IS4 in R: Understanding and Comparing Distributions (Chapter 4)

Nicholas Horton (nhorton@amherst.edu)

July 17, 2017

Introduction and Background

This document is intended to help describe how to undertake analyses introduced as examples in the Fourth Edition of *Intro Stats* (2013) 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://www.amherst.edu/~nhorton/sdm4.

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.

Note that some of the figures in this document may differ slightly from those in the IS4 book due to small differences in datasets. However in all cases the analysis and techniques in R are accurate.

Chapter 4: Understanding and Comparing Distributions

Section 4.1: Comparing Groups with Histograms
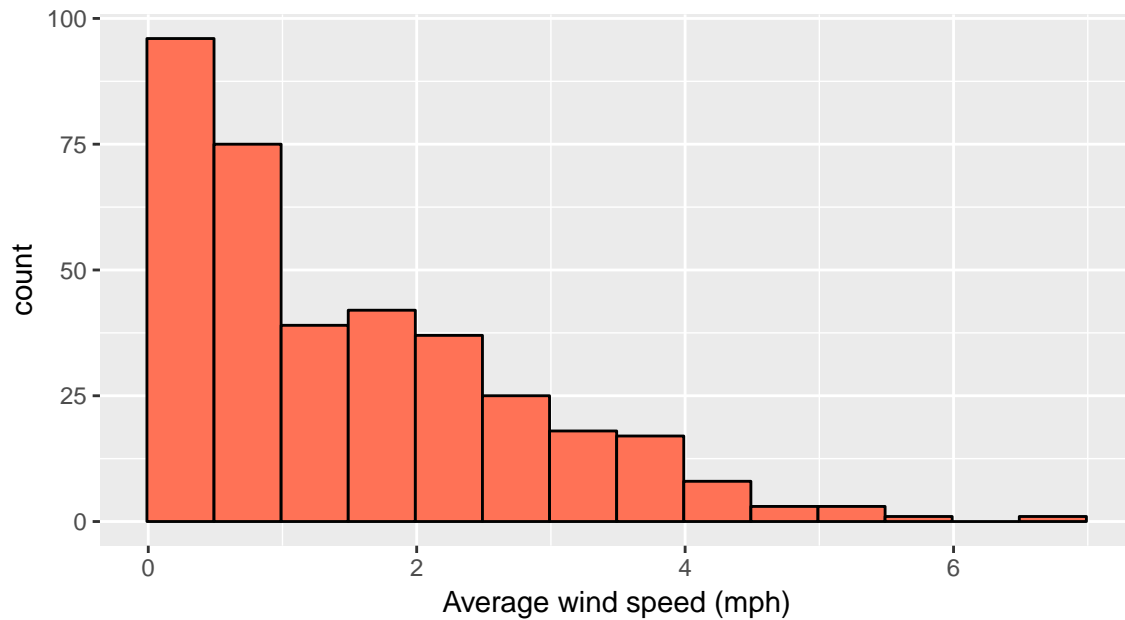
See Figure 4.1 on page 83.

```
library(mosaic); library(readr); library(ggformula)
options(digits=3)
Hopkins <- read_delim("http://www.amherst.edu/~nhorton/sdm4/data/Hopkins_Forest_2011.txt", delim="\t")
```

```
## Parsed with column specification:
## cols(
##   Season = col_character(),
##   AvgWindSpeed = col_double(),
##   Month = col_integer(),
##   Day = col_integer(),
##   DayofYear = col_integer(),
##   AvgTempC = col_double(),
##   AvgTempF = col_double(),
##   MaxWindSpeed = col_double(),
##   AvgBarom = col_double(),
##   Precip = col_double()
## )
```

```
names(Hopkins)
```

```
## [1] "Season"       "AvgWindSpeed" "Month"        "Day"
## [5] "DayofYear"    "AvgTempC"     "AvgTempF"     "MaxWindSpeed"
## [9] "AvgBarom"     "Precip"
```

```
gf_histogram(~ AvgWindSpeed, binwidth=0.5, center=0.24, data=Hopkins, fill="coral1", color=TRUE) %>%
  gf_labs(x="Average wind speed (mph)")
```


Here we reproduce Figure 4.2 on page 83.

```
Hopkins <- mutate(Hopkins,
  Summer = Month >= 4 & Month <= 9,
  Winter = !Summer
)
gf_histogram(~ AvgWindSpeed, binwidth=0.5, center=0.24, data=filter(Hopkins, Summer==TRUE), fill="blue2
  gf_labs(x="Average summer wind speed (mph)")
```
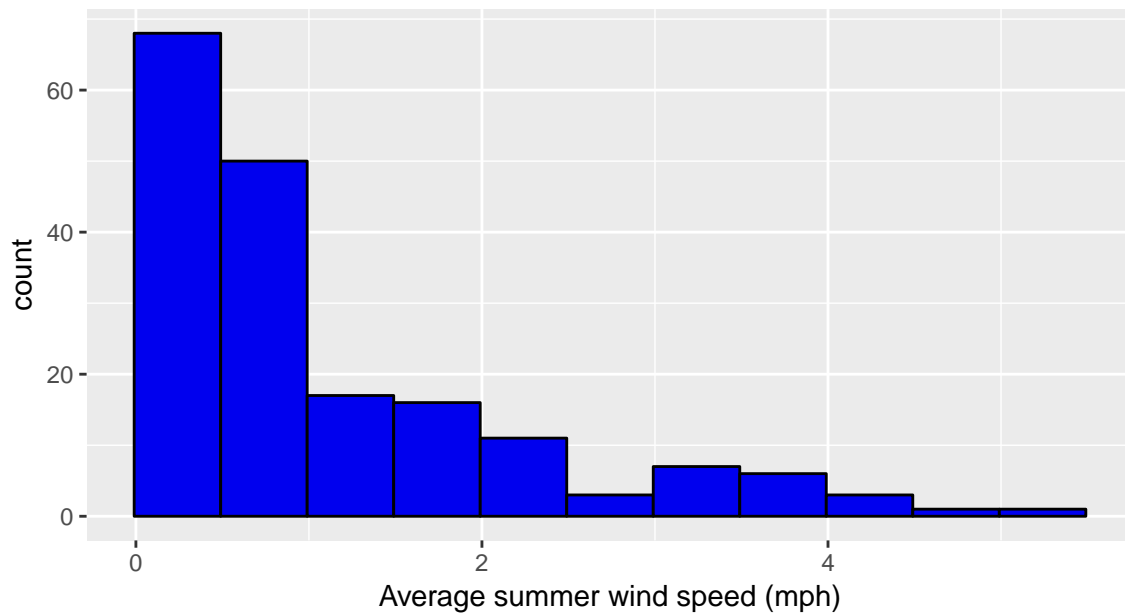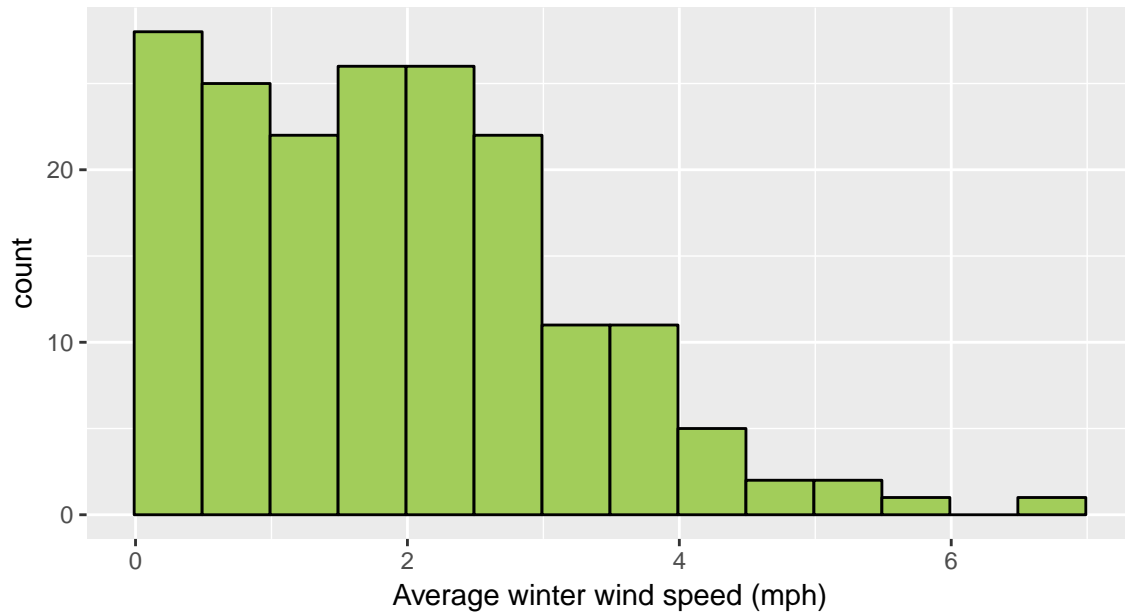


```
df_stats(~AvgWindSpeed, data=filter(Hopkins, Summer==TRUE))
```

```
##    min   Q1 median   Q3  max mean  sd    n missing
```

```
## 1    0 0.35   0.71 1.62 5.47 1.11 1.1 183        0
```
```
gf_histogram(~ AvgWindSpeed, binwidth=0.5, center=0.24, data=filter(Hopkins, Winter==TRUE),fill="darkoli
  gf_labs(x= "Average winter wind speed (mph)")
```
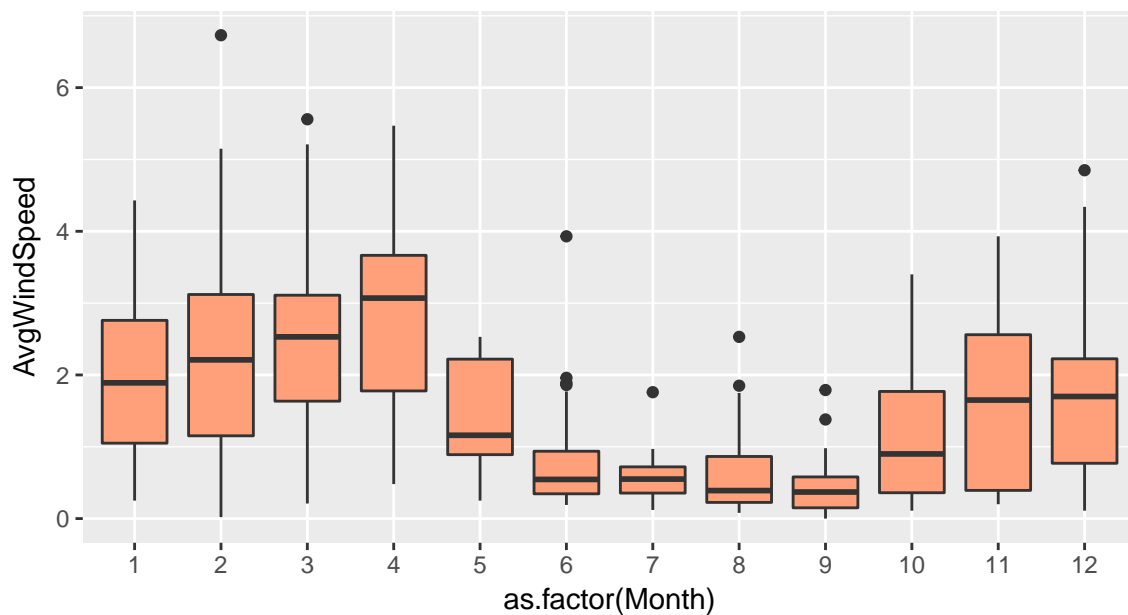


```
df_stats(~ AvgWindSpeed, data=filter(Hopkins, Winter==TRUE))
```

```
##     min  Q1 median  Q3  max mean   sd   n missing
## 1 0.02 0.84   1.72 2.66 6.73  1.9 1.29 182       0
```

Section 4.2: Comparing Groups with Boxplots

Here we reproduce Figure 4.3 on page 85.

```
gf_boxplot(AvgWindSpeed ~ as.factor(Month), data=Hopkins, fill="lightsalmon")
```


3

Section 4.3: Outliers

```
filter(Hopkins, Month==2, AvgWindSpeed > 6)    # in February
```

```
## # A tibble: 1 x 12
##   Season AvgWindSpeed Month   Day DayofYear AvgTempC AvgTempF MaxWindSpeed
##    <chr>         <dbl> <int> <int>     <int>    <dbl>    <dbl>        <dbl>
## 1 Winter         6.73     2    19        50    -5.09     22.8         39.5
## # ... with 4 more variables: AvgBarom <dbl>, Precip <dbl>, Summer <lgl>,
## #   Winter <lgl>
```

```
filter(Hopkins, Month==6, AvgWindSpeed > 3.9) # in June
```

```
## # A tibble: 1 x 12
##   Season AvgWindSpeed Month   Day DayofYear AvgTempC AvgTempF MaxWindSpeed
##    <chr>         <dbl> <int> <int>     <int>    <dbl>    <dbl>        <dbl>
## 1 Summer         3.93     6     2       153     14.7     58.5         38.8
## # ... with 4 more variables: AvgBarom <dbl>, Precip <dbl>, Summer <lgl>,
## #   Winter <lgl>
```
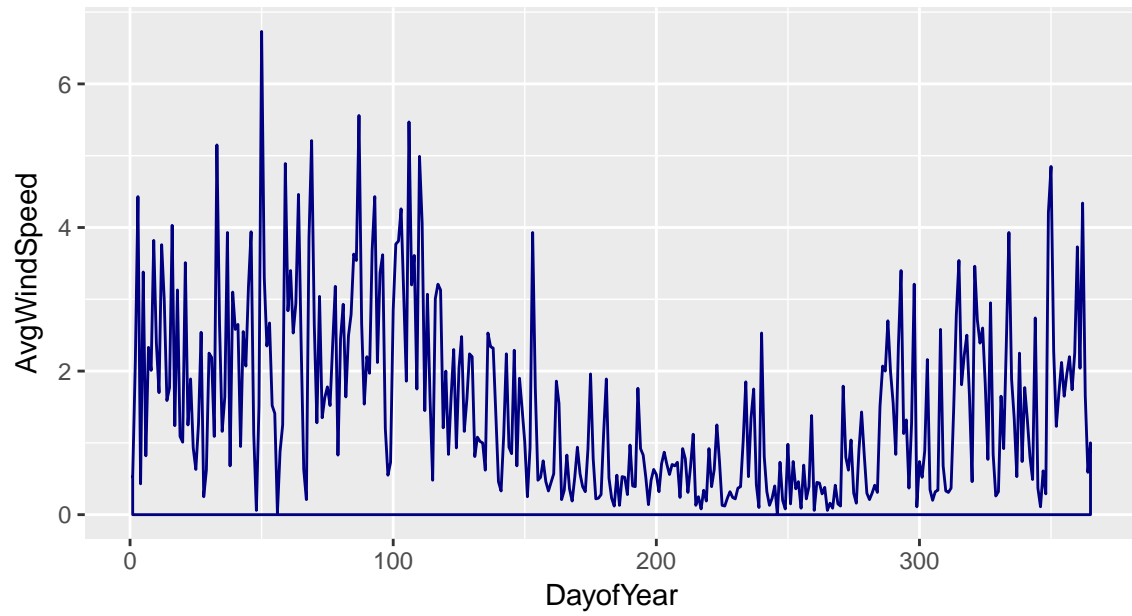
Section 4.4: Timeplots: Order, Please!

See Figures 4.4 through 4.6 starting on page 90.

```
gf_point(AvgWindSpeed ~ DayofYear, data=Hopkins, color = "mediumpurple")
```
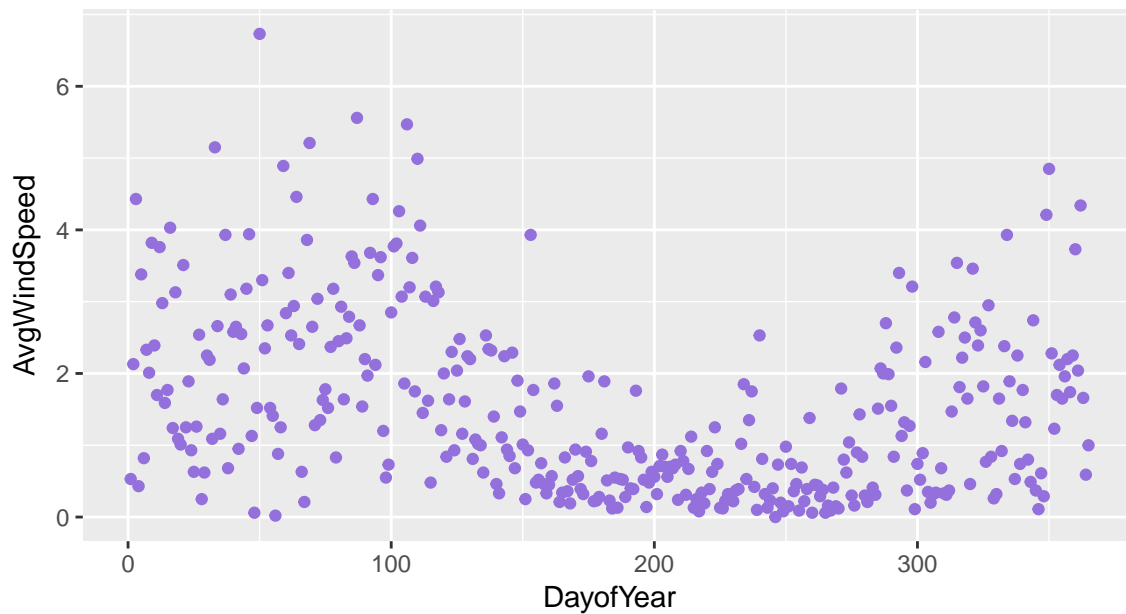


```
gf_area(AvgWindSpeed ~ DayofYear, data=Hopkins, color = "navy" ,alpha=0)
```

```
gf_point(AvgWindSpeed ~ DayofYear,data=Hopkins) %>%
  gf_smooth()
```

`geom_smooth()` using method = 'loess'