DSCI 610: Descriptive Statistics

Numerical Summaries: Descriptive Statistics

In addition to graphical summaries, you can present numerical summaries of your data in the form of descriptive statistics. The `kable()` function in `knitr` package can produce very nice tables ready for publication. The basic strategy is to construct a data frame, and then send the data frame to a package that constructs publication ready tables.

For most cases, two types of tables are generated. The first are descriptive tables that describe important features of the data being examined. The second are tables for presenting outputs from statistcial analysis.

We will demonstrate construction of descriptve tables here using the `NHANES` dataset. Statistics output tables will be deferred until later in the semester.

```
library(NHANES)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse 1.3
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------ tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
as_tibble(NHANES)
```

```
## # A tibble: 10,000 x 76
##       ID SurveyYr Gender   Age AgeDecade AgeMonths Race1 Race3 Education
##    <int> <fct>    <fct>  <int> <fct>         <int> <fct> <fct> <fct>
##  1 51624 2009_10  male      34 " 30-39"        409 White <NA>  High Sch~
##  2 51624 2009_10  male      34 " 30-39"        409 White <NA>  High Sch~
##  3 51624 2009_10  male      34 " 30-39"        409 White <NA>  High Sch~
##  4 51625 2009_10  male       4 " 0-9"           49 Other <NA>  <NA>
##  5 51630 2009_10  female    49 " 40-49"        596 White <NA>  Some Col~
##  6 51638 2009_10  male       9 " 0-9"          115 White <NA>  <NA>
##  7 51646 2009_10  male       8 " 0-9"          101 White <NA>  <NA>
##  8 51647 2009_10  female    45 " 40-49"        541 White <NA>  College ~
##  9 51647 2009_10  female    45 " 40-49"        541 White <NA>  College ~
```

```
## 10 51647 2009_10  female    45 " 40-49"        541 White <NA>  College ~
## # ... with 9,990 more rows, and 67 more variables: MaritalStatus <fct>,
## #   HHIncome <fct>, HHIncomeMid <int>, Poverty <dbl>, HomeRooms <int>,
## #   HomeOwn <fct>, Work <fct>, Weight <dbl>, Length <dbl>, HeadCirc <dbl>,
## #   Height <dbl>, BMI <dbl>, BMICatUnder20yrs <fct>, BMI_WHO <fct>,
## #   Pulse <int>, BPSysAve <int>, BPDiaAve <int>, BPSys1 <int>, BPDia1 <int>,
## #   BPSys2 <int>, BPDia2 <int>, BPSys3 <int>, BPDia3 <int>, Testosterone <dbl>,
## #   DirectChol <dbl>, TotChol <dbl>, UrineVol1 <int>, UrineFlow1 <dbl>,
## #   UrineVol2 <int>, UrineFlow2 <dbl>, Diabetes <fct>, DiabetesAge <int>,
## #   HealthGen <fct>, DaysPhysHlthBad <int>, DaysMentHlthBad <int>,
## #   LittleInterest <fct>, Depressed <fct>, nPregnancies <int>, nBabies <int>,
## #   Age1stBaby <int>, SleepHrsNight <int>, SleepTrouble <fct>,
## #   PhysActive <fct>, PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>,
## #   TVHrsDayChild <int>, CompHrsDayChild <int>, Alcohol12PlusYr <fct>,
## #   AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fct>, Smoke100 <fct>,
## #   Smoke100n <fct>, SmokeAge <int>, Marijuana <fct>, AgeFirstMarij <int>,
## #   RegularMarij <fct>, AgeRegMarij <int>, HardDrugs <fct>, SexEver <fct>,
## #   SexAge <int>, SexNumPartnLife <int>, SexNumPartYear <int>, SameSex <fct>,
## #   SexOrientation <fct>, PregnantNow <fct>
```

We will create an analysis dataset by selecting only the variables we are interesd in. Our exploratory data analysis will be based on the new dataset `df_eda`.

```
df_eda <- select(NHANES, ID, SurveyYr, Gender, Age, Race1, Poverty, HomeOwn, Weight, Height, BMI, BPSys
df_eda
```

```
## # A tibble: 10,000 x 15
##       ID SurveyYr Gender   Age Race1 Poverty HomeOwn Weight Height   BMI
##    <int> <fct>    <fct>  <int> <fct>   <dbl> <fct>    <dbl>  <dbl> <dbl>
##  1 51624 2009_10  male      34 White    1.36 Own       87.4   165.  32.2
##  2 51624 2009_10  male      34 White    1.36 Own       87.4   165.  32.2
##  3 51624 2009_10  male      34 White    1.36 Own       87.4   165.  32.2
##  4 51625 2009_10  male       4 Other    1.07 Own       17     105.  15.3
##  5 51630 2009_10  female    49 White    1.91 Rent      86.7   168.  30.6
##  6 51638 2009_10  male       9 White    1.84 Rent      29.8   133.  16.8
##  7 51646 2009_10  male       8 White    2.33 Own       35.2   131.  20.6
##  8 51647 2009_10  female    45 White    5    Own       75.7   167.  27.2
##  9 51647 2009_10  female    45 White    5    Own       75.7   167.  27.2
## 10 51647 2009_10  female    45 White    5    Own       75.7   167.  27.2
## # ... with 9,990 more rows, and 5 more variables: BPSysAve <int>,
## #   BPDiaAve <int>, TotChol <dbl>, Diabetes <fct>, SmokeNow <fct>
```

The following descriptive table has been constructed in two steps: i) construct a data frame `dftbl` with some manipulation on the contents of the final table, ii) send the data frame as an argument of the function `kable()` that will construct the publication ready table. For example, we present `mean`, `standard deviation` of total cholesterol level along with the number of participants `N` grouped by their race and gender. The first segment of the chunk calculates these statistics for each group. Note that the `pipe operator (%>%)` is used to link all the steps in the calculation without needing to save the intermediate objects. The second segment with `kable()` function actually creates the table.

```
# Step 1 : construct the data frame

dftbl <- df_eda %>%
```

```
  filter(!is.na(TotChol)) %>%
  group_by(Race1, Gender) %>%
  summarise(mean = mean(TotChol),
            stdev = sd(TotChol),
            N = n()) %>%
  ungroup() %>%

  pivot_wider(names_from = Gender, values_from = c(mean, stdev, N)) %>%
  select(Race1, ends_with("female"), everything())

# Step 2 : construct the table

  kable(dftbl, caption = "Descriptive Statistics", booktabs = TRUE,
               escape = F,
               digits = 3,
               longtable = T,
               col.names = c("Race", "Mean", "St. deviation", "N", "Mean", "St. deviation", "N")) %>%
    add_header_above(c(" " = 1, "Women" = 3, "Men" =3))%>%
  kable_styling(latex_options = "striped")
```

Table 1: Descriptive Statistics

	Women			Men		
Race	Mean	St. deviation	N	Mean	St. deviation	N
Black	4.806	0.988	501	4.597	0.999	459
Hispanic	4.816	1.006	269	4.777	1.015	231
Mexican	4.687	0.987	361	4.736	0.997	455
White	5.066	1.104	2802	4.857	1.091	2744
Other	4.732	0.989	323	4.658	1.056	329

In the following example, we present `median`, `75th percentile`, `25th percentile` of average systolic blood pressure along with the number of participants `N` grouped by their race and gender.

```
# Step 1 : construct the data frame
dftbl <- df_eda %>%
  filter(!is.na(BPSysAve)) %>%
  group_by(Race1, Gender) %>%
  summarise(median = median(BPSysAve),
            pct_75 = quantile(BPSysAve, prob=0.75),
            pct_25 = quantile(BPSysAve, prob=0.25),
            N = n()) %>%
  ungroup() %>%

  pivot_wider(names_from = Gender, values_from = c(median, pct_75, pct_25, N)) %>%
  select(Race1, ends_with("female"), everything())

# Step 2 : construct the table

  kable(dftbl, caption = "Descriptive Statistics", booktabs = TRUE,
               escape = F,
```

```
              digits = 3,
              longtable = T,
              col.names = c("Race", "Median", "75th Pct.", "25th Pct.", "N", "Median", "75th Pct.", "25
  add_header_above(c(" " = 1, "Women" = 4, "Men" = 4)) %>%
kable_styling(latex_options = "striped")
```

Table 2: Descriptive Statistics

	Women				Men			
Race	Median	75th Pct.	25th Pct.	N	Median	75th Pct.	25th Pct.	N
Black	115	129.00	106	514	119.0	132	111.00	477
Hispanic	108	121.00	100	265	116.0	126	108.25	234
Mexican	108	118.00	102	349	116.0	125	108.00	465
White	114	127.00	105	2847	119.0	129	110.00	2746
Other	107	122.25	101	324	116.5	126	108.00	330

References

1. Hadley Wickham and Garrett Grolemund. R for Data Sicence https://r4ds.had.co.nz/
2. Hao Zhu, Create Awesome LaTeX Table with knitr::kable and kableExtra