SDM4 in R: Analysis of Variance (Chapter 26)

Nicholas Horton (nhorton@amherst.edu)

August 12, 2017

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fourth Edition of *Stats: Data and Models* (2014) by De Veaux, Velleman, and Bock. More information about the book can be found at http://wps.aw.com/aw_deveaux_stats_series. This file as well as the associated R Markdown reproducible analysis source file used to create it can be found at http://nhorton.people.amherst.edu/sdm4.

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 vignettes (http://cran.r-project.org/web/packages/mosaic). A paper describing the mosaic approach was published in the *R Journal*: https://journal.r-project.org/archive/2017/RJ-2017-024.

Chapter 26: Analysis of variance

### Section 26.1: Testing whether the means of several groups are zero

The graph in Figure 26.1 (page 747) can be generated using the `bwplot()` function.

```
Soap <- read.csv("http://nhorton.people.amherst.edu/sdm4/data/Bacterial_Soap.csv")
bwplot(Bacterial.Counts ~ Method, data=Soap)
```
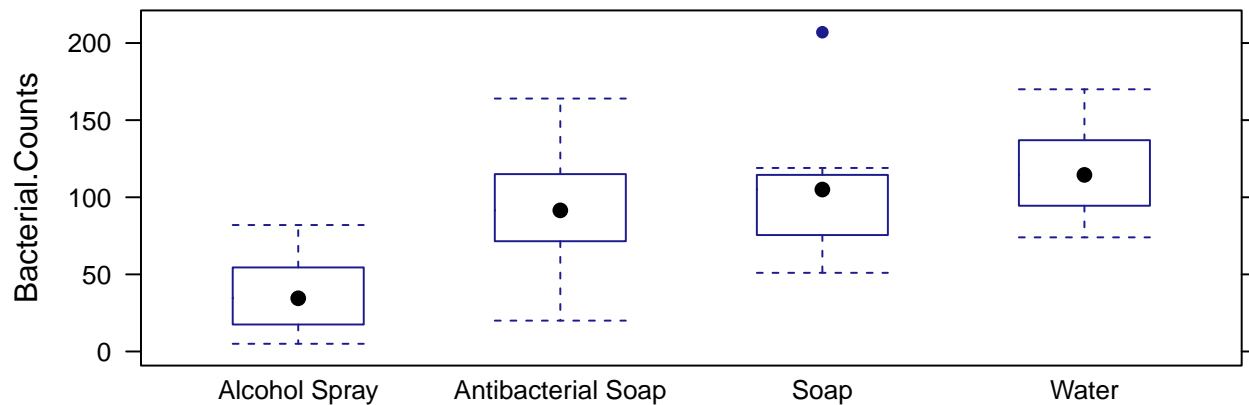

The example on page 750 considers the outcomes in hand volumes for three treatments post surgery.

```
Contrast <- read.csv("http://nhorton.people.amherst.edu/sdm4/data/Contrast_baths.csv")
bwplot(Hand.Vol.Chg ~ Treatment, data=Contrast)
```
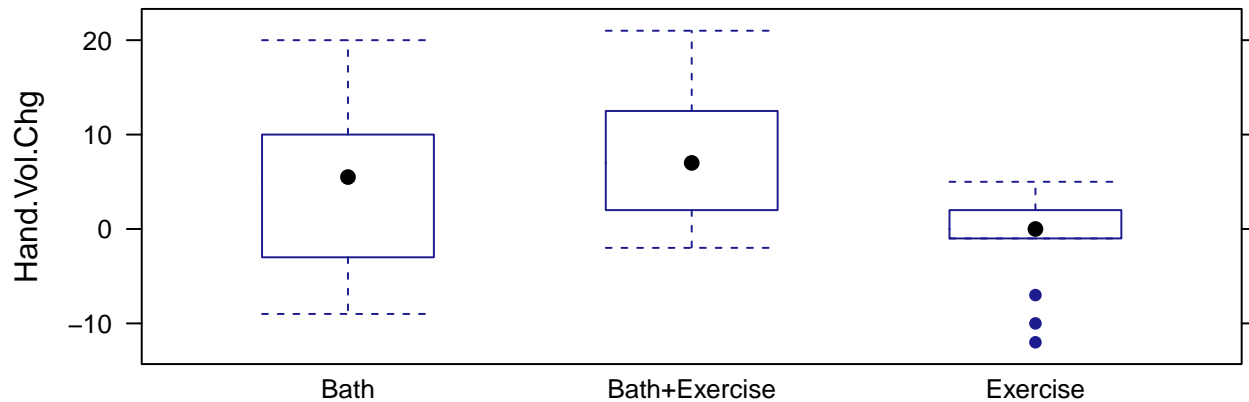
The summary statistics at the bottom of page 751 can be calculated using `favstats()`.

```
favstats(Bacterial.Counts ~ Method, data=Soap)
```

```
##               Method min    Q1 median     Q3 max  mean     sd n missing
## 1      Alcohol Spray   5 17.75   34.5  52.75  82  37.5 26.560 8       0
## 2 Antibacterial Soap  20 72.25   91.5 113.00 164  92.5 41.963 8       0
## 3               Soap  51 79.75  105.0 112.25 207 106.0 46.959 8       0
## 4              Water  74 98.25  114.5 136.00 170 117.0 31.131 8       0
```

**Section 26.2: The ANOVA table**

The `aov()` function can be used to fit an analysis of variance model.

```
aovmod <- aov(Bacterial.Counts ~ Method, data=Soap)
summary(aovmod)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Method       3  29882    9961    7.06 0.0011 **
## Residuals   28  39484    1410
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model has 3 degrees of freedom for the model (numerator) and 28 degrees of freedom for the error (denominator). The `xpf()` function can replicate the calculation of the exact p-value (and generate Figure 26.4, page 754).

```
xpf(7.0636, df1=3, df2=28)
```

2
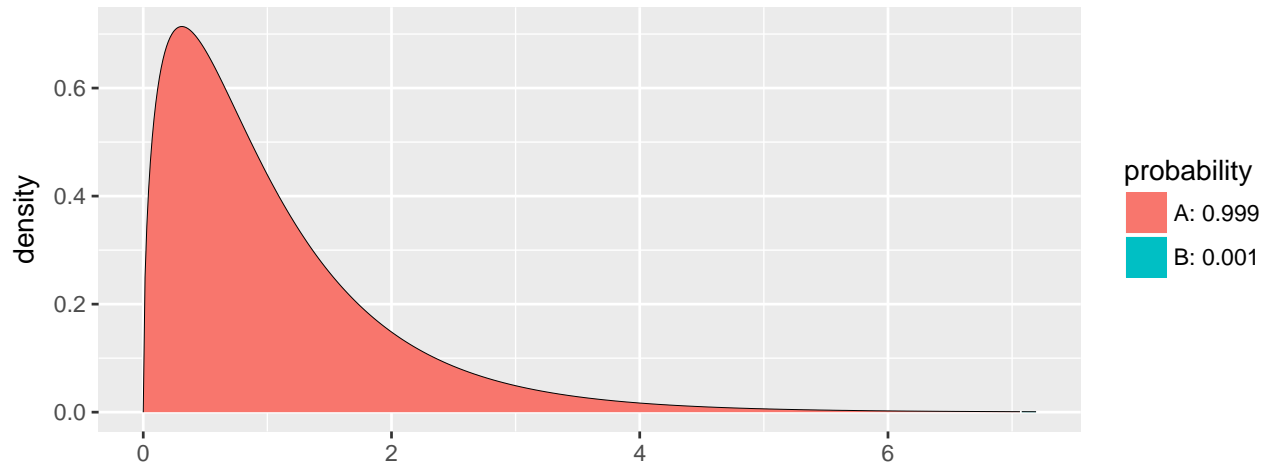
```
## [1] 0.99889
```

The treatment means can be generated using `model.tables()` (see page 757).

```r
model.tables(aovmod)
```

```
## Tables of effects
##
##  Method
## Method
##      Alcohol Spray Antibacterial Soap          Soap
##            -50.75                 4.25         17.75
##            Water
##            28.75
```

The residual standard deviation can be calculated (page 759).

```r
n <- 32; k <- 4
sp <- sqrt(sum(resid(aovmod)^2/(n-k))); sp
```

```
## [1] 37.552
```

```r
sqrt(1410)
```

```
## [1] 37.55
```

We can also see how the results are equivalent when fitting a regression model with indicators.

```r
lmmod <- lm(Bacterial.Counts ~ Method, data=Soap)
msummary(lmmod)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  37.5       13.3    2.82  0.00863 **
## MethodAntibacterial Soap     55.0       18.8    2.93  0.00669 **
## MethodSoap                   68.5       18.8    3.65  0.00107 **
```

```
## MethodWater                   79.5      18.8   4.23  0.00022 ***
##
## Residual standard error: 37.6 on 28 degrees of freedom
## Multiple R-squared:  0.431,  Adjusted R-squared:  0.37
## F-statistic: 7.06 on 3 and 28 DF,  p-value: 0.00111
```