Basic Statistics in R

ECO 6416

2022-08-27

Contents

Here are all the packages needed to get started.

```
library(gt)
library(tidyverse)
library(gtsummary)
library(plotly)



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] plotly_4.10.0   gtsummary_1.6.1 forcats_0.5.2   stringr_1.4.0
##  [5] dplyr_1.0.9     purrr_0.3.4     readr_2.1.2     tidyr_1.2.0
##  [9] tibble_3.1.7    ggplot2_3.3.6   tidyverse_1.3.2 gt_0.7.0
##
## loaded via a namespace (and not attached):
##  [1] lubridate_1.8.0     assertthat_0.2.1    digest_0.6.29
##  [4] utf8_1.2.2          R6_2.5.1            cellranger_1.1.0
```

```
##  [7] backports_1.4.1     reprex_2.0.2        evaluate_0.15
## [10] httr_1.4.3          pillar_1.7.0        rlang_1.0.3
## [13] lazyeval_0.2.2      googlesheets4_1.0.1 readxl_1.4.1
## [16] rstudioapi_0.14     data.table_1.14.2   rmarkdown_2.14
## [19] googledrive_2.0.0   htmlwidgets_1.5.4   munsell_0.5.0
## [22] broom_1.0.0         compiler_4.2.1      modelr_0.1.9
## [25] xfun_0.31           pkgconfig_2.0.3     htmltools_0.5.2
## [28] tidyselect_1.1.2    viridisLite_0.4.0   fansi_1.0.3
## [31] crayon_1.5.1        tzdb_0.3.0          dbplyr_2.2.1
## [34] withr_2.5.0         grid_4.2.1          jsonlite_1.8.0
## [37] gtable_0.3.0        lifecycle_1.0.1     DBI_1.1.3
## [40] magrittr_2.0.3      scales_1.2.0        cli_3.3.0
## [43] stringi_1.7.8       broom.helpers_1.8.0 fs_1.5.2
## [46] xml2_1.3.3          ellipsis_0.3.2      generics_0.1.3
## [49] vctrs_0.4.1         tools_4.2.1         glue_1.6.2
## [52] hms_1.1.1           fastmap_1.1.0       yaml_2.3.5
## [55] colorspace_2.0-3    gargle_1.2.0        rvest_1.0.3
## [58] knitr_1.39          haven_2.5.1
```

1 Univariate Analysis

In univariate analysis, we look at a single variable and describe 3 different things:

- Center
- Shape
- Spread

1.1 Center

Helps explain where the middle of the data is. This can be measured in 3 main ways.

1.1.1 Mean

```
grades <- c(78,79,80,81,82)

mean(grades)
```

```
## [1] 80
```

1.1.2 Median

```
median(grades)
```

```
## [1] 80
```

1.1.3 Mode

There isn't an easy way of doing this, so I created a function instead.

```
getModes <- function(x) {
  ux <- unique(x)
  tab <- tabulate(match(x, ux))
  ux[tab == max(tab)]
}
```

```
getModes(grades)
```

```
## [1] 78 79 80 81 82
```

Since there is no value that occurs most frequently, they all show. Mode is rarely used.

1.2 Shape

2 Optional - Fancier Output

Check out this fun stuff! Makes things look much cleaner.

2.1 Summary Statistics

Using our classic `mtcars` dataset.

```
mtcars%>% select(mpg, cyl,hp) %>%
  tbl_summary(statistic = list(all_continuous() ~ c("{mean} ({sd})",
                                                     "{median} ({p25}, {p75})",
                                                     "{min}, {max}"),
                               all_categorical() ~ "{n} / {N} ({p}%)"),
              type = all_continuous() ~ "continuous2"
  )
```

Characteristic	**N = 32**
mpg	
Mean (SD)	20.1 (6.0)
Median (IQR)	19.2 (15.4, 22.8)
Range	10.4, 33.9
cyl	
4	11 / 32 (34%)
6	7 / 32 (22%)
8	14 / 32 (44%)
hp	
Mean (SD)	147 (69)
Median (IQR)	123 (96, 180)
Range	52, 335

2.2 Histograms