Mosaic and the Less Volume, More Creativity Approach

Nicholas Horton (nhorton@amherst.edu)

August 2, 2020

Introduction

This is an illustrated example of the analyses presented in Nick Horton's JSM 2020 talk "The role of computing at the core of a modern introductory statistics course". See https://github.com/Amherst-Statistics/jse for slides and https://cran.r-project.org/web/packages/mosaic/vignettes/mosaic-resources.html for more information about resources related to mosaic.

Preliminary code

```
library(mosaic)
library(palmerpenguins)

twoisland <- penguins %>%
  mutate(island = as.character(island)) %>%
  filter(island != "Torgersen")
```

It's puzzling to me that `palmerpenguins` has island as a factor. This complicates dropping one of the islands so that we can easily demonstrate using a two-sample t-test. (I would generally hide this code from students in the first few weeks.)

Means of two groups

Mosaic code

```
# this should work in base R, but alas, it doesn't
mosaic::mean(
  bill_length_mm ~ island,
  na.rm = TRUE,
  data = twoisland)
```

```
## Biscoe  Dream
##  45.26  44.17
```

```
df_stats(
  bill_length_mm ~ island,
  data = twoisland) # mosaic helper function
```

```
##   island  min    Q1 median    Q3  max  mean    sd   n missing
## 1 Biscoe 34.5 42.00  45.80 48.70 59.6 45.26 4.773 167       1
## 2  Dream 32.1 39.15  44.65 49.85 58.0 44.17 5.954 124       0
```

One command provides a set of summaries (and provides sample size and missing values).

Equivalent in the tidyverse

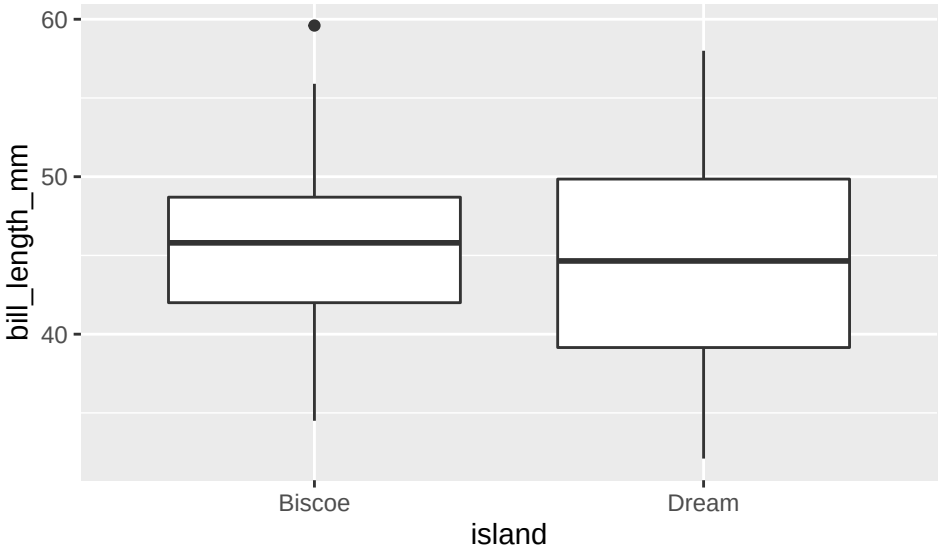
```
twoisland %>%
  group_by(island) %>%
  summarize(
    billmean = mean(
      bill_length_mm,
      na.rm = TRUE),
    n = n()
  )
```

```
## # A tibble: 2 x 3
##   island billmean     n
##   <chr>     <dbl> <int>
## 1 Biscoe     45.3   168
## 2 Dream      44.2   124
```

Graphical displays

Mosaic approach (using ggformula)

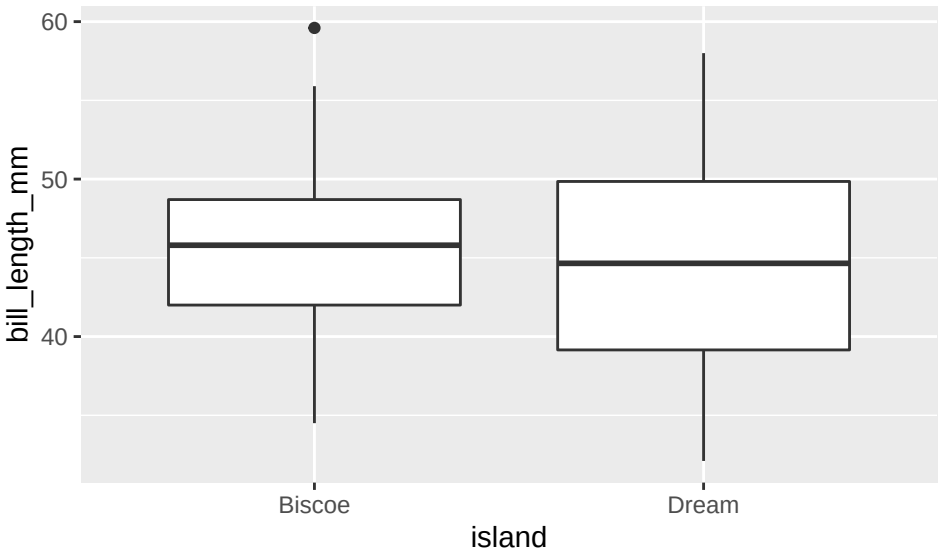
```
gf_boxplot(
  bill_length_mm ~ island,
  data = twoisland)
```



The ggformula package provides a formula interface to ggplot2 graphics

Equivalent in the tidyverse

```
ggplot(
  twoisland,
  aes(
    y = bill_length_mm,
    x = island)) +
  geom_boxplot()
```



While ggplot2 is very powerful, some aspects of the syntax (`aes()` and `+`) do not translate from the equivalent comments for descriptive statistics and modeling.

Two sample t-test

base R

```
t.test(
  bill_length_mm ~ island,
  var.equal = TRUE,
  data = twoisland)
```

```
##
##  Two Sample t-test
##
## data:  bill_length_mm by island
## t = 1.7, df = 289, p-value = 0.08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1486  2.3281
## sample estimates:
## mean in group Biscoe  mean in group Dream
##                45.26                44.17
```

another approach in base R

```
library(infer)
```

```
## Warning: package 'infer' was built under R version 4.0.2
```

```
##
## Attaching package: 'infer'
```

```
## The following objects are masked from 'package:mosaic':
##
##     prop_test, t_test
```

```
modttest <- lm(
  bill_length_mm ~ island,
  data = twoisland)
confint(modttest)
```

```
##               2.5 %  97.5 %
## (Intercept) 44.449 46.0658
## islandDream -2.328  0.1486
```

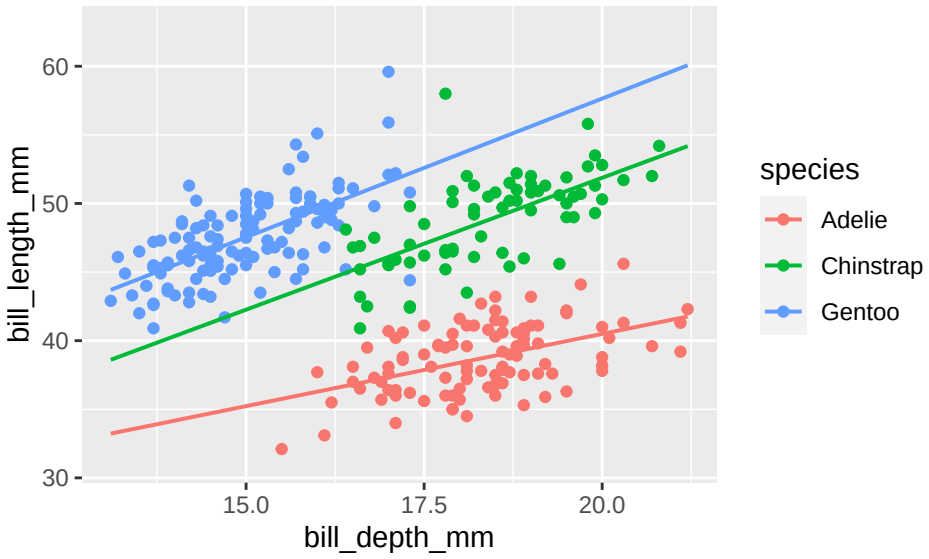
```
msummary(modttest)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   45.257      0.411  110.19   <2e-16 ***
## islandDream   -1.090      0.629   -1.73    0.084 .
##
## Residual standard error: 5.31 on 289 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.0103, Adjusted R-squared:  0.00685
## F-statistic:    3 on 1 and 289 DF,  p-value: 0.0843
```

Multiple regression

The ggformula package can be used to general scatterplots colored by species. Note that in this usage the same *pipe* operator (%>%) is used as in the tidyverse to add the regression lines for each group.

```
gf_point(
  bill_length_mm ~ bill_depth_mm,
  color = ~ species,
  data = twoisland) %>%
  gf_lm()
```




```
modmultreg <- lm(
  bill_length_mm ~ bill_depth_mm + species,
  data = twoisland)
confint(modmultreg)
```

```
##                   2.5 % 97.5 %
## (Intercept)       4.424 13.857
## bill_depth_mm     1.359  1.872
## speciesChinstrap  9.197 10.674
## speciesGentoo    13.100 15.222
```

```
msummary(modmultreg)
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.141      2.396    3.81  0.00017 ***
## bill_depth_mm       1.615      0.130   12.40  < 2e-16 ***
## speciesChinstrap    9.935      0.375   26.47  < 2e-16 ***
## speciesGentoo      14.161      0.539   26.28  < 2e-16 ***
##
## Residual standard error: 2.39 on 287 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.801,  Adjusted R-squared:  0.799
## F-statistic:  386 on 3 and 287 DF,  p-value: <2e-16
```

Bootstrapping a multiple regression model

```
set.seed(1619)
bootstraps <- do(5000) *
  lm(
    bill_length_mm ~ bill_depth_mm + species,
    data = resample(twoisland))
```

```
qdata(
  ~ speciesGentoo,
  p = c(.025, .975),
  data = bootstraps)
```

```
##  2.5% 97.5%
## 13.03 15.36
```