The Statistical Sleuth in R: Chapter 3

Linda Loi Ruobing Zhang Kate Aloisio Nicholas J. Horton*

June 15, 2016

Contents

1 Introduction

This document is intended to help describe how to undertake analyses introduced as examples in the Third Edition of the *Statistical Sleuth* (2013) by Fred Ramsey and Dan Schafer. More information about the book can be found at `http://www.proaxis.com/~panorama/home.htm`. This file as well as the associated `knitr` reproducible analysis source file can be found at `http://www.math.smith.edu/~nhorton/sleuth3`.

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 vignette (`http://cran.r-project.org/web/packages/mosaic/vignettes/MinimalR.pdf`).

To use a package within R, it must be installed (one time), and loaded (each session). The package can be installed using the following command:

*Department of Mathematics and Statistics, Smith College, nhorton@smith.edu

```
> install.packages('mosaic')                # note the quotation marks
```

Once this is installed, it can be loaded by running the command:

```
> require(mosaic)
```

This needs to be done once per session.

In addition the data files for the *Sleuth* case studies can be accessed by installing the `Sleuth3` package.

```
> install.packages('Sleuth3')                # note the quotation marks
```

```
> require(Sleuth3)
```

We also set some options to improve legibility of graphs and output.

```
> trellis.par.set(theme=col.mosaic())  # get a better color scheme for lattice
> options(digits=3, show.signif.stars=FALSE)
```

The specific goal of this document is to demonstrate how to calculate the quantities described in *Sleuth* Chapter 3: A Closer Look at Assumptions using R.

2 Cloud Seeding to Increase Rainfall

Does seeding clouds lead to more rainfall? This is the question being addressed by case study 3.1 in the *Sleuth*.

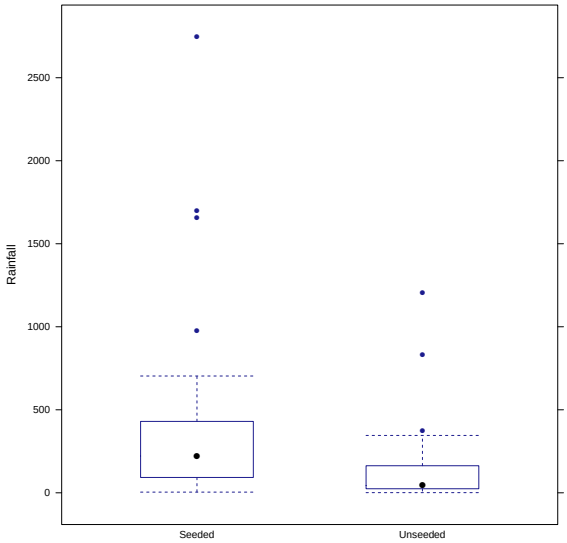
2.1 Summary statistics and graphical displays (untransformed)

We begin by reading the data and summarizing the variables.

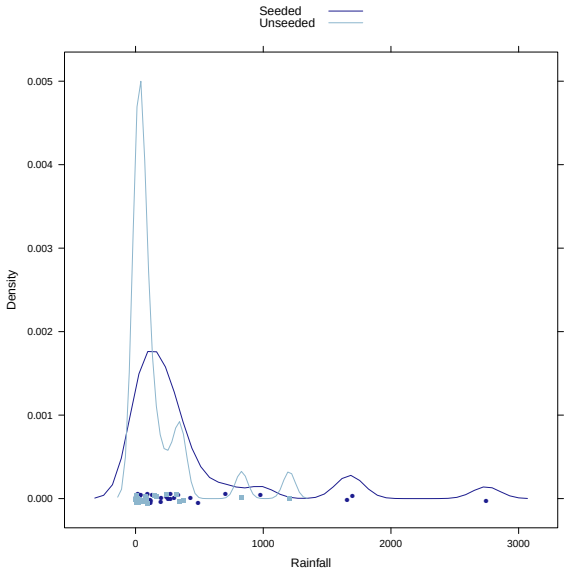
```
> summary(case0301)

    Rainfall       Treatment
 Min.   :   1   Seeded  :26
 1st Qu.:  29   Unseeded:26
 Median : 117
 Mean   : 303
 3rd Qu.: 307
 Max.   :2746

> favstats(Rainfall ~ Treatment, data=case0301)

  Treatment min   Q1 median  Q3  max mean  sd  n missing
1    Seeded 4.1 98.1  221.6 406 2746  442 651 26       0
2  Unseeded 1.0 24.8   44.2 159 1203  165 278 26       0
```

A total of 52 subjects were included in this data: 26 seeded days and 26 unseeded days (Display 3.1, page 59).

```
> bwplot(Rainfall ~ Treatment, data=case0301)
```




```
> densityplot(~Rainfall, groups=Treatment, auto.key=TRUE, data=case0301)
```



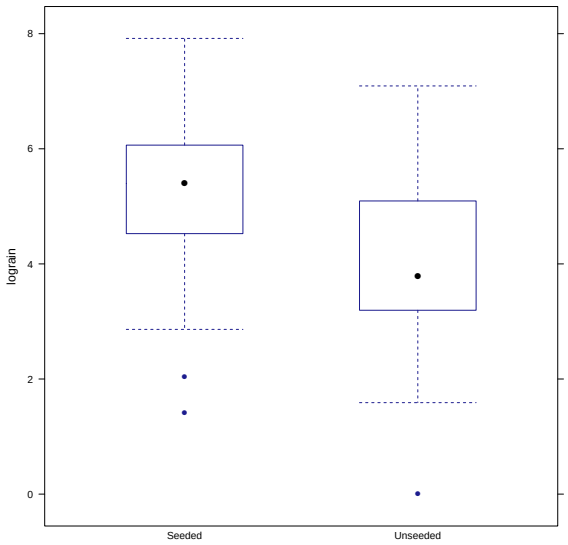
According to the boxplot and the density plot, the rainfall from seeded days seems to be larger than unseeded days. Both density curves are highly skewed to the right.

2.2 Summary statistics and graphical display (transformed)

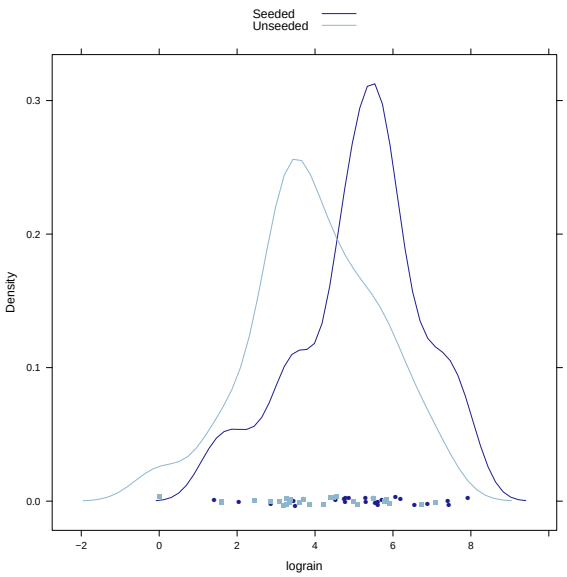
The skewness suggests that there is a need to apply a logarithmic transformation. The transformed data is shown on page 73 (Display 3.9).

```
> case0301 = transform(case0301, lograin=log(Rainfall))
> favstats(lograin ~ Treatment, data=case0301)

  Treatment  min   Q1 median   Q3  max mean   sd  n missing
1    Seeded 1.41 4.58   5.40 6.00 7.92 5.13 1.60 26       0
2  Unseeded 0.00 3.21   3.79 5.07 7.09 3.99 1.64 26       0
```

```
> bwplot(lograin ~ Treatment, data=case0301)
```




```
> densityplot(~lograin, groups=Treatment, auto.key=TRUE, data=case0301)
```



The log transformation reduces the skewness of these two distributions.

2.3 Inferential procedures (two-sample t-test)

```
> t.test(Rainfall ~ Treatment, var.equal=FALSE, data=case0301)


Welch Two Sample t-test

data:  Rainfall by Treatment
t = 2, df = 30, p-value = 0.05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -4.76 559.56
sample estimates:
  mean in group Seeded mean in group Unseeded
                   442                     165

> t.test(Rainfall ~ Treatment, var.equal=TRUE, data=case0301)


Two Sample t-test

data:  Rainfall by Treatment
t = 2, df = 50, p-value = 0.05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -1.43 556.22
sample estimates:
```

```
  mean in group Seeded mean in group Unseeded
                   442                     165
```

The following corresponds to the calculations on page 73.

```
> summary(lm(lograin ~ Treatment, data=case0301))


Call:
lm(formula = lograin ~ Treatment, data = case0301)

Residuals:
   Min     1Q Median     3Q    Max
-3.990 -0.745  0.162  1.019  3.102

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.134      0.318   16.15   <2e-16
TreatmentUnseeded   -1.144      0.450   -2.54    0.014

Residual standard error: 1.62 on 50 degrees of freedom
Multiple R-squared:  0.115,Adjusted R-squared:  0.0969
F-statistic: 6.47 on 1 and 50 DF,  p-value: 0.0141

> ttestlog = t.test(lograin ~ Treatment, data=case0301); ttestlog


Welch Two Sample t-test

data:  lograin by Treatment
t = 3, df = 50, p-value = 0.01
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.241 2.047
sample estimates:
  mean in group Seeded mean in group Unseeded
                  5.13                    3.99
```

2.4 Interpretation of log model

The following code is used to calculate the "Statistical Conclusion" on page 59. First, we want to calculate the multiplier.

```
> obslogdiff = -diff(mean(lograin ~ Treatment, data=case0301)); obslogdiff
```

```
Unseeded
    1.14

> multiplier = exp(obslogdiff); multiplier

Unseeded
    3.14
```

Next we can calculate the 95% confidence interval for the multiplier.

```
> ttestlog$conf.int

[1] 0.241 2.047
attr(,"conf.level")
[1] 0.95

> exp(ttestlog$conf.int)

[1] 1.27 7.74
attr(,"conf.level")
[1] 0.95
```

3 Effects of Agent Orange on Troops in Vietnam

Is dioxin concentration related to veteran status? This is the question being addressed by case study 3.2 in the *Sleuth*.

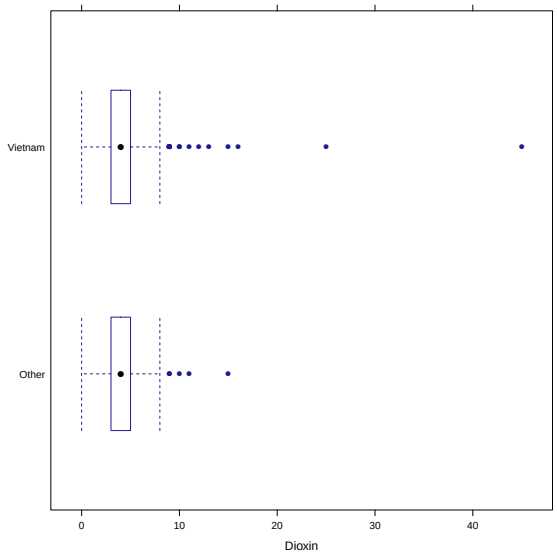
3.1 Summary statistics and graphical display

We begin by reading the data and summarizing the variables.

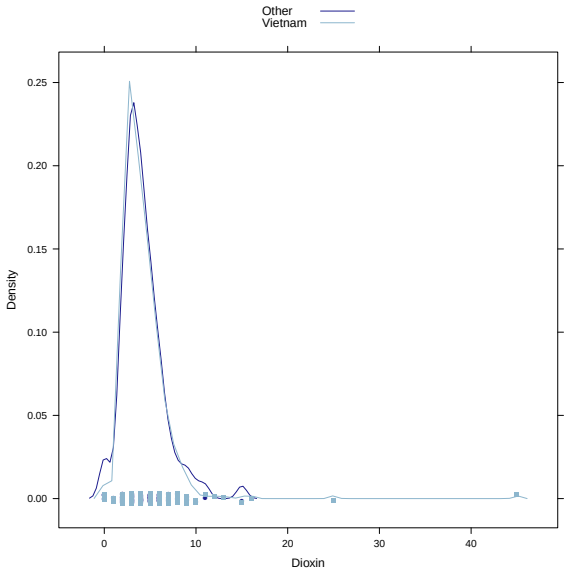
```
> summary(case0302)

     Dioxin        Veteran
 Min.   : 0.0   Other  : 97
 1st Qu.: 3.0   Vietnam:646
 Median : 4.0
 Mean   : 4.3
 3rd Qu.: 5.0
 Max.   :45.0

> favstats(Dioxin ~ Veteran, data=case0302)

  Veteran min Q1 median Q3 max mean   sd   n missing
1   Other   0  3      4  5  15 4.19 2.30  97       0
2 Vietnam   0  3      4  5  45 4.26 2.64 646       0
```

A total of 743 veterans were included in this data: 646 served in Vietnam during 1967 and 1968 and 97 served in US or Germany during 1965 and 1971.

```
> bwplot(Veteran ~ Dioxin, data=case0302)
```




```
> densityplot(~Dioxin, groups=Veteran, auto.key=TRUE, data=case0302)
```



Both distributions are highly skewed to the right.

3.2 Inferential procedures (two-sample t-test)

The following code is used to calculate the "Statistical Conclusion" on page 62.

```
> t.test(Dioxin ~ Veteran, var.equal=TRUE, alternative="less", data=case0302)


Two Sample t-test

data:  Dioxin by Veteran
t = -0.3, df = 700, p-value = 0.4
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
  -Inf 0.392
sample estimates:
  mean in group Other mean in group Vietnam
                 4.19                  4.26

> t.test(Dioxin ~ Veteran, var.equal=TRUE, data=case0302)$conf.int

[1] -0.631  0.482
attr(,"conf.level")
[1] 0.95
```

So the one-sided p-value from a two-sample t-test is 0.396. The 95% confidence interval is (-0.63, 0.48). Notice that because of the way we ordered our variables, the confidence interval shown in our analysis is different from that of the book (our confidence intervals are inverse). This is of no consequence, as the difference between the groups is still the same.

3.3 Removing outliers

We will remove two extreme observations from the data. First we remove observation 646 and perform a t-test (Display 3.7, page 70).

```
> case0302.2 = case0302[-c(646), ]
> t.test(Dioxin ~ Veteran, alternative="less", data=case0302.2)


Welch Two Sample t-test

data:  Dioxin by Veteran
t = -0.05, df = 100, p-value = 0.5
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf  0.4
sample estimates:
  mean in group Other mean in group Vietnam
                 4.19                  4.20
```

Next we remove observations 645 and 646 and perform a t-test.

```
> dim(case0302)

[1] 743   2

> case0302.3 = case0302[-c(645, 646), ]
> dim(case0302.3)

[1] 741   2

> t.test(Dioxin ~ Veteran, alternative="less", data=case0302.3)


Welch Two Sample t-test

data:  Dioxin by Veteran
t = 0.09, df = 100, p-value = 0.5
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
  -Inf 0.429
sample estimates:
  mean in group Other mean in group Vietnam
                 4.19                  4.16
```

Notice that after removing these outliers, the p-value and the confidence interval have changed but the substantive conclusion is unchanged.