





Comparing Groups

Week 10

Packages needed and a Note about Icons

Please load up the following packages. Remember to first install the ones you don't have.

You may come across the following icons. The table below lists what each means.

Icon	Description
	Indicates that an example continues on the following slide.
	Indicates that a section using common syntax has ended.
	Indicates that there is an active hyperlink on the slide.
	Indicates that a section covering a concept has ended.

The compareGroups package

- Originally designed to read, interpret, summarize, display and analyze epidemiological data.
- Allows you to create everything from data summaries for quality control.

Starting up

Let's use one of the preloaded data sets: `PREDIMED`.

- longitudinal study containing several baseline characteristics of the participants as well as events occurred during the 7 years follow-up period given by variables `event` and `toevent`.
- Each individual has been assigned to a three intervention diet randomly given by the variable `group`.
- You can read the study via [PubMed](#)

Run the following

```
data("predimed")
```

View the Data

We can take a look at the data by

```
predimed %>%  
  head()
```

```
##           group    sex age  smoke  bmi waist      wth htn diab hyperchol  
## 1      Control  Male  58  Former 33.53  122 0.7530864 No  No      Yes  
## 2      Control  Male  77  Current 31.05  119 0.7300614 Yes Yes     No  
## 4  MedDiet + V00 Female 72  Former 30.86  106 0.6543210 No  Yes     No  
## 5  MedDiet + Nuts  Male 71  Former 27.68  118 0.6941177 Yes No      Yes  
## 6  MedDiet + V00 Female 79  Never 35.94  129 0.8062500 Yes No      Yes  
## 8      Control  Male  63  Former 41.66  143 0.8033708 Yes Yes     Yes  
##   famhist hormo p14  toevent event  
## 1      No    No  10 5.374401  Yes  
## 2      No    No  10 6.097194   No  
## 4     Yes    No   8 5.946612   No  
## 5      No    No   8 2.907598  Yes  
## 6      No    No   9 4.761123   No  
## 8      No  <NA>   9 3.148528  Yes
```

Variable Names

You can take a look at the variables in the data set by running

```
names(predimed)
```

```
## [1] "group"      "sex"        "age"        "smoke"      "bmi"        "waist"  
## [7] "wth"        "htn"        "diab"       "hyperchol" "famhist"    "hormo"  
## [13] "p14"       "toevent"   "event"
```

Well that's not overtly helpful. Oh wait there's a codebook!

```
predimed_vars <-  
  read_csv("predimed_codebook.csv")
```

Code Book

Ok so let's take a look!

```
predimed_vars
```

```
## # A tibble: 15 × 3
##   Name      Label      Codes
##   <chr>    <chr>    <chr>
## 1 group    Intervention group Control; MedDiet + Nuts; MedDiet + V00
## 2 sex      Sex      Male; Female
## 3 age      Age      <NA>
## 4 smoke    Smoking  Never; Current; Former
## 5 bmi      Body mass index <NA>
## 6 waist    Waist circumference <NA>
## 7 wth      Waist-to-height ratio <NA>
## 8 htn      Hypertension No; Yes
## 9 diab     Type-2 diabetes No; Yes
## 10 hyperchol Dyslipidemia No; Yes
## 11 famhist Family history of premature CHD No; Yes
## 12 hormo    Hormone-replacement therapy No; Yes
## 13 p14      MeDiet Adherence score <NA>
## 14 toevent follow-up to main event (years) <NA>
## 15 event    AMI, stroke, or CV Death No; Yes
```

Descriptive Tables for Observations

If you want to create a quick table full of descriptives that *aren't meant for exporting*, use the `descrTable()` command

```
descrTable(group ~ ., predimed)
```

```
##
## -----Summary descriptives table by 'Intervention group'-----
##
## -----
##                               Control      MedDiet + Nuts  MedDiet + V00  p.overall
##                               N=2042      N=2100         N=2182
## -----
## Sex:                               <0.001
##   Male      812 (39.8%)    968 (46.1%)    899 (41.2%)
##   Female    1230 (60.2%)   1132 (53.9%)  1283 (58.8%)
## Age      67.3 (6.28)    66.7 (6.02)    67.0 (6.21)    0.003
## Smoking:                               0.444
##   Never    1282 (62.8%)   1259 (60.0%)  1351 (61.9%)
##   Current   270 (13.2%)    296 (14.1%)    292 (13.4%)
##   Former    490 (24.0%)    545 (26.0%)    539 (24.7%)
## Body mass index  30.3 (3.96)    29.7 (3.77)    29.9 (3.71)    <0.001
## Waist circumference  101 (10.8)    100 (10.6)    100 (10.4)    0.045
## Waist-to-height ratio  0.63 (0.07)    0.62 (0.06)    0.63 (0.06)    <0.001
## Hypertension:                               0.249
##   No      331 (16.2%)    362 (17.2%)    396 (18.1%)
```


Descriptive Tables for Analysis

If you want to create a table full of descriptives that *you can use for analysis*, use the `compareGroups()` command

```
comparison <-  
  compareGroups(group ~ ., predimed); comparison
```

```
##  
##  
## ----- Summary of results by groups of 'Intervention group'-----  
##  
##  
##      var                N      p.value  method          selection  
## 1  Sex                 6324 <0.001** categorical      ALL  
## 2  Age                 6324 0.003**  continuous normal ALL  
## 3  Smoking            6324 0.444    categorical      ALL  
## 4  Body mass index    6324 <0.001** continuous normal ALL  
## 5  Waist circumference 6324 0.045** continuous normal ALL  
## 6  Waist-to-height ratio 6324 <0.001** continuous normal ALL  
## 7  Hypertension       6324 0.249    categorical      ALL  
## 8  Type-2 diabetes    6324 0.017** categorical      ALL  
## 9  Dyslipidemia      6324 0.423    categorical      ALL  
## 10 Family history of premature CHD 6324 0.581    categorical      ALL  
## 11 Hormone-replacement therapy 5661 0.850    categorical      ALL  
## 12 MeDiet Adherence score 6324 <0.001** continuous normal ALL  
## 13 follow-up to main event (years) 6324 <0.001** continuous normal ALL
```

Subsetting

The previous example gave us the gambit. In `compareGroups(group ~ ., predimed)`, all of the variables were compared to each other. What if we just want to look at a few variables?

In this first example, we'll look at the impact of age, smoking, waist size, and hypercholesterol together on the group

```
compareGroups(group ~ age + smoke + waist + hyperchol,  
              data = predimed)
```

```
##  
##  
## ----- Summary of results by groups of 'Intervention group'-----  
##  
##  
##   var                N    p.value method      selection  
## 1 Age                6324 0.003** continuous normal ALL  
## 2 Smoking            6324 0.444   categorical  ALL  
## 3 Waist circumference 6324 0.045** continuous normal ALL  
## 4 Dyslipidemia       6324 0.423   categorical  ALL  
## -----  
## Signif. codes:  0 '**' 0.05 '*' 0.1 ' ' 1
```

Notice by using the p -value from the column `p.value`, we have our first indicator that something happened. It is NOT a guarantee!

A Quick Note About the p -value

You may have read articles where the outcomes of a study are labeled as a fact because the results were *statistically significant*.

- **Why isn't this right?** Historically and even to this day, p -values are commonly used to test and dismiss H_0 , which generally states that there is no
 - difference between two groups, or
 - correlation between a pair of characteristics.
-

Traditionally, the mistake has been in the interpretation and reliance on the notion that

the smaller the p -value, the less likely an observed set of values would occur by chance.

So $p \leq 0.05$ is generally taken to mean that a finding is statistically significant and therefore warrants publication which the American Statistical Association and anyone who knows better than to rely on a single measure can tell you is nonsense (what is called dumpster or garbage stats).

Ok That Wasn't a Quick A Note About the p -value

At best the p -value is what we call an *indicator* of something happening. Essentially it is one piece of evidence of many!

- **What it doesn't mean?** Firstly $p \leq 0.05$ *does not imply that there is a 95% chance that H_0 is correct.*
- **What it does mean!** It signifies that if the H_0 is true and all other assumptions made are valid, then there is a 5% chance of obtaining a result at least as extreme as the one observed.
 - **Most important?** A p -value cannot indicate the importance of a finding
 - *Example:* a medication can have a statistically significant effect on patients' blood glucose levels without having a therapeutic effect.
- **Time to get rid of it?** Well no. It is an indicator but just because its not the end all be all measure doesn't mean it's not useful. So use it *but* also use others with it!
 - *Examples:* There are many but confidence intervals are another piece of information to use. Other approaches include Bayesian methods and effect sizes.

The actual last slide strictly about p -values

Here is a good summary...well a summary at least:

p -values do NOT

- indicate *reproducibility* or *evidence*
- *prove* or *disprove* a hypothesis
- tell you to *accept a hypotheses*

p -values do

- indicate that *something is happening*
- imply a *probability exists*
- get misinterpreted a lot (and I mean a lot!) yielding *Type I* and *Type II Errors*

Back to Subsetting

Now that we hopefully have an idea what the p -value implies, let's look at the impact of age, smoking, waist size, and hypercholesterol together on the the sample of females

```
compareGroups(group ~ age + smoke + waist + hyperchol,  
              data = predimed,  
              subset = sex == "Female")
```

```
##  
##  
## ----- Summary of results by groups of 'group'-----  
##  
##  
##   var                N    p.value method      selection  
## 1 Age                 3645 0.056*  continuous normal sex == "Female"  
## 2 Smoking             3645 0.907   categorical      sex == "Female"  
## 3 Waist circumference 3645 0.016** continuous normal sex == "Female"  
## 4 Dyslipidemia        3645 0.319   categorical      sex == "Female"  
## -----  
## Signif. codes:  0 '**' 0.05 '*' 0.1 ' ' 1
```

It seems that *Age* and *Waist circumference* may impact the *Female* population in the study (i.e. sample). We'd have to investigate *all* of the variables more to know for sure.

Getting all of the p -values

If we wanted to get an idea if the variables impact each other, we can use

```
pvals <- getResults(comparison,  
                    "p.overall"); pvals
```

```
##           Sex           Age  
##      8.138384e-05      2.665539e-03  
##           Smoking      Body mass index  
##      4.443536e-01      3.405257e-06  
##      Waist circumference      Waist-to-height ratio  
##      4.464591e-02      7.388314e-05  
##      Hypertension      Type-2 diabetes  
##      2.487579e-01      1.725231e-02  
##      Dyslipidemia      Family history of premature CHD  
##      4.229670e-01      5.813070e-01  
##      Hormone-replacement therapy      MeDiet Adherence score  
##      8.500945e-01      1.249646e-10  
## follow-up to main event (years)      AMI, stroke, or CV Death  
##      2.076029e-25      6.386460e-02
```

Remember this is considering all of the variables, not those we subsetted!

APA Tables ...

We can also create an APA 7th edition formatted table!

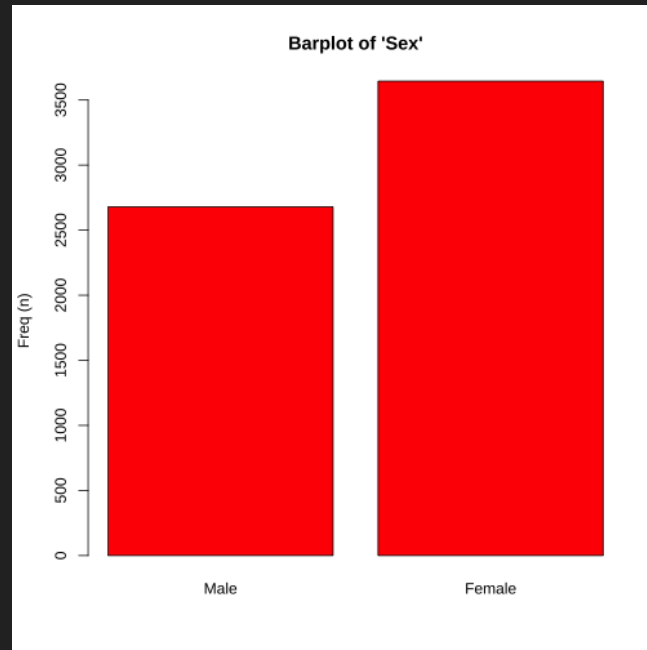
```
export_comparision <-  
  createTable(comparision); export_comparision
```

```
##  
## -----Summary descriptives table by 'Intervention group'-----  
##  
## -----  
##                               Control      MedDiet + Nuts  MedDiet + V00  p.overall  
##                               N=2042      N=2100          N=2182  
## -----  
## Sex:                               <0.001  
##   Male      812 (39.8%)    968 (46.1%)    899 (41.2%)  
##   Female    1230 (60.2%)   1132 (53.9%)  1283 (58.8%)  
## Age      67.3 (6.28)    66.7 (6.02)    67.0 (6.21)    0.003  
## Smoking:                               0.444  
##   Never    1282 (62.8%)   1259 (60.0%)  1351 (61.9%)  
##   Current  270 (13.2%)    296 (14.1%)   292 (13.4%)  
##   Former   490 (24.0%)    545 (26.0%)   539 (24.7%)  
## Body mass index  30.3 (3.96)   29.7 (3.77)   29.9 (3.71)   <0.001  
## Waist circumference  101 (10.8)   100 (10.6)    100 (10.4)    0.045  
## Waist-to-height ratio  0.63 (0.07)  0.62 (0.06)   0.63 (0.06)   <0.001  
## Hypertension:                               0.249
```


... and Plot ...

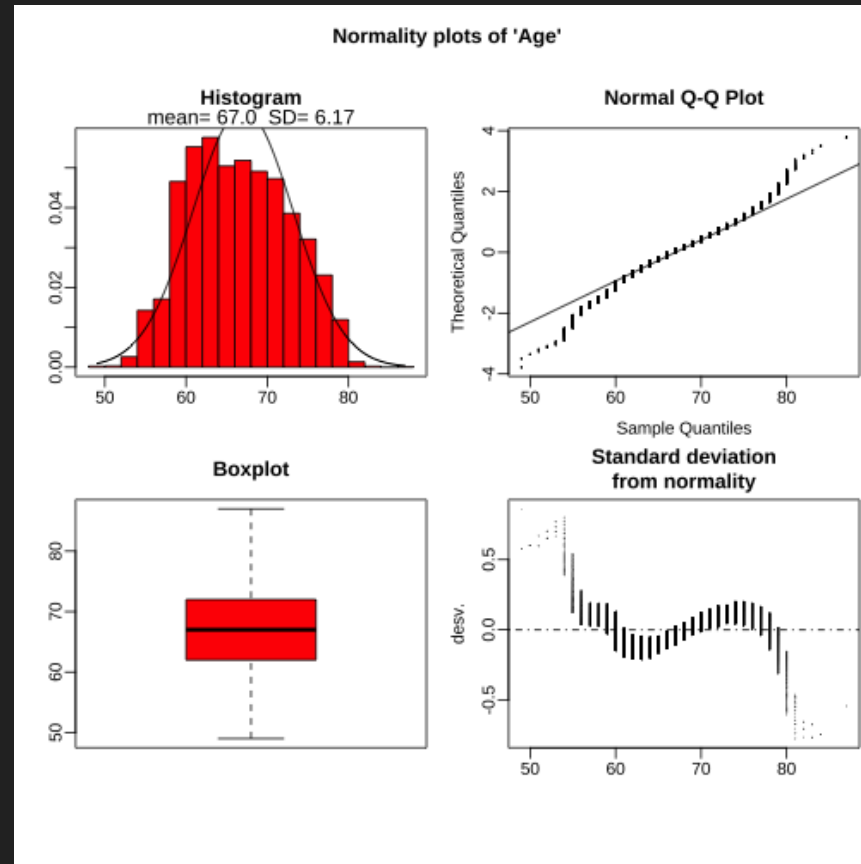
And we can also create an APA 7th edition formatted plot!

```
plot(export_comparison["sex"]) # barplot by sex
```



... and Other Plots

```
plot(export_comparison["age"]) # histogram and normality plot by age
```



Exporting

Finally you can export your items! Here are some common ways to export tables

```
export2csv(export_comparison,  
           file = "comparison.csv") # as a csv file  
  
export2word(export_comparison,  
            file = "comparison.docx") # as a word file  
  
export2xls(export_comparison,  
           file = "comparison.xls") # as a word file  
  
export2pdf(export_comparison,  
           file = "comparison.pdf") # as a pdf file
```

One More Thing: The GUI

If you do not like the command line interface of R or in general, there is an experimental click-click based built in app you can by typing

```
cGroupsGUI(predimed)
```

It appears to work fine on a PC. However if you have a Mac and *did not* install XQuartz as originally instructed, there is a *statistically significant* chance it may (a) not load or (b) have quirks if it does.

Thats it!