.1.1      evaluate_0.21
```

1 National Park Visitor Data

Go to Webcourses and download the `national_park_attendance.xlsx` file. Be sure to save the file in an easy to locate folder.

```
parks <- read_xlsx("C:/Users/jo585802/Downloads/national_park_attendance.xlsx")
```

```
head(parks)
```

```
## # A tibble: 6 x 2
##   `National Park`                          `Number of Visitors, 2017`
##   <chr>                                                         <dbl>
## 1 1. Great Smoky Mountains National Park                      11388893
## 2 2. Grand Canyon National Park                                6254238
## 3 3. Zion National Park                                        4504812
## 4 4. Rocky Mountain National Park                              4437215
## 5 5. Yosemite National Park                                    4366890
## 6 6. Yellowstone National Park                                 4116524
```

```
tail(parks)
```

```
## # A tibble: 6 x 2
##   `National Park`                        `Number of Visitors, 2017`
##   <chr>                                                       <dbl>
## 1 54. Katmai National Park                                    37818
## 2 55. North Cascades National Park                            30326
## 3 56. Isle Royale National Park                               28196
## 4 57. Lake Clark National Park                                22755
## 5 58. Kobuk Valley National Park                              15500
## 6 59. Gates of the Arctic National Park                       11177
```

1.1 Dimensions of the Data

Describe the three dimensions of the data in the above table (circle correct answers)

- univariate / bivariate / multivariate
- two categorical / two quantitative / one categorical one quantitative
- cross section / time series

1.2 Confidence Interval

```
summary(parks$`Number of Visitors, 2017`)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    11177   304003   642809  1443314  1789159 11388893
```

```
sd(parks$`Number of Visitors, 2017`)
```

```
## [1] 1946009
```

a) Give the output above, what is the best prediction of attendance at any park at the 95% level of confidence? Show the result as (low, high) and round to the nearest 100 thousand.

b) How comfortable are you with your prediction? Justify your answer.

1.3 Choosing Measure of Center

Which statistic provides the best measure of center? You can use the chart below to help.

```
hist(parks$`Number of Visitors, 2017`,
     main = "Histogram of Number of Visitors, 2017",
     xlab = "Number of Visitors, 2017")
```