IS4 in R: Displaying and Summarizing Quantitative Data (Chapter 3)

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 3: Displaying and Summarizing Quantitative Data

Section 3.1: Displaying Quantitative Variables

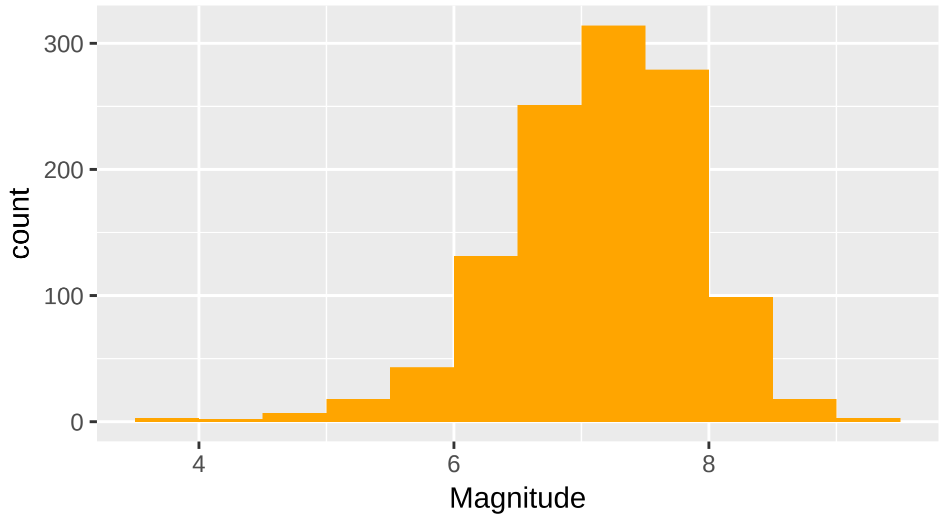
See Figure 3.1 on page 44.

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

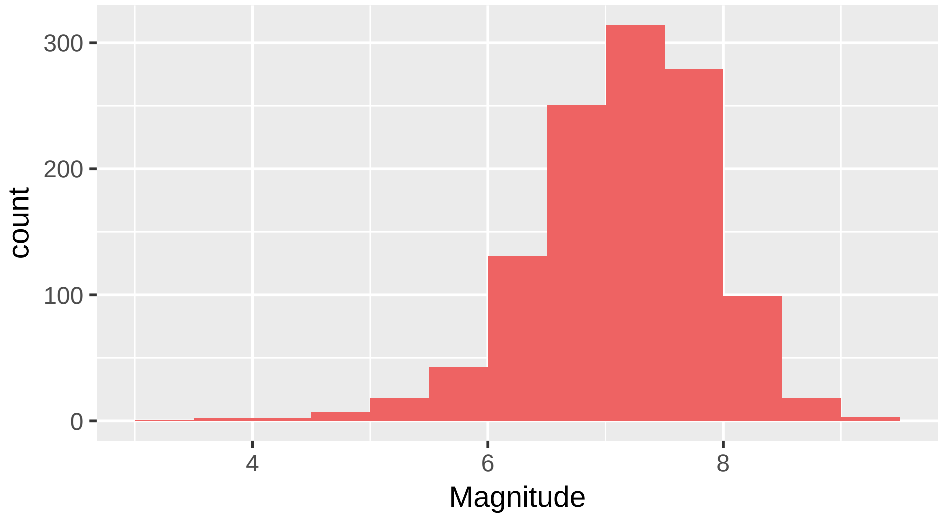
```r
nrow(Tsunami)
```

```
## [1] 1168
```

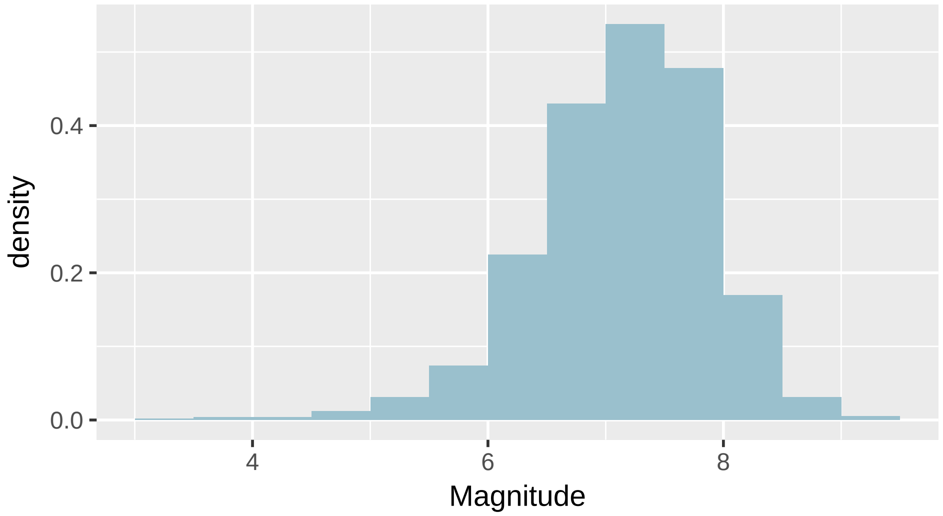
```r
gf_histogram(~ Magnitude, binwidth=0.5, center=0.5/2, type="count", data=Tsunami,fill="orange")
```



```
gf_histogram(~ Magnitude, binwidth=0.5, center=0.5/2+0.001, data=Tsunami,fill="indianred2")
```



```
gf_histogram(..density.. ~ Magnitude, binwidth=0.5, center=0.5/2+0.001, data=Tsunami,fill="lightblue3")
```



Note that Figure 3.3 on page 47 displays a histogram (with the y-axis measured by percent in each bar. The first histogram displays the count and the last the density (where the total area of the bars adds up to 1).

```
Pulse_rates <- read_delim("http://www.amherst.edu/~nhorton/sdm4/data/Pulse_rates.txt",
  delim="\t")
```

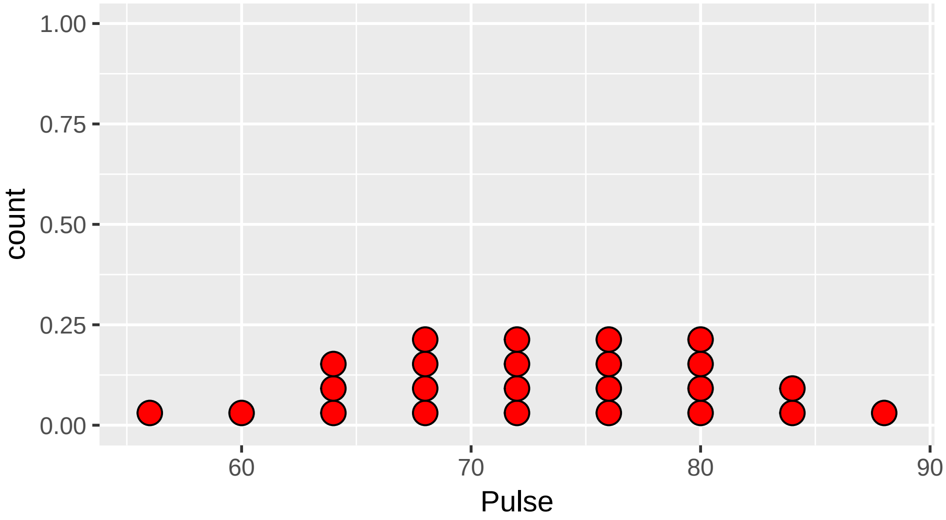
```
## Parsed with column specification:
## cols(
##   Pulse = col_integer()
## )
```

```
with(Pulse_rates, stem(Pulse))
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   5 | 6
##   6 | 04448888
##   7 | 22226666
##   8 | 0000448
```

```
gf_dotplot(~Pulse, data=Pulse_rates, fill="red")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



Or on page 47.

```
with(Pulse_rates, stem(Pulse, scale=2))
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   5 | 6
##   6 | 0444
##   6 | 8888
##   7 | 2222
##   7 | 6666
##   8 | 000044
##   8 | 8
```

Section 3.2: Shape

Section 3.3: Center

See calculation and Figure 3.11 on page 51.

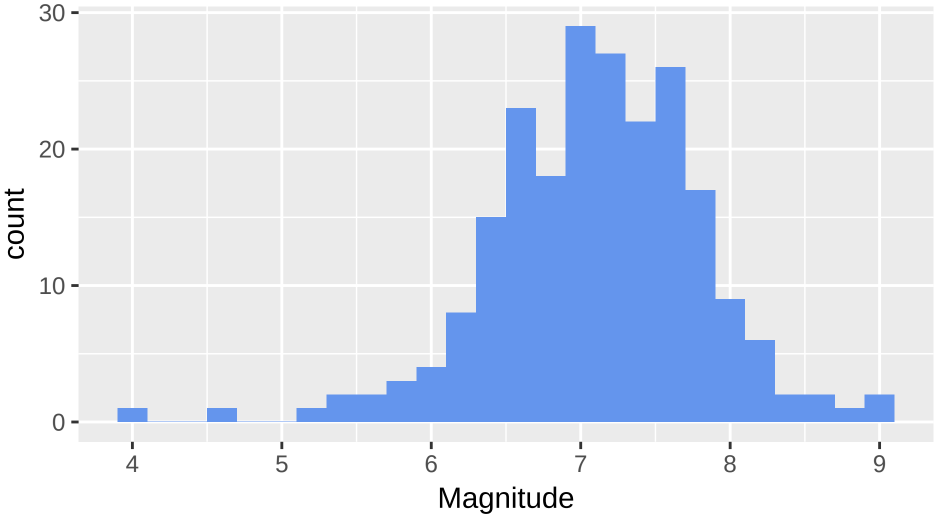
```
recent <- filter(Tsunami, Year >= 1989, Year <= 2013)
nrow(recent)
```

```
## [1] 221
```

```
median(~ Magnitude, data=recent)
```

```
## [1] 7.2
```

```
gf_histogram(~ Magnitude, binwidth=0.2, data=recent, fill="cornflowerblue")
```



Section 3.4: Spread

See statistics reported on pages 54-55.

```
df_stats(~ Magnitude, data=recent)
```

```
##   min  Q1 median  Q3 max mean    sd   n missing
## 1   4 6.7    7.2 7.6 9.1 7.15 0.702 221       0
```

```
range(~ Magnitude, data=recent)
```

```
## [1] 4.0 9.1
```

```
diff(range(~ Magnitude, data=recent))
```

```
## [1] 5.1
```

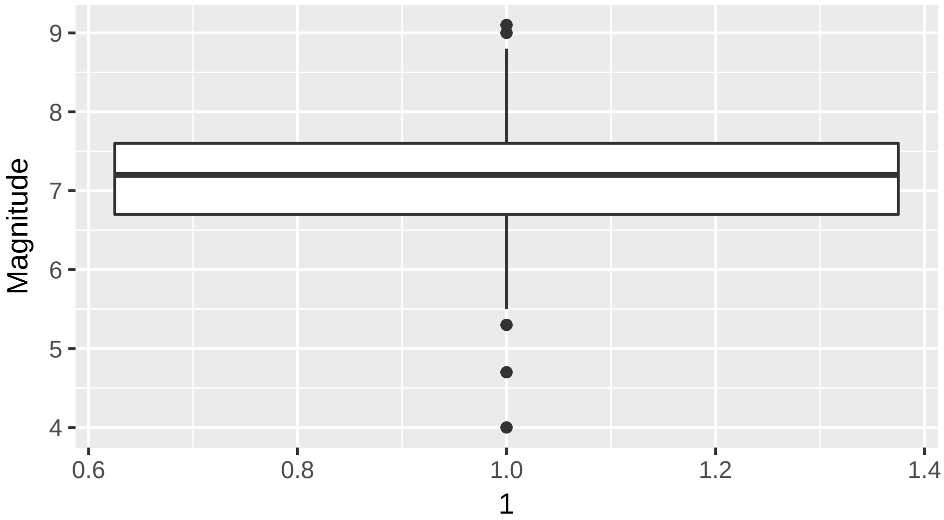
```
IQR(~ Magnitude, data=recent)
```

```
## [1] 0.9
```

Section 3.5: Boxplots and 5-Number Summaries

See display on page 55.

```
gf_boxplot(Magnitude ~ 1, data=recent)
```



Note that boxplots of a single distribution aren't usually very interesting (more useful displays will be seen in Chapter 4 when we start comparing groups).

Section 3.6: The Center of Symmetric Distributions: The Mean

See calculation on page 57.

```
mean(~ Magnitude, data=recent)
```

```
## [1] 7.15
```

```
median(~ Magnitude, data=recent)
```

```
## [1] 7.2
```

Section 3.7: The Spread of Symmetric Distributions: The Standard Deviation

To check the claim made on page 60.

```
sd(~ Magnitude, data=recent)
```

```
## [1] 0.702
```

```
var(~ Magnitude, data=recent)
```

```
## [1] 0.493
```

```
sqrt(var(~ Magnitude, data=recent))
```

```
## [1] 0.702
```

```
0.702^2
```

```
## [1] 0.493
```

The standard deviation squared equals the variance.