The Statistical Sleuth in R: Chapter 2

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:

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

2 Evidence Supporting Darwin's Theory of Natural Selection

Do birds evolve to adapt to their environments? That's the question being addressed by Case Study 2.1 in the *Sleuth*.

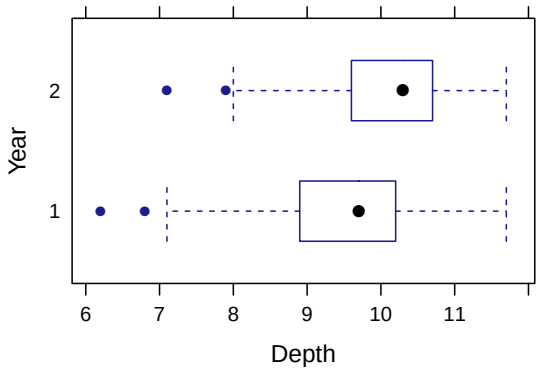
2.1 Statistical summary and graphical display

We begin by reading the data and summarizing the variables.

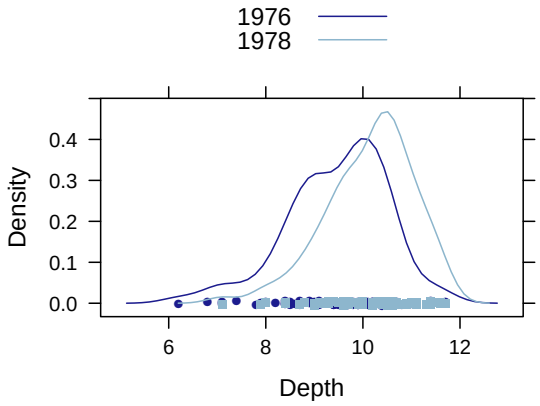
```
> summary(case0201)

      Year          Depth
 Min.   :1976   Min.   : 6.2
 1st Qu.:1976   1st Qu.: 9.1
 Median :1977   Median : 9.9
 Mean   :1977   Mean   : 9.8
 3rd Qu.:1978   3rd Qu.:10.5
 Max.   :1978   Max.   :11.7

> fav = favstats(Depth ~ Year, data=case0201); fav

  Year min  Q1 median   Q3  max  mean    sd  n missing
1 1976 6.2 8.9    9.7 10.2 11.7  9.47 1.035 89       0
2 1978 7.1 9.6   10.3 10.7 11.7 10.14 0.906 89       0
```

A total of 178 subjects are included in the data: 89 are finches that were caught in 1976 and 89 are finches that were caught in 1978. The following figure replicates Display 2.1 on page 30.

```
> bwplot(Year ~ Depth, data=case0201)
```




```
> densityplot(~ Depth, groups=Year, auto.key=TRUE, data=case0201)
```



The distributions are approximately normally distributed, with some evidence for a long left tail.

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 41, Display 2.8).

```
> # Calculate Pooled SD
> n1 = fav["1976", "n"]; n1

[1] NA

> n2 = fav["1978", "n"]; n2

[1] NA
```

```
> s1 = fav["1976", "sd"]; s1

[1] NA

> s2 = fav["1978", "sd"]; s2

[1] NA

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

[1] NA

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

[1] NA
```

So the pooled SD is NA and the standard error is NA.
Based on this information, we can construct a 95% confidence interval (page 43, Display 2.9).

```
> Y1 = fav["1976", "mean"]; Y1

[1] NA

> Y2 = fav["1978", "mean"]; Y2

[1] NA

> Yd = Y2-Y1; Yd

[1] NA

> df = n1+n2-2; df

[1] NA

> qt = qt(0.975, df); qt

[1] NA

> hw = qt*SE; hw

[1] NA

> lower = Yd-hw; lower

[1] NA

> upper = Yd+hw; upper

[1] NA
```

So the 95% confidence interval of the difference between means is (NA, NA)
Now we want to calculate the t-statistic and p-value (as shown on page 46, Display 2.10).

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

[1] NA

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

[1] NA

> twopval = 2*onepval; twopval

[1] NA
```

The one-sided p-value is approximately NA and the two-sided p-value is also approximately NA.
We can get the results of "Summary of Statistical Findings" (page 29) by using the following code:

```
> t.test(Depth ~ Year, var.equal=TRUE, data=case0201)


Two Sample t-test

data:  Depth by Year
t = -5, df = 200, p-value = 9e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.956 -0.381
sample estimates:
mean in group 1976 mean in group 1978
              9.47               10.14

> confint(lm(Depth ~ Year, data=case0201))

               2.5 %   97.5 %
(Intercept) -935.61 -366.488
Year           0.19    0.478
```

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)

   Unaffected      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
```

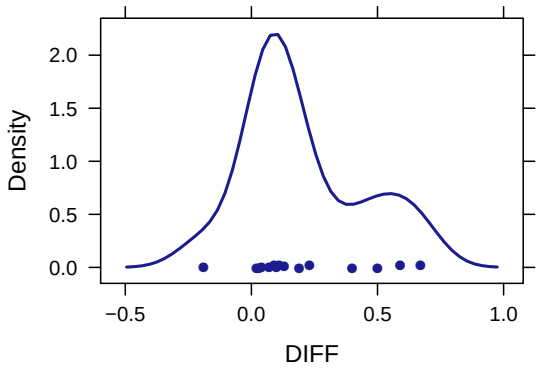
A total of 15 subjects are included in the data. There are 15 pairs of twins; one of the twins has schizophrenia, and the other does not. So there are 15 affected subjects and 15 unaffected subjects.

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

```
> case0202 = transform(case0202, DIFF = Unaffected - Affected)
> favstats(~ DIFF, data=case0202)

   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 31, Display 2.2.

```
> densityplot(~ DIFF, data=case0202)
```



3.2 Inferential procedures (two-sample t-test)

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

```
> # Calculate t-statistics
> difmean = mean(~ DIFF, data=case0202); difmean

[1] 0.199
```

```
> difsd = sd(~ DIFF, data=case0202); difsd

[1] 0.238

> difn = nrow(case0202); difn

[1] 15

> difSE = difsd/sqrt(difn); difSE

[1] 0.0615

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

[1] 3.23

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

[1] 0.00606

> # Construct confidence interval
> tstar = qt(0.975, difn-1); tstar

[1] 2.14

> schizolower = difmean - tstar*difSE; schizolower

[1] 0.0667

> schizoupper = difmean + tstar*difSE; schizoupper

[1] 0.331
```

So the two-sided p-value is approximately 0.006 and the 95% confidence interval is (0.07, 0.33). We can also get the results displayed on page 32 by conducting a paired t-test:

```
> with(case0202, t.test(Unaffected, Affected, paired=TRUE))

Warning in sub("^x$", deparse(x_lazy$expr), res$data.name):  argument 'replacement'
has length > 1 and only the first element will be used
Warning in sub("^x and y$", paste(deparse(x_lazy$expr), "and", deparse(y_lazy$expr)),
:  argument 'replacement' has length > 1 and only the first element will be used


Paired t-test

data:  c(1.94, 1.44, 1.56, 1.58, 2.06, 1.66, 1.75, 1.77, 1.78, 1.92,  and c(1.27, 1.63, 1.47,
```

```
t = 3, df = 10, p-value = 0.006
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
```