BIOSTATS 640 – Introduction to R
Fall 2023
https://people.umass.edu/biep640w/webpages/demonstrations.html

Source: *https://www.thisoldhouse.com/furniture/21017595/how-to-turn-a-wood-slab-into-a-table*

10
Beautiful Tables
November 10, 2023

Dataset used
hersdata.Rdata

Packages used: `Hmisc, tidyverse, gtsummary, gt, compareGroups`

1. At a Glance

{gtsummary} and {gt}

Descriptives -Every Variable	**No frills**
	`tbl_summary(dataframe)`
	
	Aesthetics added
	`tbl_summary(mydata,`
	`type = all_dichotomous() ~ "categorical",`
	`statistic=list(`
	`all_continuous() ~ " {mean} ({sd})")`
	`) %>%`
	`bold_labels( ) %>%`
	`as_gt() %>%`
	`tab_header(title="Table 1. Baseline Characteristics")`

Descriptives - by Group	**No frills**
	`tbl_summary(dataframe, by=groupvariable)`
	
	Aesthetics added
	`tbl_summary(mydata, by=HT,`
	`type = all_dichotomous() ~ "categorical",`
	`missing_text = "(Missing)",`
	`statistic=list(`
	`all_continuous() ~ " {mean} ({sd})")`
	`) %>%`
	`add_p(test = all_continuous() ~ "t.test") %>%`
	`bold_labels() %>%`
	`as_gt() %>%`
	`tab_header(title="Baseline Characteristics, by Group")`

Good to Know	**Code these as options of tbl_summary( ):**
	`( dataframe,`
	`by = survivef,`
	`type = list(age ~ 'continuous2'),`
	`statistic = list(`
	`all_continuous() ~ c("{N_nonmiss",`
	`"{mean} ({sd})",`
	`"({min}, {max})")),`
	`percent = c("row"),`
	`digits = all_continuous() ~ 2,`
	`label = c(age ~ "Age, years",`
	`BMI ~ "Body Mass Index"),`
	`missing_text = "(Missing)" ) %>%`
	
	Code these as their own lines of code:
	`add_p(test = all_continuous() ~ "t.test") %>%`
	`bold_labels() %>%`
	
	These lines of code come AFTER `as_gt( ) %>%`
	`as_gt() %>%`
	`tab_header(title="Baseline Characteristics, by Group")`

{compareGroups}

Preliminary: Label variables	library(Hmisc) **Example** label(mydata$HT) <- "Hormone Therapy" label(mydata$age) <- "Age,years"
Descrptives - every variable	library(compareGroups) **No frills** descrTable(dataframe) **Aesthetics added (e.g., changing the column heading from "all")** mytable <- descrTable(dataframe) print(mytable, header.labels = c("all" = "My preferred title"))
Descriptives - by Group Note. Requires 2 steps.	library(compareGroups) **Step 1: Compute statistics** mytable <- compareGroups(groupvar ~ var1 + var2 + etc, data = mydata, byrow=TRUE) # default is column % **Step 2: Produced basic display and show** createTable(mytable) **Aesthetics added (e.g., changing p.overall** print(createTable(mytable), header.labels = c("p.overall" = "p-value"))
Export	export2pdf(tab, file = "example.pdf") export2xls(tab, file = "example.xlsx") export2word(tab, file = "example.docx") export2latex(tab, file = "example.tex")

A nice example	**Produce crude OR's for a Y=0/1 (survival)in relationship to several predictors**
	mystats <- compareGroups(data=keepdata2,
	survivef ~ age + age_quintilef + femalef + classf,
	byrow = TRUE, # Show row %
	fact.ratio = c(age=10)) # Show OR for +10 years
	
	mytable2 <- createTable(mystats, show.ratio = TRUE,
	show.p.overall = FALSE,
	type=2) # show freqs and %
	
	print(mytable2, header.labels = c(p.ratio = "p-value")) # change labeling

2. Introduction to The Heart and Estrogen/progestin Replacement Study (HERS)

hersdata.Rdata

In this illustration, we are working with a larger subset of the HERS study.

Source: Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B and Vittinghoff E (1998). Randomized trial of estrogren plus progestin for secondary prevention of heart diessase in postmenopausal women. The Heart and Estrogen/progestin Replacement Study. *Journal of the American Medical Association*, **280**(7), 605-613.

In the HERS study, Hulley et al. (1998) sought to determine if exercise, a modifiable behavior, might lower the risk of diabetes in non-diabetic women who were at risk of developing the disease. The question is a complex one because there are many risk factors for diabetes. Moreover, the type of woman who chooses to exercise may be related in other ways to risk of diabetes, apart from the fact of her exercise habit. For example, women who exercise regularly are typically younger and have lower body mass index (BMI); these characteristics also confer a risk benefit with respect to diabetes. Finally, the benefit of exercise may be mediated through a reduction of body mass index. Vittinghoff, Glidden, Shiboski and McCullogh (2005) consider portions of this data in their 2005 text, *Regression Methods in Biostatistics: Linear.Logistic, Survival and Repeated Measures Models* (Springer).

The subset **hersdata.Rdata** has n=2,763 observations on 37 variables. We will be using the following 8 variables

Data Dictionary

Position	Variable	Variable Label	Type	Codes	Missing data
1	HT	Hormone Therapy	numeric	0 = placebo 1 = hormone therapy	None
2	age	Age in years	numeric	Range: [44, 79]	None
3	raceth	Race/ethnicity	character	"1" = White "2" = African American "3" = Other	None
4	exercise	Exercise at least 3x/week	numeric	0 = no 1 = yes	None
5	diabetes	Diabetes	numeric	0 = no 1 = yes	None
6	BMI	Body Mass Index, kg/m²	numeric	Range: [15.49, 49.51]	Yes
7	glucose	Glucose, mg/dl	numeric	Range: [67, 294]	None
8	LDL	LDL Cholesteraol, mg/dl	numeric	Range: [36.8, 365.2]	Yes

3. Highlights of Lesson 09
Introduction to Analysis of Variance

One Way Analysis of Variance

Create R factor variables	**Example:**
	`library(tidyverse)`
	`source$racethf <- factor(source$raceth,`
	`levels=c(1,2,3),`
	`labels=c("White", "African-American", "Other Race"))`

Create 0/1 indicator variables	**Example:**
	`source <- source %>%`
	`mutate(African_American = ifelse(racethf=="African-American",1,0)) %>%`
	`mutate(Other_Race = ifelse(racethf=="Other Race",1,0))`

Data Description	**Example:**
	`library(summarytools)`
	
	`with(source, # with(dataframe,`
	`stby(data = sbp, # data = yvar`
	`INDICES = racethf, # INDICES = factor var`
	`FUN = descr,`
	`stats=c("n.valid", "pct.valid", "mean", "sd", "min", "max"),`
	`transpose=TRUE))`

Fit Model	**Example:**
	`m1_anova <- aov(sbp ~ racethf, data=source) # racethf must be type factor`
	`m1_regression <- lm(sbp ~ as.factor(raceth), data=source) # equivalent`
	`m1_regression2 <- lm(sbp ~ African_American + Other_Race, data=source) # Using 0/1's`

Diagnostics	**Example:**
	`hist(fit.resid) # histogram residuals`
	`plot(fit, which=2) # qqplot`
	`shapiro.test(fit.resid) # test of null: normality`
	`bartlett.test(fit.resid) # test of null: constant variance`
	`HH:hovPlot(sbp~racethf) # Plot variance residuals`
	`pairwise.t.test(yvar,groupvar) # pairwise t-tests`
	`TukeyHSD(aov(yvar~groupvar) # Tukey pairwise t-tests`

Data Visualization*	**Side-by-side box plot w overlay scatter:**
	`ggplot(data=source) +`
	`aes(x=racethf) + # x = factor predictor`
	`aes(y=sbp) + # y = outcome`
	`geom_boxplot( ) +`
	`geom_jitter( )`

*See R Lesson 09 for additional visualizations.

Two Way Analysis of Variance

Create R factor variables	**Example:**
	`library(tidyverse)`
	`source$racethf <- factor(source$raceth,`
	`levels=c(1,2,3),`
	`labels=c("White", "African-American", "Other Race"))`

Create 0/1 indicator variables	**Example:**
	`source <- source %>%`
	`mutate(African_American = ifelse(racethf=="African-American",1,0)) %>%`
	`mutate(Other_Race = ifelse(racethf=="Other Race",1,0))`

Data Description	**Example:**
	`library(FSA)`
	`Summarize(sbp ~ racethf + activityf, digits=2, data=source)`

Fit Model	**Example - Fit 2 way fully factorial anova:**
	`m2_anova <- aov(sbp ~ racethf + activityf + racethf:activityf, data=source)`
	`m2_anova <- aov(sbp ~ racethf*activityf), data=source)  # equivalent`

Data Visualization*	**Side-by-side box plot w overlay scatter, stratified:**
	`ggplot(data=source) +`
	`aes(x=racethf) +  # x = factor predictor`
	`aes(y=sbp) +  # y = outcome`
	`aes(fill=activityf) +  # fill = stratification variable`
	`geom_boxplot( ) +`
	`geom_jitter( ) +`
	`facet_grid(.~activityf)  #  panels in 1 row`

4. Introduction to {gtsummary} and {gt}

Please note. Some of the tables produced in {gtsummary} did not appear in my knitted document and so are not shown here. But the code ran successfully.

initialize session

```
setwd("/cloud/project")                                    # Set working directory
#getwd()                                                  # Check working directory (remove hashtag to execute)
options(scipen=999)                                        # Turn off scientific notation
rm(list = ls())                                            # Clear the Decks
```

Load R Data

```
load(file="hersdata.Rdata")                                # load(file="NAME".Rdata) to load an R dataset
hersdata <- as.data.frame(hersdata)
str(hersdata)
```

```
## 'data.frame':    2763 obs. of  37 variables:
##  $ HT      : chr  "placebo" "placebo" "hormone therapy" "placebo" ...
##  $ age     : num  70 62 69 64 65 68 70 69 61 62 ...
##  $ raceth  : chr  "African American" "African American" "White" "White" ...
##  $ nonwhite: chr  "yes" "yes" "no" "no" ...
##  $ smoking : chr  "no" "no" "no" "yes" ...
.... several rows not shown ...
##   .. ..- attr(*, "class")= chr [1:2] "collector_guess" "collector"
##   ..$ skip    : num 1
##   ..- attr(*, "class")= chr "col_spec"
```

IMPORTANT - When creating a beautiful table, work with ONLY the variables you want

```
library(tidyverse)
mydata <- hersdata %>%                                     # mydata will have ONLY the vars of interest
          dplyr::select(HT,diabetes,age,BMI)               # if need be preface with dplyr::
head(mydata, n=10)
```

```
##                HT diabetes age   BMI
## 1         placebo       no  70 23.69
## 2         placebo       no  62 28.62
## 3  hormone therapy      yes  69 42.51
## 4         placebo       no  64 24.39
## 5         placebo       no  65 21.90
## 6  hormone therapy       no  68 29.05
## 7         placebo      yes  70 34.45
## 8  hormone therapy       no  69 23.16
## 9  hormone therapy       no  61 30.26
## 10 hormone therapy       no  62 45.68
```

```
summary(mydata)
```

```
##       HT             diabetes             age             BMI
##  Length:2763        Length:2763        Min.   :44.00   Min.   :15.21
##  Class :character   Class :character   1st Qu.:62.00   1st Qu.:24.64
##  Mode  :character   Mode  :character   Median :67.00   Median :27.75
##                                        Mean   :66.65   Mean   :28.58
##                                        3rd Qu.:72.00   3rd Qu.:31.73
##                                        Max.   :79.00   Max.   :54.13
##                                                        NA's   :5
```

PRELIMINARY: Convert character to factor with levels in desired order. Label variables

```
library(Hmisc)                                                  # label() in package {Hmisc}

cat("\nShow character levels of original var HT")
table(mydata$HT)                                                # Quick look before creating factor vars

## Show character levels of original var HT
## hormone therapy         placebo
##            1380            1383

mydata$HT <- factor(mydata$HT,                                  # original var HT is character
                    levels=c("placebo","hormone therapy"),      # original levels
                    labels=c("Placebo", "Hormone Therapy"))     # new level labels (for display)

mydata$diabetes <- factor(mydata$diabetes,                      # original var diabetes is character
                   levels=c("no", "yes"),                       # original
                   labels=c("No","Yes"))                        # new level labels (for display)


label(mydata$HT) <- "Hormone Therapy"
label(mydata$age) <- "Age,years"
label(mydata$diabetes) <- "Diabetes"
label(mydata$BMI) <- "Body Mass Index"

summary(mydata)

##                HT       diabetes        age              BMI
##  Placebo        :1383   No :2032   Min.   :44.00   Min.   :15.21
##  Hormone Therapy:1380   Yes: 731   1st Qu.:62.00   1st Qu.:24.64
##                                    Median :67.00   Median :27.75
##                                    Mean   :66.65   Mean   :28.58
##                                    3rd Qu.:72.00   3rd Qu.:31.73
##                                    Max.   :79.00   Max.   :54.13
##                                                    NA's   :5
```

4.1. Descriptives on Every Variable

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

# Recommended:  Start with basic table.  Add aesthetics one at a time.
tbl_summary(mydata)
```

Characteristic	N = 2,763[1]
Hormone Therapy	
Placebo	1,383 (50%)
Hormone Therapy	1,380 (50%)
Diabetes	731 (26%)
Age,years	67 (62, 72)
Body Mass Index	27.8 (24.6, 31.7)
Unknown	5

[1] n (%); Median (IQR)

```
# AESTHETIC:  Force display of ALL levels of categorical vars that are dichotomous
# GOOD TO KNOW:  If categorical is 0/1 or 1/2 you may not get all levels displayed

tbl_summary(mydata,
          type = all_dichotomous() ~ "categorical")
```

Characteristic	N = 2,763[1]
Hormone Therapy	
Placebo	1,383 (50%)
Hormone Therapy	1,380 (50%)
Diabetes	
No	2,032 (74%)
Yes	731 (26%)
Age,years	67 (62, 72)
Body Mass Index	27.8 (24.6, 31.7)
Unknown	5

[1]n (%); Median (IQR)

```
# AESTHETIC: Show mean and sd instead of median and IQR.
# AESTHETIC:  Bold the variable names
tbl_summary(mydata,
           type = all_dichotomous() ~ "categorical",
           statistic=list(
               all_continuous() ~ " {mean} ({sd})")
                      ) %>%
           bold_labels( )
```

Characteristic	N = 2,763[1]
Hormone Therapy	
Placebo	1,383 (50%)
Hormone Therapy	1,380 (50%)
Diabetes	
No	2,032 (74%)
Yes	731 (26%)
Age,years	67 (7)
Body Mass Index	28.6 (5.5)
Unknown	5

[1]n (%); Mean (SD)

```
# AESTHETIC:  Add a title

tbl_summary(mydata,
           type = all_dichotomous() ~ "categorical",
           statistic=list(
                all_continuous() ~ " {mean} ({sd})")
                      ) %>%
           bold_labels( ) %>%
```

```
          as_gt() %>%
          tab_header(title="Table 1.  Baseline Characteristics")

# Tip. You can save your table as an object
mytable <- tbl_summary(mydata,
              type = all_dichotomous() ~ "categorical")
mytable
```

Characteristic	N = 2,763[1]
Hormone Therapy	
Placebo	1,383 (50%)
Hormone Therapy	1,380 (50%)
Diabetes	
No	2,032 (74%)
Yes	731 (26%)
Age,years	67 (62, 72)
Body Mass Index	27.8 (24.6, 31.7)
Unknown	5

[1]n (%); Median (IQR)

```
# TIP:  By enclosing code in parentheses, you save table as object and display at the same time.
(mytable <- tbl_summary(mydata,
              type = all_dichotomous() ~ "categorical"))
```

Characteristic	N = 2,763[1]
Hormone Therapy	
Placebo	1,383 (50%)
Hormone Therapy	1,380 (50%)
Diabetes	
No	2,032 (74%)
Yes	731 (26%)
Age,years	67 (62, 72)
Body Mass Index	27.8 (24.6, 31.7)
Unknown	5

[1]n (%); Median (IQR)

4.2. Descriptives by Group

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

# Recommended:  Start with basic table.  Add aesthetics one at a time.
tbl_summary(mydata, by=HT)               # by = FACTORVAR
```

Characteristic	Placebo, N = 1,383[1]	Hormone Therapy, N = 1,380[1]
Diabetes	352 (25%)	379 (27%)
Age,years	67 (62, 72)	67 (62, 72)
Body Mass Index	27.6 (24.5, 31.7)	27.9 (24.8, 31.8)
Unknown	4	1

[1]n (%); Median (IQR)

```
# AESTHETIC:  Force display of ALL levels of categorical variables
tbl_summary(mydata, by=HT,
            type = all_dichotomous() ~ "categorical")
```

Characteristic	Placebo, N = 1,383[1]	Hormone Therapy, N = 1,380[1]
Diabetes		
No	1,031 (75%)	1,001 (73%)
Yes	352 (25%)	379 (27%)
Age,years	67 (62, 72)	67 (62, 72)
Body Mass Index	27.6 (24.5, 31.7)	27.9 (24.8, 31.8)
Unknown	4	1

[1]n (%); Median (IQR)

```
AESTHETIC: Show mean and sd instead of median and IQR.
# AESTHETIC: Bold variable names
tbl_summary(mydata, by=HT,
       type = all_dichotomous() ~ "categorical",
       statistic=list(
            all_continuous() ~ " {mean} ({sd})")
                    ) %>%
       bold_labels( )                                          # this is using {gt}
```

Characteristic	Placebo, N = 1,383[1]	Hormone Therapy, N = 1,380[1]
Diabetes		
No	1,031 (75%)	1,001 (73%)
Yes	352 (25%)	379 (27%)
Age,years	67 (7)	67 (7)
Body Mass Index	28.5 (5.5)	28.6 (5.5)
Unknown	4	1

[1]n (%); Mean (SD)

```
# AESTHETIC:  Add a title
tbl_summary(mydata, by=HT,
      type = all_dichotomous() ~ "categorical",
      statistic=list(
           all_continuous() ~ " {mean} ({sd})")
                  ) %>%
           bold_labels( ) %>%
           as_gt() %>%
           tab_header(title="Baseline Characteristics, by Group")                              # this is in {gt}

# AESTHETIC:  Change "Unknown" to read "Missing"
tbl_summary(mydata, by=HT,
     type = all_dichotomous() ~ "categorical",
     missing_text = "(Missing)",
     statistic=list(
           all_continuous() ~ " {mean} ({sd})")
                  ) %>%
     bold_labels( ) %>%
     as_gt() %>%
     tab_header(title="Baseline Characteristics, by Group")
```

5. Introduction to {compareGroups}

5.1. Descriptives on Every Variable

```
library(compareGroups)
library(Hmisc)

# Recommended:  Start with basic table.  Add aesthetics one at a time.
descrTable(mydata)

##
## --------Summary descriptives table ---------
##
## _____________________________________
##                        [ALL]      N
##                        N=2763
## -------------------------------------
## Hormone Therapy:                 2763
##     Placebo          1383 (50.1%)
##     Hormone Therapy  1380 (49.9%)
## Diabetes:                        2763
##     No               2032 (73.5%)
##     Yes              731 (26.5%)
## Age,years            66.6 (6.65)  2763
## Body Mass Index      28.6 (5.52)  2758
## -------------------------------------

# AESTHETIC: Change header from "all" to "HERS Study Cohort"
# NOTE:  This requires saving table as an object first
mytable2 <-  descrTable(mydata)
print(mytable2, header.labels = c("all" = "HERS Study Cohort"))

##
## --------Summary descriptives table ---------
##
## _____________________________________________
##                     HERS Study Cohort  N
##                          N=2763
## ---------------------------------------------
## Hormone Therapy:                        2763
##     Placebo            1383 (50.1%)
##     Hormone Therapy    1380 (49.9%)
## Diabetes:                               2763
##     No                 2032 (73.5%)
##     Yes                 731 (26.5%)
## Age,years               66.6 (6.65)    2763
## Body Mass Index         28.6 (5.52)    2758
## ---------------------------------------------

# You can export your table to any of the following:
# export2pdf(tab, file = "example.pdf")
# export2xls(tab, file = "example.xlsx")
# export2word(tab, file = "example.docx")
# export2latex(tab, file = "example.tex")

# Export to a Word Document
# export2word(mytable2, file = "mytable2.docx")
```

5.2. Descriptives by Group

```
library(compareGroups)
library(Hmisc)

# Recommended:  Start with basic table.  Add aesthetics one at a time.
# NOTE:  Descriptives by group requires 2 steps
# STEP 1:  compareGroups() to generate statistics
# Step 2:  createTable() to create table
mytable3 <- compareGroups(HT ~ age + BMI + diabetes,         # group ~ var1 + var2 + etc
        data = mydata,
        byrow=TRUE)
createTable(mytable3)

##
## --------Summary descriptives table by 'Hormone Therapy'---------
##
## _____________________________________________________
##                   Placebo    Hormone Therapy p.overall
##                    N=1383        N=1380
## -----------------------------------------------------
## Age,years       66.8 (6.68)    66.5 (6.62)     0.331
## Body Mass Index 28.5 (5.52)    28.6 (5.52)     0.596
## Diabetes:                                      0.248
##     No          1031 (50.7%)  1001 (49.3%)
##     Yes         352 (48.2%)    379 (51.8%)
## -----------------------------------------------------

# Change "p.overall" to "p-value" in printout
print(createTable(mytable3),header.labels = c(p.overall = "p-value"))

##
## --------Summary descriptives table by 'Hormone Therapy'---------
##
## _____________________________________________________
##                   Placebo    Hormone Therapy p-value
##                    N=1383        N=1380
## -----------------------------------------------------
## Age,years       66.8 (6.68)    66.5 (6.62)    0.331
## Body Mass Index 28.5 (5.52)    28.6 (5.52)    0.596
## Diabetes:                                     0.248
##     No          1031 (50.7%)  1001 (49.3%)
##     Yes         352 (48.2%)    379 (51.8%)
## -----------------------------------------------------
```