IS5 in R: Stats Starts Here (Chapter 1)

Margaret Chien and Nicholas Horton (nhorton@amherst.edu)

June 10, 2018

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fifth Edition of *Intro Stats* (2018) by De Veaux, Velleman, and Bock. More information about the book can be found at http://wps.aw.com/aw_deveaux_stats_series. This file as well as the associated R Markdown reproducible analysis source file used to create it can be found at http://nhorton.people.amherst.edu/is5.

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 t he *R Journal*: https://journal.r-project.org/archive/2017/RJ-2017-024.

Chapter 1: Stats Starts Here

Section 1.1: What is Statistics?

Section 1.2: Data

Section 1.3: Variables

See table on page 7.

```
library(mosaic)
library(readr)
options(digits = 3)
Tour <-
  read_csv("http://nhorton.people.amherst.edu/is5/data/Tour_de_France_2016.csv")
```

```
## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   Winner = col_character(),
##   Country = col_character(),
##   Age = col_integer(),
##   Team = col_character(),
##   `Total Time(h.min.sec)` = col_character(),
##   `Total Time(h)` = col_double(),
##   Average.Speed = col_double(),
##   Stages = col_integer(),
##   `Total Distance Ridden` = col_double(),
##   `Starting Riders` = col_integer(),
##   `Finishing Riders` = col_integer()
## )
```

```
names(Tour)
```

```
##  [1] "Year"                "Winner"
##  [3] "Country"             "Age"
##  [5] "Team"                "Total Time(h.min.sec)"
##  [7] "Total Time(h)"       "Average.Speed"
##  [9] "Stages"              "Total Distance Ridden"
## [11] "Starting Riders"     "Finishing Riders"
```

```
glimpse(Tour)
```

```
## Observations: 103
## Variables: 12
## $ Year                  <int> 1903, 1904, 1905, 1906, 1907, 1908, 19...
## $ Winner                <chr> "Maurice Garin", "Henri Cornet", "Loui...
## $ Country               <chr> "France", "France", "France", "France"...
## $ Age                   <int> 32, 20, 24, 27, 24, 25, 22, 21, 27, 24...
## $ Team                  <chr> "La Fran\u008daise", "Cycles JC", "Peu...
## $ `Total Time(h.min.sec)` <chr> "94.33.00", "96.05.56", "110.26.58", "...
## $ `Total Time(h)`       <dbl> 94.5, 96.1, 110.4, 189.6, 158.8, 156.9...
## $ Average.Speed         <dbl> 25.7, 25.3, 27.1, 24.5, 28.5, 28.7, 28...
## $ Stages                <int> 6, 6, 11, 13, 14, 14, 14, 15, 15, 15, ...
## $ `Total Distance Ridden` <dbl> 2428, 2428, 2994, 4637, 4488, 4488, 44...
## $ `Starting Riders`     <int> 60, 88, 60, 82, 93, 112, 150, 110, 84,...
## $ `Finishing Riders`    <int> 21, 27, 24, 14, 33, 36, 55, 41, 28, 41...
```

```
head(Tour, 3)
```

```
## # A tibble: 3 x 12
##    Year Winner    Country   Age Team      `Total Time(h.mi~ `Total Time(h)`
##   <int> <chr>     <chr>   <int> <chr>     <chr>                       <dbl>
## 1  1903 Maurice ~ France     32 "La Fra~  94.33.00                     94.6
## 2  1904 Henri Co~ France     20 Cycles ~  96.05.56                     96.1
## 3  1905 Louis Tr~ France     24 Peugeot   110.26.58                    110.
## # ... with 5 more variables: Average.Speed <dbl>, Stages <int>, `Total
## #   Distance Ridden` <dbl>, `Starting Riders` <int>, `Finishing
## #   Riders` <int>
```

```
tail(Tour, 8) %>%
  select(Winner, Year, Country)
```

```
## # A tibble: 8 x 3
##   Winner              Year Country
##   <chr>             <int> <chr>
## 1 Contador Alberto   2009 Spain
## 2 Andy Schleck       2010 Luxembourg
## 3 Cadel Evans        2011 Australia
## 4 Bradley Wiggins    2012 Great Britain
## 5 Christopher Froome 2013 Great Britain
## 6 Vincezo Nibali     2014 Italy
## 7 Cristopher Froome  2015 Great Britain
## 8 Cristopher Froome  2016 Great Britain
```

Let's find who was the winner in 1998

```r
filter(Tour, Year == 1998) %>%
  select(Winner, Year, Country)
```

```
## # A tibble: 1 x 3
##   Winner         Year Country
##   <chr>         <int> <chr>
## 1 Marco Pantani  1998 Italy
```

How many stages did Alberto Contador win in the years when he won the Tour?

```r
filter(Tour, Winner == "Contador Alberto") %>%
  select(Winner, Year, Stages)
```

```
## # A tibble: 2 x 3
##   Winner            Year Stages
##   <chr>            <int>  <int>
## 1 Contador Alberto  2007     21
## 2 Contador Alberto  2009     21
```

Note that the following command generates the same output.

```r
Tour %>%
  filter(Winner == "Contador Alberto") %>%
  select(Winner, Year, Stages)
```

```
## # A tibble: 2 x 3
##   Winner            Year Stages
##   <chr>            <int>  <int>
## 1 Contador Alberto  2007     21
## 2 Contador Alberto  2009     21
```

The pipe operator (%>%) can be used to connect one dataframe or command to another.

What was the slowest average speed of any tour? Fastest?

```r
filter(Tour, Average.Speed == min(Average.Speed)) %>%
  select(Year, Average.Speed)
```

```
## # A tibble: 1 x 2
##    Year Average.Speed
##   <int>         <dbl>
## 1  1919          24.1
```

```r
filter(Tour, Average.Speed == max(Average.Speed)) %>%
  select(Year, Average.Speed)
```

```
## # A tibble: 1 x 2
##    Year Average.Speed
##   <int>         <dbl>
## 1  2005          41.7
```

How can we summarize the distribution of Average Speeds?

```
favstats(~ Average.Speed, data = Tour)
```

```
##    min   Q1 median   Q3  max mean  sd   n missing
##   24.1 29.5   35.4 38.7 41.7 34.1 5.2 103       0
```