The Statistical Sleuth in R:
Chapter 3

Ruobing Zhang    Kate Aloisio    Nicholas J. Horton*

September 7, 2013

Contents

1 Introduction 1

2 Cloud Seeding to Increase Rainfall 2
   2.1 Summary statistics and graphical displays (untransformed) 2
   2.2 Summary statistics and graphical display (transformed) 4
   2.3 Inferential procedures (two-sample t-test) 5
   2.4 Interpretation of log model 6

3 Effects of Agent Orange on Troops in Vietnam 7
   3.1 Summary statistics and graphical display 7
   3.2 Inferential procedures (two-sample t-test) 9
   3.3 Removing outliers 9

1 Introduction

This document is intended to help describe how to undertake analyses introduced as examples in the
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)

<table>
<thead>
<tr>
<th>Rainfall</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. : 1</td>
<td>Unseeded:26</td>
</tr>
<tr>
<td>1st Qu.: 29</td>
<td>Seeded :26</td>
</tr>
<tr>
<td>Median : 117</td>
<td></td>
</tr>
<tr>
<td>Mean : 303</td>
<td></td>
</tr>
<tr>
<td>3rd Qu. : 307</td>
<td></td>
</tr>
<tr>
<td>Max. : 2746</td>
<td></td>
</tr>
</tbody>
</table>

> favstats(Rainfall ~ Treatment, data = case0301)

<table>
<thead>
<tr>
<th>min</th>
<th>Q1</th>
<th>median</th>
<th>Q3</th>
<th>max</th>
<th>mean</th>
<th>sd</th>
<th>n</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unseeded</td>
<td>1.0</td>
<td>24.8</td>
<td>44.2</td>
<td>159</td>
<td>165</td>
<td>278</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>Seeded</td>
<td>4.1</td>
<td>98.1</td>
<td>221.6</td>
<td>406</td>
<td>442</td>
<td>651</td>
<td>26</td>
<td>0</td>
</tr>
</tbody>
</table>
A total of 52 subjects were included in this data: 26 seeded days and 26 unseeded days (Display 3.1, page 57).

\[
\text{bwplot(Rainfall ~ Treatment, data = case0301)}
\]

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).

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

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>Q1</th>
<th>median</th>
<th>Q3</th>
<th>max</th>
<th>mean</th>
<th>sd</th>
<th>n</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unseeded</td>
<td>0.00</td>
<td>3.21</td>
<td>3.79</td>
<td>5.07</td>
<td>7.09</td>
<td>3.99</td>
<td>1.64</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>Seeded</td>
<td>1.41</td>
<td>4.58</td>
<td>5.40</td>
<td>6.00</td>
<td>7.92</td>
<td>5.13</td>
<td>1.60</td>
<td>26</td>
<td>0</td>
</tr>
</tbody>
</table>
```

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

```r
> 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)

```r
> 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:
```
2 CLOUD SEEDING TO INCREASE RAINFALL

<table>
<thead>
<tr>
<th>mean in group Unseeded</th>
<th>mean in group Seeded</th>
</tr>
</thead>
<tbody>
<tr>
<td>165</td>
<td>442</td>
</tr>
</tbody>
</table>

The following corresponds to the calculations on page 71.

```r
> 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) 3.9900     0.3181  12.55   <2e-16
TreatmentSeeded 1.1442     0.4502   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

```r
> 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.

Statistical Sleuth in R: Chapter 3
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.

```r
> 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
```

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

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

```r
> 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

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>Q1</th>
<th>median</th>
<th>Q3</th>
<th>max</th>
<th>mean</th>
<th>sd</th>
<th>n</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnam</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>45</td>
<td>4.26</td>
<td>2.64</td>
<td>646</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>15</td>
<td>4.19</td>
<td>2.30</td>
<td>97</td>
<td>0</td>
</tr>
</tbody>
</table>

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)`

Both distributions are highly skewed to the right.

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

Statistical Sleuth in R: Chapter 3
3.2 Inferential procedures (two-sample t-test)

The following code is used to calculate the “Summary of Statistical Findings” on page 60.

```r
> 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).

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