IPS9 in R: One-way analysis of variance (Chapter 12)

Shukry Zablah (szablah20@amherst.edu) and Nicholas Horton (nhorton@amherst.edu)

July 25, 2018

Introduction and background

These documents are intended to help describe how to undertake analyses introduced as examples in the Ninth Edition of *Introduction to the Practice of Statistics* (2017) by Moore, McCabe, and Craig.

More information about the book can be found here. The data used in these documents can be found under Data Sets in the Student Site. This file as well as the associated R Markdown reproducible analysis source file used to create it can be found at https://nhorton.people.amherst.edu/ips9/.

This work leverages initiatives undertaken by Project MOSAIC (http://www.mosaic-web.org), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the `mosaic` package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignettes (http://cran.r-project.org/web/packages/mosaic). A paper describing the mosaic approach was published in the *R Journal*: https://journal.r-project.org/archive/2017/RJ-2017-024.

Chapter 12: One-way analysis of variance

This file replicates the analyses from Chapter 12: One-way analysis of variance.

First, load the packages that will be needed for this document:

```r
library(mosaic)
library(readr)
library(DescTools)
```

> Complicated computations do not guarantee a valid statistical analysis. (648)

Section 12.1: Inference for one-way analysis of variance

We begin with Example 12.3 in page 648. Let's read in our data

```r
#Ex12.3
Friends <- read_csv("https://nhorton.people.amherst.edu/ips9/data/chapter12/EG12-03FRIENDS.csv")
head(Friends)
```

```
## # A tibble: 6 x 3
##   Friends Participant Score
##     <int>       <int> <dbl>
## 1     102           1   3.8
## 2     102           2   3.6
## 3     102           3   3.2
## 4     102           4   2.4
## 5     102           5   4.8
## 6     102           6   3
```

We want to get a nice summary table like the one in example 12.3. To do this we use the `favstats()` function.

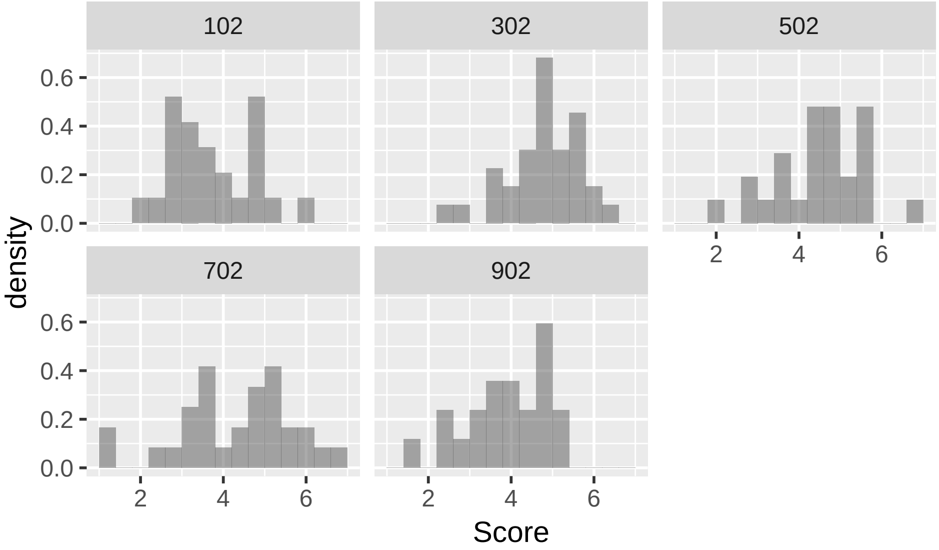
```
#Ex12.3
SummaryFriends <- favstats(Score ~ Friends, data = Friends)
SummaryFriends
```

```
##   Friends min   Q1 median  Q3 max     mean        sd  n missing
## 1     102 2.2 3.00    3.6 4.8 6.0 3.816667 0.9989850 24       0
## 2     302 2.6 4.40    5.0 5.6 6.4 4.878788 0.8513803 33       0
## 3     502 2.0 3.85    4.7 5.4 6.8 4.561538 1.0703558 26       0
## 4     702 1.0 3.60    4.6 5.4 7.0 4.406667 1.4282696 30       0
## 5     902 1.6 3.40    4.2 5.0 5.2 3.990476 1.0226949 21       0
```

There are some helpful visualizations starting in page 649 that will help us in our analysis of our Friends data.

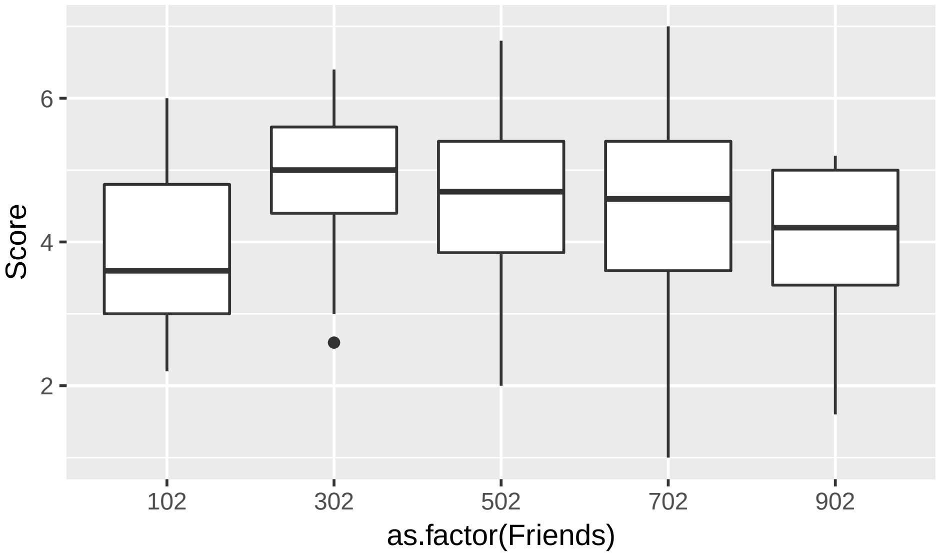
The `gf_histogram()` function can be used to recreate the histogram (note the `|` used to facet based on the number of friends).

```
#Fig12.4
gf_dhistogram(~ Score | Friends, data = Friends, binwidth = 0.4)
```



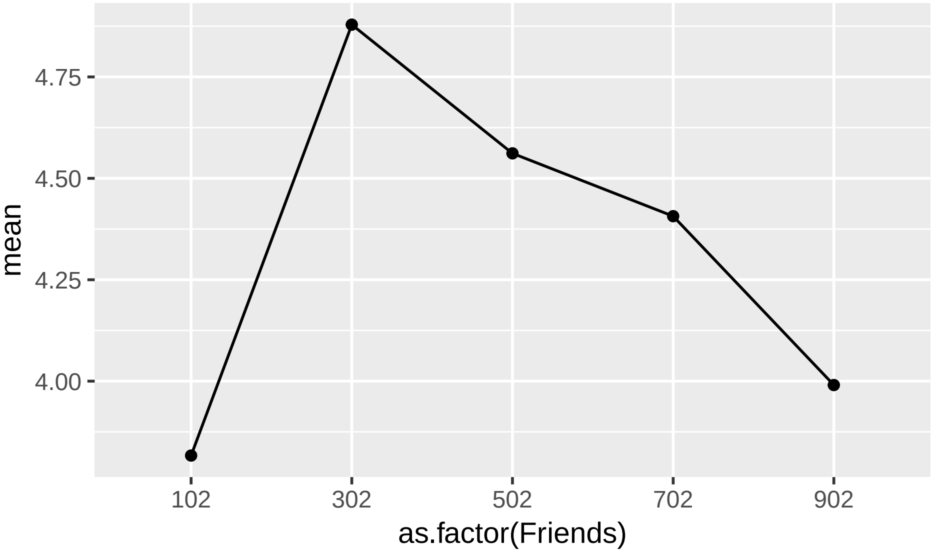
To recreate the boxplot (note: we have to convert Friends to a factor for the boxplots to work):

```
#Fig12.4
gf_boxplot(Score ~ as.factor(Friends), data = Friends)
```



And to recreate the linegraph (note the piping through gf_point to add the points to the line):

```
#Fig12.4
gf_line(mean ~ as.factor(Friends), data = SummaryFriends, group = 1) %>% gf_point()
```



These visualizations are a quick way to know what is going on with your data. The `favstats()` command can be used to generate summary statistics by group.

```
#Ex12.7
favstats(Score ~ Friends, data = Friends)
```

```
##   Friends min   Q1 median  Q3 max     mean        sd  n missing
## 1     102 2.2 3.00    3.6 4.8 6.0 3.816667 0.9989850 24       0
## 2     302 2.6 4.40    5.0 5.6 6.4 4.878788 0.8513803 33       0
## 3     502 2.0 3.85    4.7 5.4 6.8 4.561538 1.0703558 26       0
## 4     702 1.0 3.60    4.6 5.4 7.0 4.406667 1.4282696 30       0
## 5     902 1.6 3.40    4.2 5.0 5.2 3.990476 1.0226949 21       0
```

A more general approach uses the `group_by()` and `summarize()` functions as part of the tidyverse/dplyr packages.

```r
#Ex12.7
Friends %>%
  group_by(Friends) %>%
  summarize(N = n(),
            Mean = mean(Score),
            Std.Dev = sd(Score),
            Minimum = min(Score),
            Maximum = max(Score))
```

```
## # A tibble: 5 x 6
##   Friends     N  Mean Std.Dev Minimum Maximum
##     <int> <int> <dbl>   <dbl>   <dbl>   <dbl>
## 1     102    24  3.82   0.999     2.2     6
## 2     302    33  4.88   0.851     2.6     6.4
## 3     502    26  4.56   1.07      2       6.8
## 4     702    30  4.41   1.43      1       7
## 5     902    21  3.99   1.02      1.6     5.2
```

And we can get the confidence interval for one of the groups.

```r
#Ex12.7
confint(lm(Score ~ 1, data = filter(Friends, Friends == 102)))
```

```
##                 2.5 %   97.5 %
## (Intercept) 3.394832 4.238501
```

The ANOVA analysis can be done in a straightforward manner in R. We will create a linear model out of the Scores given their Friends group and pipe it into the anova function. (Note: the anova function takes an lm object. This is useful to us since we would maybe want to use the lm object for other purposes too.)

```r
#Ex12.8
modFriends <- lm(Score ~ Friends, data = Friends)
modFriends %>%
  anova()
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df  Sum Sq Mean Sq F value Pr(>F)
## Friends     1   0.102 0.10154  0.0767 0.7822
## Residuals 132 174.655 1.32315
```

Let's turn to page 663 and look at Example 12.15.

```r
#Ex12.15
Eyes <- read_csv("https://nhorton.people.amherst.edu/ips9/data/chapter12/EG12-15EYES.csv")
```

We will recreate the ouput from the Excel spreadsheet. With the help of the `group_by()` and `summarize()` idiom we can use the aggregating functions to summarize our dataset. To recreate the anova part of the output we use again the `anova()` function that takes an `lm` object.

```r
#Ex12.15
favstats(Score ~ Group, data = Eyes)
```

```
##   Group min  Q1 median   Q3 max     mean       sd  n missing
## 1  Blue   1 1.6    2.9 4.35 7.0 3.194030 1.754724 67       0
## 2 Brown   1 2.3    3.7 5.10 7.0 3.724324 1.715356 37       0
## 3  Down   1 1.8    2.8 4.20 6.8 3.107317 1.525351 41       0
## 4 Green   1 2.7    3.8 5.10 7.0 3.859740 1.665933 77       0
```

```r
modEyes <- lm(Score ~ Group, data = Eyes)
modEyes %>%
  anova()
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df Sum Sq Mean Sq F value  Pr(>F)
## Group       3  24.42  8.1399  2.8941 0.03618 *
## Residuals 218 613.14  2.8126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Section 12.2: Comparing the means

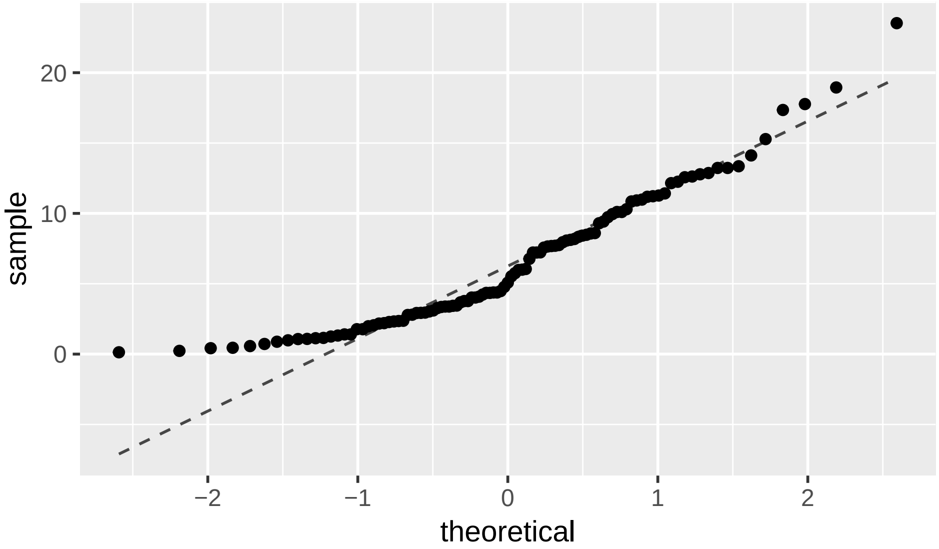
Anova gives us the answer to the question “are the differences between the means of the groups statistically significant?” However, it gives no information on what these differences are. This section covers those PostHocTests.

Let’s read in the dataset for the times for times people spend on Facebook.

```r
#Ex12.17
Facetym <- read_csv("https://nhorton.people.amherst.edu/ips9/data/chapter12/EG12-17FACETYM.csv")
```

The analysis starts by checking the data. We will check the distribution of our data with a qqplot.

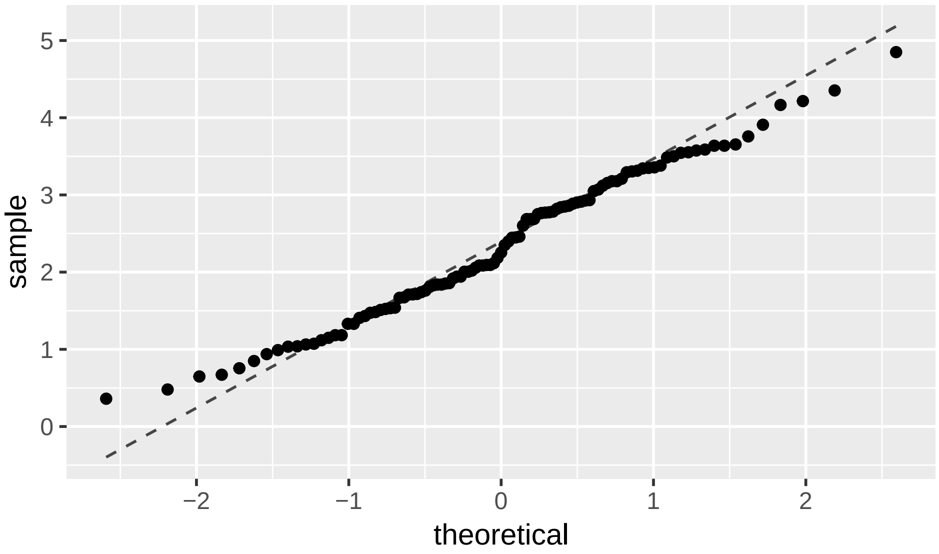
```r
#Fig12.11
gf_qq(~ Time, data = Facetym) %>%
  gf_qqline()
```



The `gf_qq()` is piped to `gf_qqline()` in order to get the reference line through the middle of the plot. Note the skewdness of the data.

We now visualize the transformed data.

```r
#Fig12.13
gf_qq(~ SqrtTime, data = Facetym) %>%
  gf_qqline()
```



Much better. Next step in our analysis is to get some descriptive statistics about the data. After that we perform the Anova test with `anova()`.

```
#Fig12.12
favstats(SqrtTime ~ Grp, data = Facetym)
```

```
##   Grp       min       Q1   median       Q3      max     mean        sd  n
## 1   1 0.8485281 1.852026 2.687006 3.304542 3.757659 2.517536 0.8503620 21
## 2   2 1.0392305 1.918333 2.751363 2.927456 4.353160 2.584771 0.8924149 21
## 3   3 1.0630146 1.708801 2.092845 3.209361 4.165333 2.404783 0.9207980 21
## 4   4 0.4795832 2.004994 2.840775 3.178050 4.849742 2.614896 1.0413608 21
## 5   5 0.3605551 1.034408 1.523155 2.004994 3.653765 1.600392 0.8340752 21
##   missing
## 1       0
## 2       0
## 3       0
## 4       0
## 5       0
```

```
modFacetym <- lm(SqrtTime ~ factor(Grp), data = Facetym)
modFacetym %>%
  anova()
```

```
## Analysis of Variance Table
##
## Response: SqrtTime
##              Df Sum Sq Mean Sq F value   Pr(>F)
## factor(Grp)   4  15.08  3.7701   4.545 0.002051 **
## Residuals   100  82.95  0.8295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the result is statistically significant, there is enough evidence to reject the null hypothesis that the mean differences are all 0. However, we now need PostHocTests in order to find out what those differences are.

We will use a package called `DescTools` which has many multiple comparisons tests for after you reject the null hypothesis based on your `anova()` output.

To use the `PostHocTest()` function we have to pass it an object returned by the `aov()` function, which does the ANOVA test for a formula (and not an `lm` object). After that we only have to specify the `method` parameter, which is equal to the name of the multiple comparisons test that you want to perform.

```
PostHocTest(aov(SqrtTime ~ factor(Grp), data = Facetym), method = "bonferroni")
```

```
##
##   Posthoc multiple comparisons of means : Bonferroni
##     95% family-wise confidence level
##
## $`factor(Grp)`
##            diff     lwr.ci       upr.ci   pval
## 2-1  0.06723591 -0.7396166  0.874088419 1.0000
## 3-1 -0.11275242 -0.9196049  0.694100090 1.0000
## 4-1  0.09736089 -0.7094916  0.904213396 1.0000
## 5-1 -0.91714318 -1.7239957 -0.110290666 0.0151 *
## 3-2 -0.17998833 -0.9868408  0.626864181 1.0000
## 4-2  0.03012498 -0.7767275  0.836977488 1.0000
## 5-2 -0.98437909 -1.7912316 -0.177526574 0.0069 **
## 4-3  0.21011331 -0.5967392  1.016965818 1.0000
## 5-3 -0.80439076 -1.6112433  0.002461756 0.0513 .
## 5-4 -1.01450406 -1.8213566 -0.207651551 0.0048 **
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
PostHocTest(aov(SqrtTime ~ factor(Grp), data = Facetym), method = "lsd")
```

```
##
##   Posthoc multiple comparisons of means : Fisher LSD
##     95% family-wise confidence level
##
## $`factor(Grp)`
##            diff     lwr.ci     upr.ci    pval
## 2-1  0.06723591 -0.4903979  0.6248697 0.81143
## 3-1 -0.11275242 -0.6703863  0.4448814 0.68916
## 4-1  0.09736089 -0.4602729  0.6549947 0.72977
## 5-1 -0.91714318 -1.4747770 -0.3595093 0.00151 **
## 3-2 -0.17998833 -0.7376222  0.3776455 0.52340
## 4-2  0.03012498 -0.5275089  0.5877588 0.91486
## 5-2 -0.98437909 -1.5420129 -0.4267453 0.00069 ***
## 4-3  0.21011331 -0.3475205  0.7677471 0.45649
## 5-3 -0.80439076 -1.3620246 -0.2467569 0.00513 **
## 5-4 -1.01450406 -1.5721379 -0.4568702 0.00048 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

XX Anova power?