Basic Statistics and Graphics II

C.-Y. Kuo, T. Eiting, M. Rosario

10.05.2012

1 Introduction

1.1 Welcome

During our meeting last week, we demonstrated how to do descriptive stats, check for normality and perform two-group comparisons. This week is a continuation of the previous week, in which we will show how to perform 1- and 2-factor analysis of variance (ANOVA), regression, correlation and associated graphics in R.

1.2 Getting started

Again, there are a few things that we need to do before doing any stats. If you have been participating in our meetings, you must have got very used to them now.

```
# some prep work
rm(list = ls())
setwd("D:/R")

# read-in data file, check it is done correctly and attach the
# dataset
cal <- read.csv("calcium.csv", header = T)
head(cal)

##   calcium gender hormone height
## 1    16.5      f       n   1.60
## 2    18.4      f       n   1.71
## 3    12.7      f       n   1.42
## 4    14.0      f       n   1.66
## 5    12.8      f       n   1.76
## 6    14.5      m       n   1.73

attach(cal)
```

2 Basic statistics and graphics

The dataset we just imported is the concentrations of calcium ions in the blood in 10 male and 10 female subjects that received either hormone treatment (treatment "y") or placebo (treatment "n"). We would like to find out about a few things:

1. Whether males and females differ in blood calcium concentration.
2. Whether hormone treatment increases blood calcium concentration.
3. Whether there is an interaction between gender and treatment. In other words, do males and females react to hormone treatment differently.

Analysis of variance (ANOVA) is an excellent statistical tool to answer the above questions. There are many variations of ANOVA, each of which is suitable for a particular kind of experimental design. Again, this is beyond the scope of this meeting, and we will only introduce two of the simpler types of ANOVAs below.

2.1 One-factor ANOVA

When there is only one factor of interest (e.g., pretend for now that treatment didn't exist in the calcium dataset), we can use a one-factor (or one-way) ANOVA to see if the means values differ between groups that vary with regard to the factor in question. In our case here, we can use one-way ANOVA to test whether males and females have different blood calcium concentration.

```
# one-factor ANOVA with gender as the factor of interest
anova(lm(calcium ~ gender))

## Analysis of Variance Table
##
## Response: calcium
##           Df Sum Sq Mean Sq F value Pr(>F)
## gender     1     70    70.3    0.72   0.41
## Residuals 18   1757    97.6
```

A p-value greater than 0.05 means that we cannot reject the null hypothesis that there's no difference in blood calcium between males and females. Therefore, blood calcium concentration is likely not different between genders.

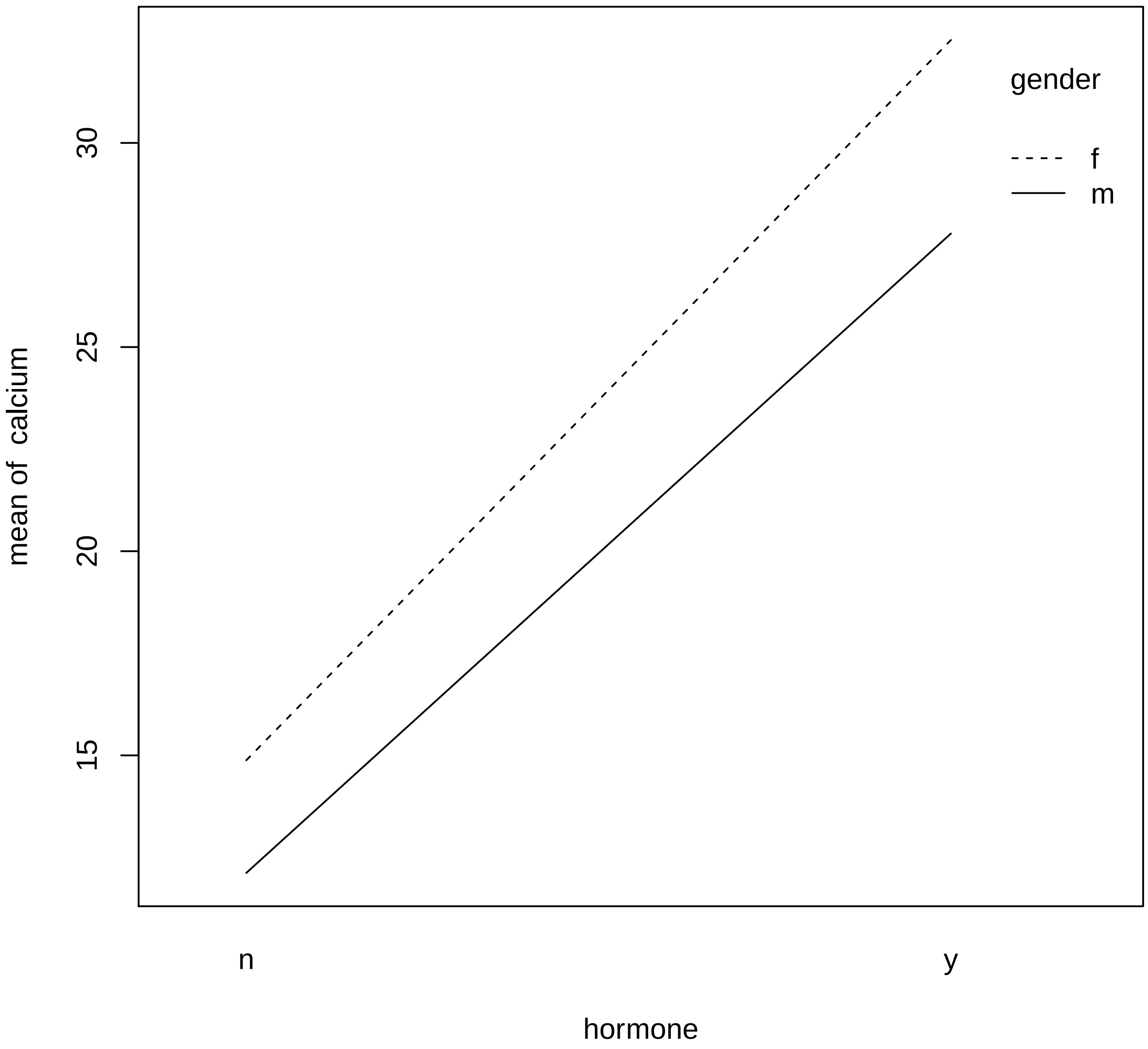
2.2 Two-factor ANOVA

Now back to reality: there are acutally two factors in the calcium dataset: gender and treatment. To fully address the three questions above, we need to use a two-factor ANOVA.

```
# two-factor ANOVA with gender and hormone as factors and their
# interaction
anova(lm(calcium ~ gender * hormone))

## Analysis of Variance Table
##
## Response: calcium
##                Df Sum Sq Mean Sq F value  Pr(>F)
## gender          1     70      70    3.07   0.099 .
## hormone         1   1386    1386   60.53 7.9e-07 ***
## gender:hormone  1      5       5    0.21   0.650
## Residuals      16    366      23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results tell us that there's no difference in blood calcium between males and females, although the p-value for gender becomes much lower (from 0.4 to 0.1) because we better partitioned the sources of variation in our ANOVA model. However, there is a significant difference between subjects with different hormone treatments. The line gender:hormone in the output tells us that there is no interaction between gender and treatment; males and females react the same way to hormone treatments. A good graphic way to help visualize the above results is an interaction plot.

```
# make an interaction plot
interaction.plot(hormone, gender, calcium)
```



(Just as a sidenote: in some experimental design we can be sure that there is no interaction between the two factors. In that case, the R code will be slightly different.)

```
# two-factor ANOVA without interaction
anova(lm(calcium ~ gender + hormone))

## Analysis of Variance Table
##
## Response: calcium
##           Df Sum Sq Mean Sq F value  Pr(>F)
## gender     1     70      70    3.22   0.091 .
## hormone    1   1386    1386   63.47 3.9e-07 ***
## Residuals 17    371      22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.3 Graphical presentation of means between groups: barplot

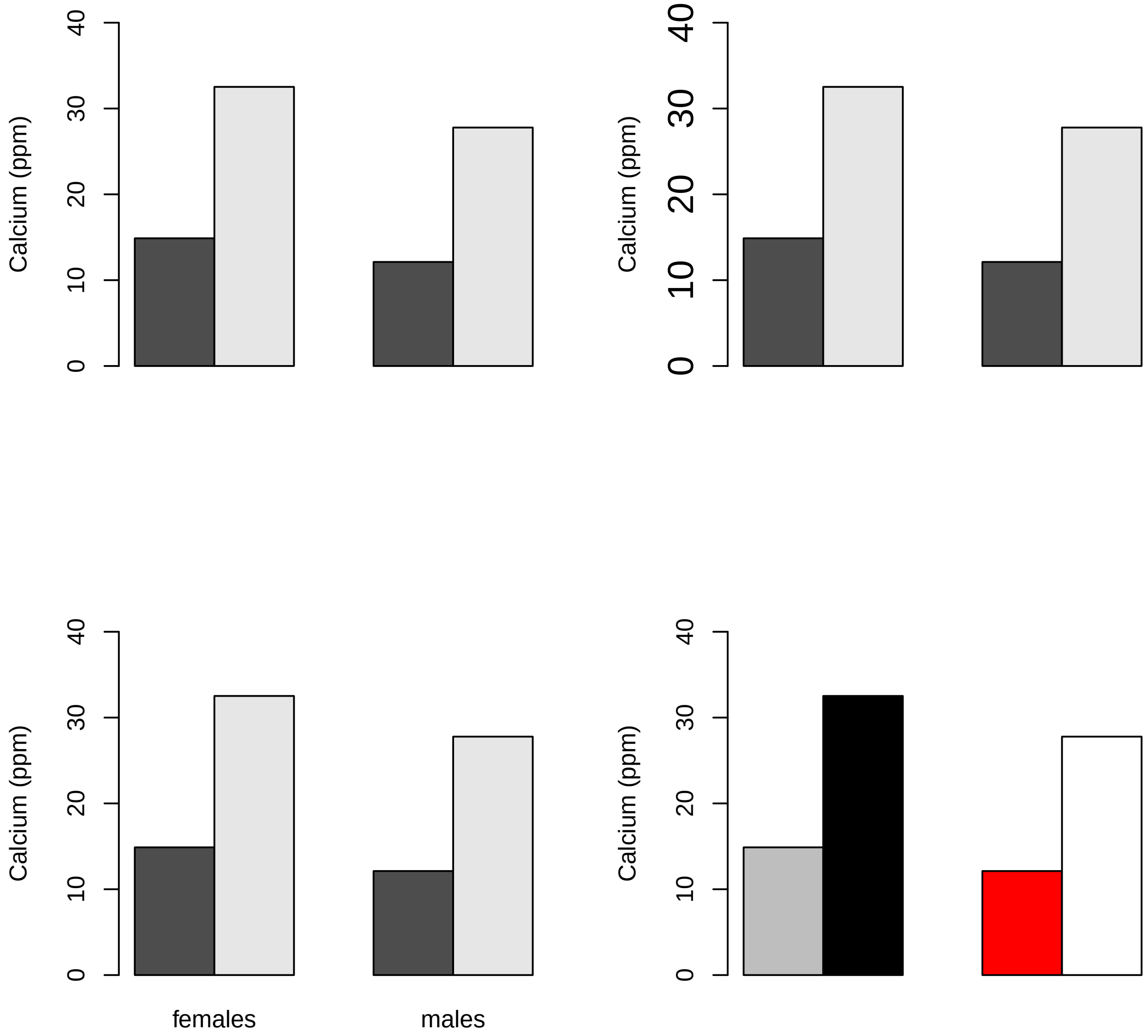
After ANOVA, we want to make a figure showing the mean blood calcium concentration of the following 4 groups: female w/o hormone, female w/ hormone, male w/o hormone and male w/o hormone. Making a barplot in R is easy, but the preparatory work can be a bit tedious at times. First of all, we can't just use our original dataset. Instead, we have to create a mini-dataset for making a barplot.

```
# create a matrix consists of the 4 means (gender by treatment)
cal.bar <- matrix(c(14.88, 32.52, 12.12, 27.78), nrow = 2, byrow = F)
cal.bar

##       [,1]  [,2]
## [1,] 14.88 12.12
## [2,] 32.52 27.78
```

There is a very important point. The values that we would like to appear right next to each other in the barplot need to be in the same column!

After making this mini-dataset, we can go ahead and make the barplot

```
# make a barplot; the 'beside' option specifies whether bars of the
# same group should be stacked; the 'ylab' option tells R what to
# use as the label for the y-axis; the 'ylim' option specifies the
# lower and upper limit of the y-axis
barplot(cal.bar, beside = T, ylab = "Calcium (ppm)", ylim = c(0, 40))
# there are some other useful options in the barplot function.
par(mfrow = c(2, 2))
# the original barplot
barplot(cal.bar, beside = T, ylab = "Calcium (ppm)", ylim = c(0, 40))
#'cex.axis' option allows us to adjust the size of the numeric axis label
barplot(cal.bar, beside = T, ylab = "Calcium (ppm)", ylim = c(0, 40),
    cex.axis = 1.5)
#'names.arg' option allows us to put names under each group/bar
barplot(cal.bar, beside = T, ylab = "Calcium (ppm)", ylim = c(0, 40),
    names.arg = c("females", "males"))
#'col' option allows us to choose the color for each bar
barplot(cal.bar, beside = T, ylab = "Calcium (ppm)", ylim = c(0, 40),
    col = c("grey", "black", "red", "white"))
```



When presenting data like this, we almost always want have error bars in addition to the means. Unfortunately, there is no built-in option in barplot function to add error bars. Again, we will have to rely on custom fuctions. The good news is that many external R packages for graphics allow easy addition of error bars, and we will introduce them when we talk about advanced graphics.

2.4 Regression and correlation

Let's go back to the calcium dataset. Pretend again that gender and treatmetnt didn't exist. We would like to know:

1. Whether the values of blood calcium is dependent on height, and
2. How much of the variation in blood calcium can be explained by variation in height.

To do so, we can use a regression analysis with height as the predictor (or independent) variable and blood calcium as the response (or dependent) variable.

```
# simple linear regression
summary(lm(calcium ~ height))

##
## Call:
## lm(formula = calcium ~ height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11.32  -8.40  -2.28   8.68  19.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    34.32      26.47    1.30     0.21
## height         -7.25      15.31   -0.47     0.64
##
## Residual standard error: 10 on 18 degrees of freedom
## Multiple R-squared: 0.0123,Adjusted R-squared: -0.0426
## F-statistic: 0.224 on 1 and 18 DF,  p-value: 0.641
```

The p-value for height is far greater than 0.05, which means that we cannot reject the null hypothesis that the slope in this regression is 0. Therefore, blood calcium does not seem to vary with height. In addition, the variation in height only explains 0.04% of the variation in blood calcium.

If we don't assume that there is a dependent relationship between height and blood calcium, we can simply ask whether the values of blood calcium and height vary together. In this case, we are testing the correlation between blood calcium and height without one being dependent on the other.

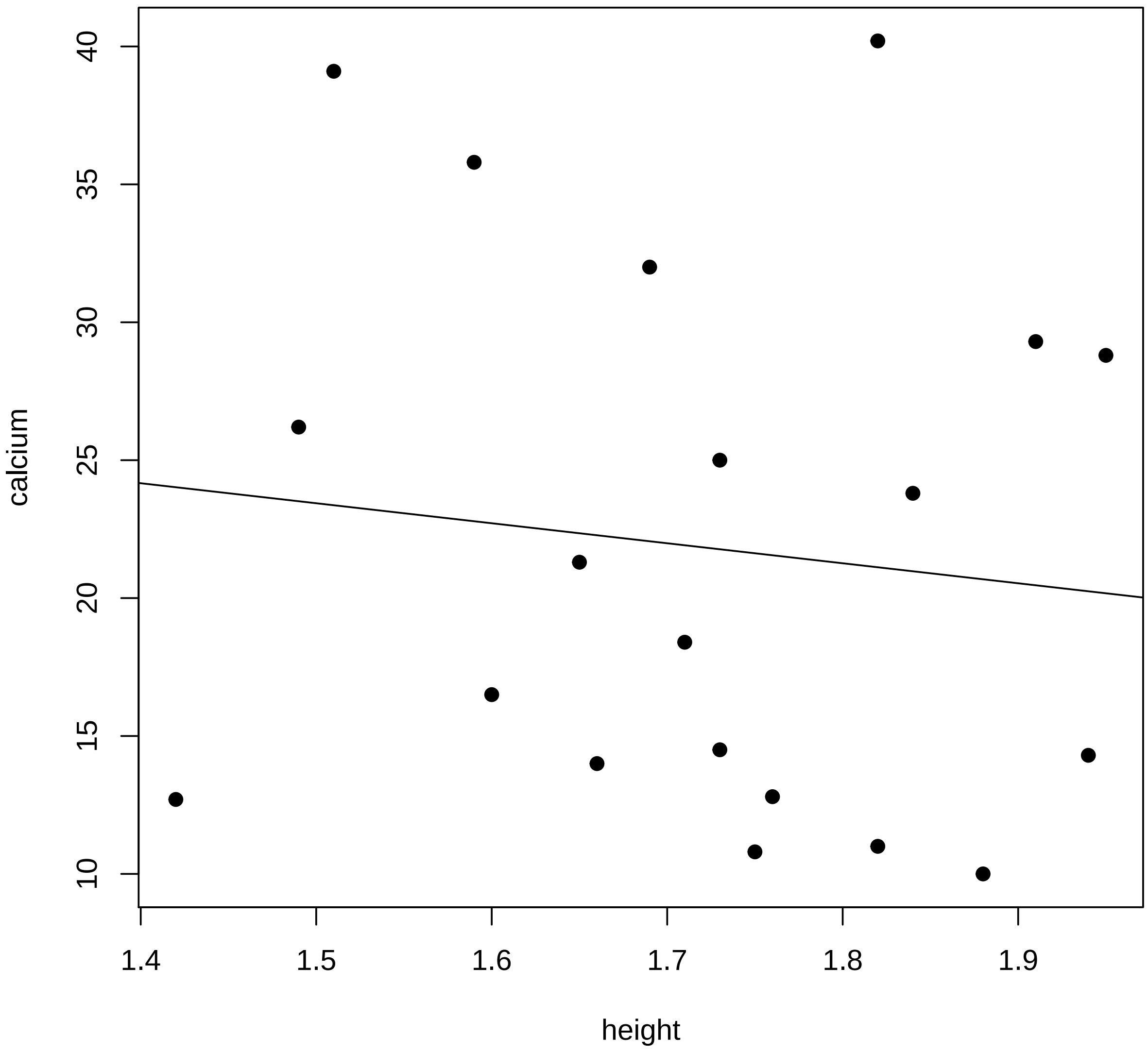
```
# Pearson's simple linear correlation the method option allows us
# to use the parametric Pearson correlation or two other
# nonparametric methods (Spearman and Kendall)
cor.test(calcium, height, method = "pearson")

##
##  Pearson's product-moment correlation
##
## data:  calcium and height
## t = -0.4738, df = 18, p-value = 0.6413
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5276  0.3487
## sample estimates:
##    cor
## -0.111
```

Based on the results (p-value greater than 0.05), although the correlation coefficient is -0.11, it is actually not significant different from 0. We can conclude that there's no relationship between blood calcium and height.

2.5 Graphical presentation for regression: scatter plot

As a graphical presentation, we can plot all the data points on the same graph according to their calcium and height values, with height as the x-axis and calcium the y-axis.

```
# create a scatter plot. The 'pch' option allows us to change the
# looks of the points.
par(mfrow = c(1, 1))
plot(height, calcium, pch = 19)
# add a regression line. The 'lty' option specifies the type of the
# line.
abline(lm(calcium ~ height), lty = 1)
```



3 Exercise

3.1 In-class

1. Play with the options we introduced in barplot, plot and abline functions. Find out what kind of difference you can make to a plot by tweaking those options.

3.2 Take-home

1. Read-in the milipedes dataset

2. Test the following null hypotheses:

- There's no difference in alanine concentration among three species.

- There's no difference in alanine concentration between sexes.

- There's no interaction between species and sex in mean alanine concentration.