Introduction to the Practice of Statistics using R: Chapter 12

Nicholas J. Horton* Ben Baumer

April 10, 2013

Contents

Introduction

This document is intended to help describe how to undertake analyses introduced as examples in the Sixth Edition of *Introduction to the Practice of Statistics* (2002) by David Moore, George McCabe and Bruce Craig. More information about the book can be found at `http://bcs.whfreeman.com/ips6e/`. This file as well as the associated `knitr` reproducible analysis source file can be found at `http://www.math.smith.edu/~nhorton/ips6e`.

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`).

Additional examples of fitting multiple regression models can be found in the companion site which implements the examples within *The Statistical Sleuth* in R (`http://www.math.smith.edu/~nhorton/sleuth`).

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

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

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

The # character is a comment in R, and all text after that on the current line is ignored.
Once the package is installed (one time only), it can be loaded by running the command:

```
> require(mosaic)
> require(gmodels)
```

This needs to be done once per session.
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)
```

The specific goal of this document is to demonstrate how to replicate the analysis described in Chapter 12: One-Way Analysis of Variance.

1 Inference for One-way ANOVA

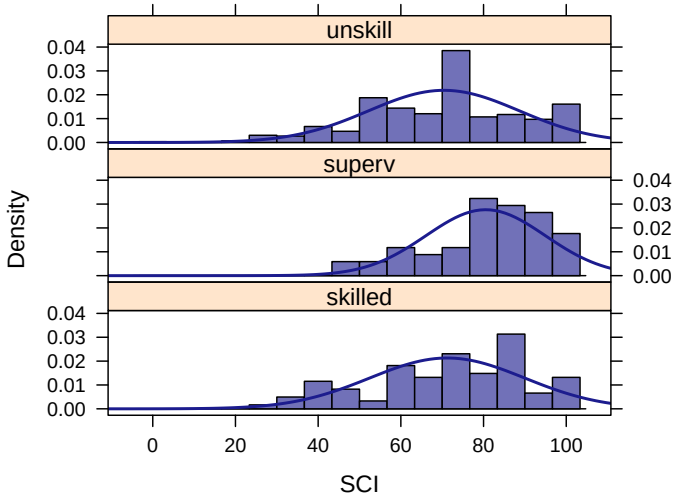
1.1 Exploratory analysis

We consider the case study on workplace safety introduced on page 641 (Example 12.3).

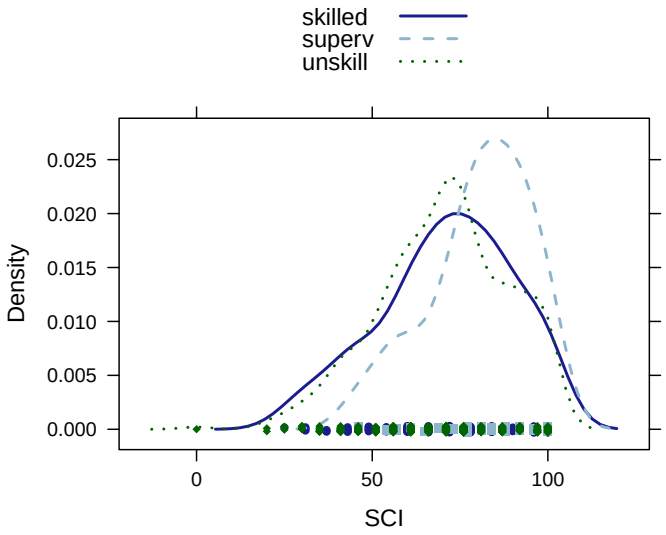
```
> ds = read.csv("http://www.math.smith.edu/ips6e/Ch12/ex12_003.csv")
> favstats(SCI ~ jobcat, data=ds)

         min Q1 median   Q3 max mean   sd   n missing
skilled   25 60     72 84.0 100 71.2 18.8  91       0
superv    46 73     81 92.0 100 80.5 14.6  51       0
unskill    0 61     71 82.5 100 70.4 18.3 448       0
```

Variants of the graphical displays (from Figure 12.3, page 642) are reproduced below. Note that the histograms (with overlaid normal curve) can be generated using separate stacked figures, or a single display can be created using overlapping density plots.

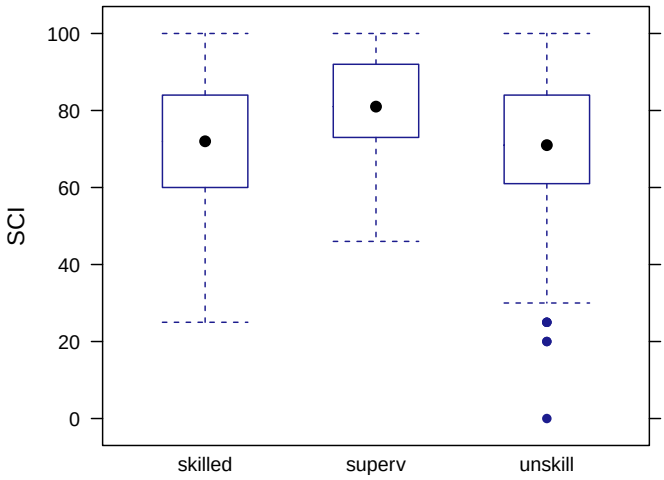
```
> xhistogram(~ SCI| jobcat, fit="normal", layout=c(1, 3), data=ds)

Loading required package:  MASS
```




```
> densityplot(~ SCI, groups=jobcat, auto.key=TRUE, data=ds)
```




```
> bwplot(SCI ~ jobcat, data=ds)
```



1.2 Pooled standard deviation

The pooled standard deviation can be easily calculated through the `lm()` command:

```
> ex12.5 = lm(SCI ~ jobcat, data=ds)
> summary(ex12.5)


Call:
lm(formula = SCI ~ jobcat, data = ds)

Residuals:
   Min     1Q Median     3Q    Max
-70.42 -11.21   0.58  12.79  29.58

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    71.209      1.895   37.59   <2e-16 ***
jobcatsuperv    9.301      3.161    2.94   0.0034 **
jobcatunskill  -0.785      2.078   -0.38   0.7059
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.1 on 587 degrees of freedom
Multiple R-squared: 0.0237,Adjusted R-squared: 0.0204
F-statistic: 7.14 on 2 and 587 DF,  p-value: 0.000866
```

The value of 18.07 matches the results in Example 12.5 (page 647).

1.3 ANOVA table

The ANOVA table (Figure 12.8, page 649) can be generated from this linear model object.

```
> anova(ex12.5)

Analysis of Variance Table

Response: SCI
           Df Sum Sq Mean Sq F value  Pr(>F)
jobcat      2   4662    2331    7.14 0.00087 ***
Residuals 587 191729     327
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.4 Decomposition

As always, the total variability (SST) can be decomposed into part explained by the model (SSM or SSG, as described in Example 12.9, page 651) and part unexplained (SSE, or sums of squares for error).

```
> meanval = mean(~ SCI, data=ds)
> SST = with(ds, sum((SCI - meanval)^2))
> SST

[1] 196391

> SSM = sum((fitted(ex12.5) - meanval)^2)
> SSM

[1] 4662

> SSE = sum((residuals(ex12.5)^2))
> SSE

[1] 191729

> SSM + SSE

[1] 196391
```

We can use these results to verify the value of s_p (the pooled estimate of the parameter σ) from our model.

```
> MSE = SSE / (nrow(ds) - 2 - 1); MSE

[1] 327
```

```
> sqrt(MSE)

[1] 18.1
```

This matches the value on page 652 (Example 12.11).

1.5 The F test

The F distribution is used to test the overall hypotheses (and other multiple degree of freedom tests). The p-value is the probability that a random variable having the $F(I-1, N-I)$ distribution is greater or equal to the calculated value of the F statistic. The values from Example 12.12 (page 653) can be found using the `qf()` function:

```
> qf(c(.90, .95, .975, .99, .999), df1 = 2, df2 = 587)

[1] 2.31 3.01 3.71 4.64 6.99
```

(Note that the values in the book are only available for denominator degrees of freedom equal to 200, so the results are conservative).

1.6 Coefficient of determination

As usual, the R^2 (or coefficient of determination) can be calculated in multiple ways:

```
> r.squared(ex12.5)

[1] 0.0237

> SSM/SST

[1] 0.0237
```

2 Comparing the means

2.1 Contrasts

Contrasts can be used to calculate specific one degree of freedom tests of hypotheses. Recall the means from the worker data:

```
> mean(SCI ~ jobcat, data=ds)

skilled  superv unskill
   71.2    80.5    70.4
```

We can also calculate these in terms of the regression parameter estimates:

```
> mycoef = coef(ex12.5); mycoef

  (Intercept)  jobcatsuperv jobcatunskill
       71.209         9.301        -0.785

> mycoef[1]

(Intercept)
       71.2

> mycoef[1] + mycoef[2]

(Intercept)
       80.5

> mycoef[1] + mycoef[3]

(Intercept)
       70.4
```

Contrasts can be fit using the `fit.contrast()` function within the `gmodels` package.

```
> require(gmodels)

Loading required package: gmodels

> fit.contrast(ex12.5, "jobcat", c(-1/2, 1, -1/2))

                         Estimate Std. Error t value Pr(>|t|)
jobcat c=( -0.5 1 -0.5 )     9.69       2.74    3.54 0.000427
```

This matches the results for the first contrast (Example 12.18, pages 658–659).
A similar process is used to test the second contrast (Example 12.20, page 659):

```
> fit.contrast(ex12.5, "jobcat", c(-1, 0, 1))

                     Estimate Std. Error t value Pr(>|t|)
jobcat c=( -1 0 1 )   -0.785       2.08  -0.378    0.706
```

These values can be used to calculate a 95% confidence interval for the difference in means:

```
> -0.79 + c(-1, 1)*qt(.975, df=587) * 2.08

[1] -4.88  3.30
```

2.2 Multiple comparisons

A number of packages support the comparison of multiple tests using R (see for example the `multcomp` package).

2.3 Power

The `power.anova.test()` function can be used to calculate power and sample size for a one-way ANOVA. For the power of a reading comprehension study (Example 12.27, pages 668-669), this yields power of approximately 35%.

```
> power.anova.test(groups=3, n=10, within.var=7^2, between.var=var(c(41, 47, 44)))


     Balanced one-way analysis of variance power calculation

         groups = 3
              n = 10
    between.var = 9
     within.var = 49
      sig.level = 0.05
          power = 0.349

 NOTE: n is number in each group
```