The Statistical Sleuth in R: Chapter 3

Ruobing Zhang Kate Aloisio Nicholas J. Horton*

September 7, 2013

Contents

1 Introduction

This document is intended to help describe how to undertake analyses introduced as examples in the Second Edition of the *Statistical Sleuth* (2002) 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.amherst.edu/~nhorton/sleuth`.

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, Amherst College, nhorton@amherst.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 Sleuth2 package.

```
> install.packages("Sleuth2")   # note the quotation marks
```

```
> require(Sleuth2)
```

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*.

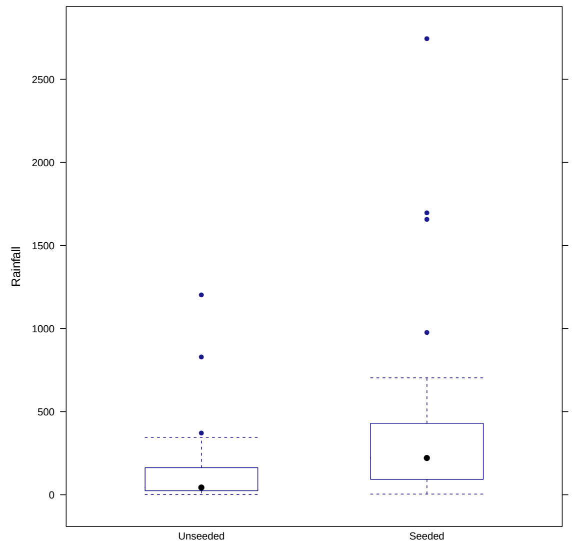
2.1 Summary statistics and graphical displays (untransformed)

We begin by reading the data and summarizing the variables.

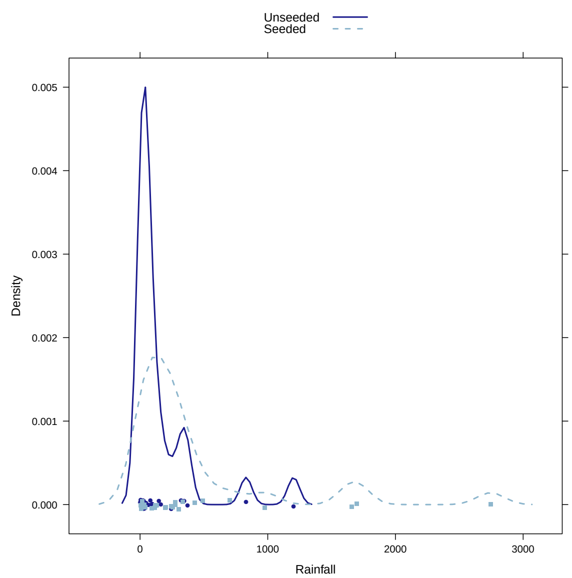
```
> summary(case0301)

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

> favstats(Rainfall ~ Treatment, data = case0301)

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

A total of 52 subjects were included in this data: 26 seeded days and 26 unseeded days (Display 3.1, page 57).

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




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



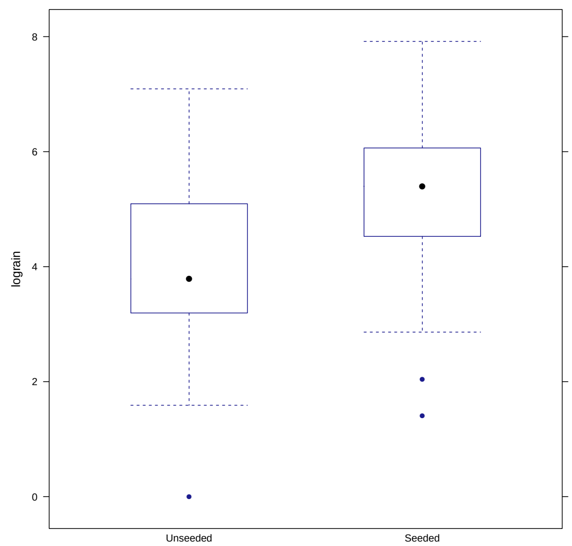
According to the boxplot and the density plot, the rainfall from seeded days seems to be larger than unseeded days. Both density curves are highly skewed to the right.

2.2 Summary statistics and graphical display (transformed)

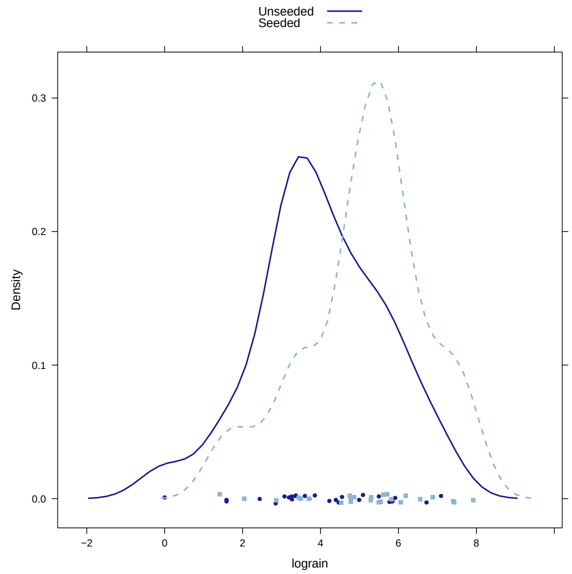
The skewness suggests there is a need to apply the logarithmic transformation. The transformed data is shown on page 71 (Display 3.9).

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

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

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




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



The log transformation reduces 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 = 33.9, p-value = 0.05377
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -559.56    4.76
sample estimates:
mean in group Unseeded   mean in group Seeded
                   165                    442

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


Two Sample t-test

data:  Rainfall by Treatment
t = -2, df = 50, p-value = 0.05114
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -556.22    1.43
sample estimates:
```

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

The following corresponds to the calculations on page 71.

```
> 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)        3.990      0.318   12.55   <2e-16
TreatmentSeeded    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 = -2.54, df = 50, p-value = 0.01408
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.047 -0.241
sample estimates:
mean in group Unseeded   mean in group Seeded
                  3.99                   5.13
```

The two-sided p-value is $p = 0.014$ and the 95% confidence interval is between -2.05 and -0.24.

2.4 Interpretation of log model

The following code is used to calculate the "Summary of Statistical Findings" on page 57. First, we want to calculate the multiplier.

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

Seeded
 -1.14

> multiplier = exp(obslogdiff)
> multiplier

Seeded
 0.319
```

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

```
> ttestlog$conf.int

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

> exp(ttestlog$conf.int)

[1] 0.129 0.786
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

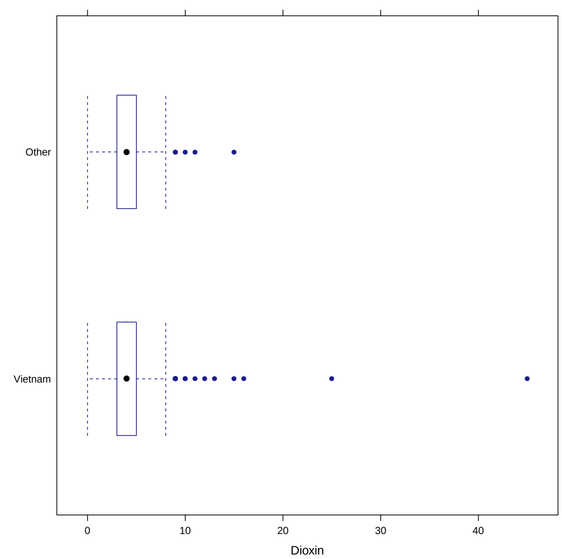
We begin by reading the data and summarizing the variables.

```
> summary(case0302)

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

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

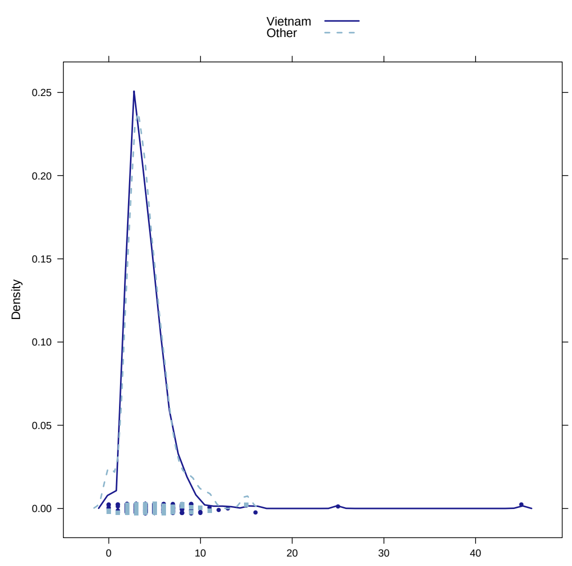
```
        min Q1 median Q3 max mean   sd   n missing
Vietnam   0  3      4  5  45 4.26 2.64 646       0
Other     0  3      4  5  15 4.19 2.30  97       0
```

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 “Summary of Statistical Findings” on page 60.

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


Two Sample t-test

data:  Dioxin by Veteran
t = 0.263, df = 741, p-value = 0.6037
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
  -Inf 0.541
sample estimates:
mean in group Vietnam   mean in group Other
                 4.26                  4.19

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

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

So the one-sided p-value from a two-sample t-test is 0.604. The 95% confidence interval is (-0.48, 0.63).

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 67).

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


Welch Two Sample t-test

data:  Dioxin by Veteran
t = 0.0457, df = 121, p-value = 0.5182
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
  -Inf 0.422
sample estimates:
mean in group Vietnam   mean in group Other
                 4.20                  4.19
```

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 = -0.0853, df = 117, p-value = 0.4661
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
  -Inf 0.387
sample estimates:
mean in group Vietnam   mean in group Other
                 4.16                  4.19
```

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