The Statistical Sleuth in R: Chapter 2

Ruobing Zhang Kate Aloisio Nicholas J. Horton*

September 16, 2012

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.math.smith.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:

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

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

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

```
> 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 Chapter 2: Inference Using t-Distributions using R.

2 Bumpus's Data on Natural Selection

Is humerus length related to whether the bird would survive or perish? That's the question being addressed by Case Study 2.1 in the *Sleuth*.

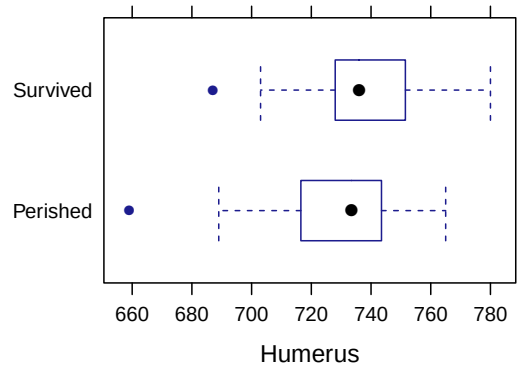
2.1 Statistical summary and graphical display

We begin by reading the data and summarizing the variables.

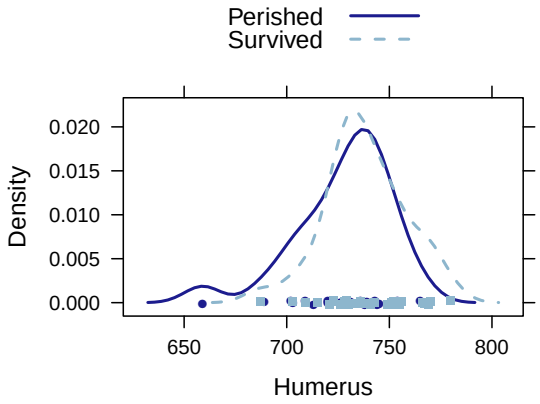
```
> summary(case0201)

    Humerus         Status
 Min.   :659   Perished:24
 1st Qu.:724   Survived:35
 Median :736
 Mean   :734
 3rd Qu.:747
 Max.   :780

> fav = favstats(Humerus ~ Status, data = case0201)
> fav

          min  Q1 median  Q3 max mean   sd  n missing
Perished 659 718    734 743 765  728 23.5 24       0
Survived 687 728    736 752 780  738 19.8 35       0
```

A total of 59 subjects are included in the data: 35 are adult male sparrows that survived and 24 that perished. The following figure replicates Display 2.1 on page 29.

```
> bwplot(Status ~ Humerus, data = case0201)
```




```
> densityplot(~Humerus, groups = Status, auto.key = TRUE, data = case0201)
```



Both distributions are approximately normally distributed.

2.2 Inferential procedures (two-sample t-test)

First, we calculate the pooled SD and the standard error between these two different sample average (page 40, Display 2.8).

```
> # Calculate Pooled SD
> n1 = fav["Perished", "n"]
> n1

[1] 24

> n2 = fav["Survived", "n"]
> n2

[1] 35
```

```
> s1 = fav["Perished", "sd"]
> s1

[1] 23.5

> s2 = fav["Survived", "sd"]
> s2

[1] 19.8

> Sp = sqrt(((n1 - 1) * (s1)^2 + (n2 - 1) * (s2)^2)/(n1 + n2 - 2))
> Sp

[1] 21.4

> # Calculate standard error
> SE = Sp * sqrt(1/n1 + 1/n2)
> SE

[1] 5.67
```

So the pooled SD is 21.41 and the standard error is 5.7.
Based on this information, we can construct a 95% confidence interval (page 41, Display 2.9).

```
> Y1 = fav["Perished", "mean"]
> Y1

[1] 728

> Y2 = fav["Survived", "mean"]
> Y2

[1] 738

> Yd = Y2 - Y1
> Yd

[1] 10.1

> df = n1 + n2 - 2
> df

[1] 57

> qt = qt(0.975, 57)
> qt

[1] 2
```

```
> hw = qt * SE
> hw

[1] 11.4

> lower = Yd - hw
> lower

[1] -1.28

> upper = Yd + hw
> upper

[1] 21.4
```

So the 95% confidence interval of the difference between means is (-1.3, 21.4)

Now we want to calculate the t-statistic and p-value (as shown on page 44, Display 2.10).

```
> tstats = (Yd - 0)/SE
> tstats  # The hypothesis difference=0

[1] 1.78

> onepval = 1 - pt(tstats, df)
> onepval

[1] 0.0405

> twopval = 2 * onepval
> twopval

[1] 0.0809
```

The one-sided p-value is 0.04 and the two-sided p-value is 0.08.

We can get the results of "Summary of Statistical Findings" (page 29) by using the following code:

```
> t.test(Humerus ~ Status, var.equal = TRUE, data = case0201)


Two Sample t-test

data:  Humerus by Status
t = -1.78, df = 57, p-value = 0.0809
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.45   1.28
```

```
sample estimates:
mean in group Perished mean in group Survived
                   728                    738


> confint(lm(Humerus ~ Status, data = case0201))

                2.5 % 97.5 %
(Intercept)    719.17  736.7
StatusSurvived  -1.28   21.4
```

3 Anatomical Abnormalities Associated with Schizophrenia

Is the area of brain related to the development of schizophrenia? That's the question being addressed by case study 2.2 in the *Sleuth*.

3.1 Statistical summary and graphical display

We begin by reading the data and summarizing the variables.

```
> summary(case0202)

    Unaffect       Affected
 Min.   :1.25   Min.   :1.02
 1st Qu.:1.60   1st Qu.:1.31
 Median :1.77   Median :1.59
 Mean   :1.76   Mean   :1.56
 3rd Qu.:1.94   3rd Qu.:1.78
 Max.   :2.08   Max.   :2.02
```

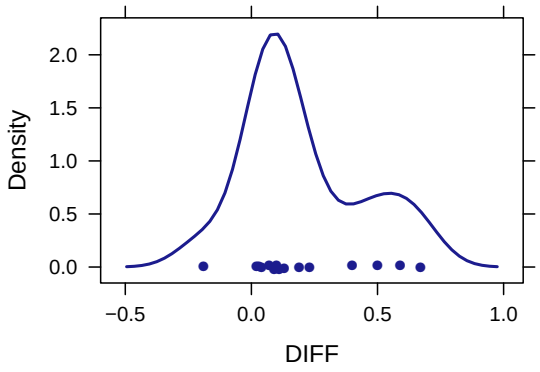
A total of 15 subjects are included in the data. There are 15 pairs of twins; one of the twins has schizophrenia, and the other does not. So there are 15 affected subjects and 15 unaffected subjects.

The difference in area of left hippocampus of these pairs of twins is:

```
> DIFF = case0202[, "Unaffect"] - case0202[, "Affected"]
> favstats(DIFF)

   min    Q1 median    Q3  max  mean    sd  n missing
 -0.19 0.055   0.11 0.315 0.67 0.199 0.238 15       0
```

This matches the results on page 30, Display 2.2.

```
> densityplot(DIFF)
```



3.2 Inferential procedures (two-sample t-test)

We want to calculate the paired t-test and 95% confidence interval.

```
> # Calculate t-statistics
> difmean = favstats(DIFF)[, "mean"]
> difmean

[1] 0.199

> difsd = favstats(DIFF)[, "sd"]
> difsd

[1] 0.238

> difSE = difsd/sqrt(15)
> difSE

[1] 0.0615

> tscore = (difmean - 0)/difSE
> tscore  # hypothesis difference=0

[1] 3.23

> twopvalue = 2 * (1 - pt(tscore, 15 - 1))
> twopvalue

[1] 0.00606

> # Construct confidence interval
> q = qt(0.975, 15 - 1)
> q
```

```
[1] 2.14

> schizolower = difmean - q * difSE
> schizolower

[1] 0.0667

> schizoupper = difmean + q * difSE
> schizoupper

[1] 0.331
```

So the two-sided p-value is 0.006 and the 95% confidence interval is (0.07, 0.33).
Or we can get the results displayed on page 31 by conducting a paired t-test.

```
> t.test(case0202[, "Unaffect"], case0202[, "Affected"], paired = TRUE)


Paired t-test

data:  case0202[, "Unaffect"] and case0202[, "Affected"]
t = 3.23, df = 14, p-value = 0.006062
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0667 0.3306
sample estimates:
mean of the differences
                  0.199
```