Descriptive statistics

Summarized by SY Ohn

## References:

1.https://www.statsandr.com/blog/descriptive-statistics-in-r/#line-plot 2.https://www.statmethods.net/stats/descriptives.html#:~:text=R%20provides%20a%20wide%20range,with%20a%20specified%20summary%20statistic.&text=Possible%20functions%20used%20in%20sapply,median%2C%20range%2C%20and%20quantile.

# Descriptive statistics

1. Base package
2. Using Hmisc , pastecs, psych package

## Using Base package

Use sapply with summary statistics 1. numeric variables: mean, sd, var, min, max, range, and quatiles 2. nominal variables: contingency table, mode

We are using mtcars data set

library(dplyr)

## Warning: 패키지 'dplyr'는 R 버전 4.1.3에서 작성되었습니다

##   
## 다음의 패키지를 부착합니다: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data("mtcars")  
mt1 <- mtcars %>% as\_tibble # pipe  
# convert into tibble  
  
mt1$cyl <- factor(mt1$cyl)  
mt1$vs <- factor(mt1$vs)  
mt1$am <- factor(mt1$am)   
mt1$gear <- factor(mt1$gear)   
mt1$carb <- factor(mt1$carb)   
  
class(mt1)

## [1] "tbl\_df" "tbl" "data.frame"

anyNA(mt1)

## [1] FALSE

mt1

## # A tibble: 32 x 11  
## mpg cyl disp hp drat wt qsec vs am gear carb   
## <dbl> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct> <fct>  
## 1 21 6 160 110 3.9 2.62 16.5 0 1 4 4   
## 2 21 6 160 110 3.9 2.88 17.0 0 1 4 4   
## 3 22.8 4 108 93 3.85 2.32 18.6 1 1 4 1   
## 4 21.4 6 258 110 3.08 3.22 19.4 1 0 3 1   
## 5 18.7 8 360 175 3.15 3.44 17.0 0 0 3 2   
## 6 18.1 6 225 105 2.76 3.46 20.2 1 0 3 1   
## 7 14.3 8 360 245 3.21 3.57 15.8 0 0 3 4   
## 8 24.4 4 147. 62 3.69 3.19 20 1 0 4 2   
## 9 22.8 4 141. 95 3.92 3.15 22.9 1 0 4 2   
## 10 19.2 6 168. 123 3.92 3.44 18.3 1 0 4 4   
## # ... with 22 more rows

### Numeric variables

mt1\_num <- mt1[,sapply(mt1, is.numeric)]  
mt1\_num

## # A tibble: 32 x 6  
## mpg disp hp drat wt qsec  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 21 160 110 3.9 2.62 16.5  
## 2 21 160 110 3.9 2.88 17.0  
## 3 22.8 108 93 3.85 2.32 18.6  
## 4 21.4 258 110 3.08 3.22 19.4  
## 5 18.7 360 175 3.15 3.44 17.0  
## 6 18.1 225 105 2.76 3.46 20.2  
## 7 14.3 360 245 3.21 3.57 15.8  
## 8 24.4 147. 62 3.69 3.19 20   
## 9 22.8 141. 95 3.92 3.15 22.9  
## 10 19.2 168. 123 3.92 3.44 18.3  
## # ... with 22 more rows

sapply(mt1\_num, mean, na.rm = TRUE)

## mpg disp hp drat wt qsec   
## 20.090625 230.721875 146.687500 3.596563 3.217250 17.848750

# repeat for mean, sd, var, min, max, range, and quatiles for numeric variables

### Nominal variables

table, mode-최빈값

mt1\_nom <- mt1[,sapply(mt1, is.factor)]  
  
mt1\_table <- sapply(mt1\_nom, table) # contingency table  
mt1\_table

## $cyl  
##   
## 4 6 8   
## 11 7 14   
##   
## $vs  
##   
## 0 1   
## 18 14   
##   
## $am  
##   
## 0 1   
## 19 13   
##   
## $gear  
##   
## 3 4 5   
## 15 12 5   
##   
## $carb  
##   
## 1 2 3 4 6 8   
## 7 10 3 10 1 1

max\_row\_index <- sapply(mt1\_table, which.max)  
max\_row\_index

## cyl.8 vs.0 am.0 gear.3 carb.2   
## 3 1 1 1 2

# mode values  
mapply(function(x,y){return(rownames(x)[y])}, mt1\_table, max\_row\_index)

## cyl vs am gear carb   
## "8" "0" "0" "3" "2"

### summary()

quartiles: mean,median,25th and 75th quartiles,min,max

summary(mt1)

## mpg cyl disp hp drat   
## Min. :10.40 4:11 Min. : 71.1 Min. : 52.0 Min. :2.760   
## 1st Qu.:15.43 6: 7 1st Qu.:120.8 1st Qu.: 96.5 1st Qu.:3.080   
## Median :19.20 8:14 Median :196.3 Median :123.0 Median :3.695   
## Mean :20.09 Mean :230.7 Mean :146.7 Mean :3.597   
## 3rd Qu.:22.80 3rd Qu.:326.0 3rd Qu.:180.0 3rd Qu.:3.920   
## Max. :33.90 Max. :472.0 Max. :335.0 Max. :4.930   
## wt qsec vs am gear carb   
## Min. :1.513 Min. :14.50 0:18 0:19 3:15 1: 7   
## 1st Qu.:2.581 1st Qu.:16.89 1:14 1:13 4:12 2:10   
## Median :3.325 Median :17.71 5: 5 3: 3   
## Mean :3.217 Mean :17.85 4:10   
## 3rd Qu.:3.610 3rd Qu.:18.90 6: 1   
## Max. :5.424 Max. :22.90 8: 1

## Using the psych package

library(psych)

## Warning: 패키지 'psych'는 R 버전 4.1.3에서 작성되었습니다

describe(mtcars)

## vars n mean sd median trimmed mad min max range skew  
## mpg 1 32 20.09 6.03 19.20 19.70 5.41 10.40 33.90 23.50 0.61  
## cyl 2 32 6.19 1.79 6.00 6.23 2.97 4.00 8.00 4.00 -0.17  
## disp 3 32 230.72 123.94 196.30 222.52 140.48 71.10 472.00 400.90 0.38  
## hp 4 32 146.69 68.56 123.00 141.19 77.10 52.00 335.00 283.00 0.73  
## drat 5 32 3.60 0.53 3.70 3.58 0.70 2.76 4.93 2.17 0.27  
## wt 6 32 3.22 0.98 3.33 3.15 0.77 1.51 5.42 3.91 0.42  
## qsec 7 32 17.85 1.79 17.71 17.83 1.42 14.50 22.90 8.40 0.37  
## vs 8 32 0.44 0.50 0.00 0.42 0.00 0.00 1.00 1.00 0.24  
## am 9 32 0.41 0.50 0.00 0.38 0.00 0.00 1.00 1.00 0.36  
## gear 10 32 3.69 0.74 4.00 3.62 1.48 3.00 5.00 2.00 0.53  
## carb 11 32 2.81 1.62 2.00 2.65 1.48 1.00 8.00 7.00 1.05  
## kurtosis se  
## mpg -0.37 1.07  
## cyl -1.76 0.32  
## disp -1.21 21.91  
## hp -0.14 12.12  
## drat -0.71 0.09  
## wt -0.02 0.17  
## qsec 0.34 0.32  
## vs -2.00 0.09  
## am -1.92 0.09  
## gear -1.07 0.13  
## carb 1.26 0.29

## Using the Hmisc package

library(Hmisc)

## Warning: 패키지 'Hmisc'는 R 버전 4.1.3에서 작성되었습니다

## 필요한 패키지를 로딩중입니다: lattice

## 필요한 패키지를 로딩중입니다: survival

## 필요한 패키지를 로딩중입니다: Formula

## 필요한 패키지를 로딩중입니다: ggplot2

## Warning: 패키지 'ggplot2'는 R 버전 4.1.3에서 작성되었습니다

##   
## 다음의 패키지를 부착합니다: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

##   
## 다음의 패키지를 부착합니다: 'Hmisc'

## The following object is masked from 'package:psych':  
##   
## describe

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

describe(mt1)

## mt1   
##   
## 11 Variables 32 Observations  
## --------------------------------------------------------------------------------  
## mpg   
## n missing distinct Info Mean Gmd .05 .10   
## 32 0 25 0.999 20.09 6.796 12.00 14.34   
## .25 .50 .75 .90 .95   
## 15.43 19.20 22.80 30.09 31.30   
##   
## lowest : 10.4 13.3 14.3 14.7 15.0, highest: 26.0 27.3 30.4 32.4 33.9  
## --------------------------------------------------------------------------------  
## cyl   
## n missing distinct   
## 32 0 3   
##   
## Value 4 6 8  
## Frequency 11 7 14  
## Proportion 0.344 0.219 0.438  
## --------------------------------------------------------------------------------  
## disp   
## n missing distinct Info Mean Gmd .05 .10   
## 32 0 27 0.999 230.7 142.5 77.35 80.61   
## .25 .50 .75 .90 .95   
## 120.83 196.30 326.00 396.00 449.00   
##   
## lowest : 71.1 75.7 78.7 79.0 95.1, highest: 360.0 400.0 440.0 460.0 472.0  
## --------------------------------------------------------------------------------  
## hp   
## n missing distinct Info Mean Gmd .05 .10   
## 32 0 22 0.997 146.7 77.04 63.65 66.00   
## .25 .50 .75 .90 .95   
## 96.50 123.00 180.00 243.50 253.55   
##   
## lowest : 52 62 65 66 91, highest: 215 230 245 264 335  
## --------------------------------------------------------------------------------  
## drat   
## n missing distinct Info Mean Gmd .05 .10   
## 32 0 22 0.997 3.597 0.6099 2.853 3.007   
## .25 .50 .75 .90 .95   
## 3.080 3.695 3.920 4.209 4.314   
##   
## lowest : 2.76 2.93 3.00 3.07 3.08, highest: 4.08 4.11 4.22 4.43 4.93  
## --------------------------------------------------------------------------------  
## wt   
## n missing distinct Info Mean Gmd .05 .10   
## 32 0 29 0.999 3.217 1.089 1.736 1.956   
## .25 .50 .75 .90 .95   
## 2.581 3.325 3.610 4.048 5.293   
##   
## lowest : 1.513 1.615 1.835 1.935 2.140, highest: 3.845 4.070 5.250 5.345 5.424  
## --------------------------------------------------------------------------------  
## qsec   
## n missing distinct Info Mean Gmd .05 .10   
## 32 0 30 1 17.85 2.009 15.05 15.53   
## .25 .50 .75 .90 .95   
## 16.89 17.71 18.90 19.99 20.10   
##   
## lowest : 14.50 14.60 15.41 15.50 15.84, highest: 19.90 20.00 20.01 20.22 22.90  
## --------------------------------------------------------------------------------  
## vs   
## n missing distinct   
## 32 0 2   
##   
## Value 0 1  
## Frequency 18 14  
## Proportion 0.562 0.438  
## --------------------------------------------------------------------------------  
## am   
## n missing distinct   
## 32 0 2   
##   
## Value 0 1  
## Frequency 19 13  
## Proportion 0.594 0.406  
## --------------------------------------------------------------------------------  
## gear   
## n missing distinct   
## 32 0 3   
##   
## Value 3 4 5  
## Frequency 15 12 5  
## Proportion 0.469 0.375 0.156  
## --------------------------------------------------------------------------------  
## carb   
## n missing distinct   
## 32 0 6   
##   
## lowest : 1 2 3 4 6, highest: 2 3 4 6 8  
##   
## Value 1 2 3 4 6 8  
## Frequency 7 10 3 10 1 1  
## Proportion 0.219 0.312 0.094 0.312 0.031 0.031  
## --------------------------------------------------------------------------------

## Using the pastecs package

library(pastecs)

## Warning: 패키지 'pastecs'는 R 버전 4.1.3에서 작성되었습니다

##   
## 다음의 패키지를 부착합니다: 'pastecs'

## The following objects are masked from 'package:dplyr':  
##   
## first, last

stat.desc(mt1)

## mpg cyl disp hp drat wt  
## nbr.val 32.0000000 NA 3.200000e+01 32.0000000 32.00000000 32.0000000  
## nbr.null 0.0000000 NA 0.000000e+00 0.0000000 0.00000000 0.0000000  
## nbr.na 0.0000000 NA 0.000000e+00 0.0000000 0.00000000 0.0000000  
## min 10.4000000 NA 7.110000e+01 52.0000000 2.76000000 1.5130000  
## max 33.9000000 NA 4.720000e+02 335.0000000 4.93000000 5.4240000  
## range 23.5000000 NA 4.009000e+02 283.0000000 2.17000000 3.9110000  
## sum 642.9000000 NA 7.383100e+03 4694.0000000 115.09000000 102.9520000  
## median 19.2000000 NA 1.963000e+02 123.0000000 3.69500000 3.3250000  
## mean 20.0906250 NA 2.307219e+02 146.6875000 3.59656250 3.2172500  
## SE.mean 1.0654240 NA 2.190947e+01 12.1203173 0.09451874 0.1729685  
## CI.mean.0.95 2.1729465 NA 4.468466e+01 24.7195501 0.19277224 0.3527715  
## var 36.3241028 NA 1.536080e+04 4700.8669355 0.28588135 0.9573790  
## std.dev 6.0269481 NA 1.239387e+02 68.5628685 0.53467874 0.9784574  
## coef.var 0.2999881 NA 5.371779e-01 0.4674077 0.14866382 0.3041285  
## qsec vs am gear carb  
## nbr.val 32.0000000 NA NA NA NA  
## nbr.null 0.0000000 NA NA NA NA  
## nbr.na 0.0000000 NA NA NA NA  
## min 14.5000000 NA NA NA NA  
## max 22.9000000 NA NA NA NA  
## range 8.4000000 NA NA NA NA  
## sum 571.1600000 NA NA NA NA  
## median 17.7100000 NA NA NA NA  
## mean 17.8487500 NA NA NA NA  
## SE.mean 0.3158899 NA NA NA NA  
## CI.mean.0.95 0.6442617 NA NA NA NA  
## var 3.1931661 NA NA NA NA  
## std.dev 1.7869432 NA NA NA NA  
## coef.var 0.1001159 NA NA NA NA

## Summary Statistics by Group

A simple way of generating summary statistics by grouping variable is available in the psych package.

library(psych)  
describeBy(mtcars$mpg, group = mtcars$vs)

##   
## Descriptive statistics by group   
## group: 0  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 18 16.62 3.86 15.65 16.42 2.97 10.4 26 15.6 0.48 -0.05 0.91  
## ------------------------------------------------------------   
## group: 1  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 14 24.56 5.38 22.8 24.34 6 17.8 33.9 16.1 0.41 -1.4 1.44

### - End -