Week 8 Problem Based Learning and Practical Solutions

Jeffrey O. Hanson[1]
[1]*School of Biological Sciences, The University of Queensland, Brisbane, QLD, Australia*
Correspondance should be addressed to jeffrey.hanson@uqconnect.edu.au

03 May 2016

Contents

Problem based learning workshop

General notes

Background reading

1. Quinn and Keough Chapter 19
2. Whitlock, M and Schluter D. 2009. The Analysis of Biological Data. Roberts and Company, Chapter 2.

Goals

1. To understand the reporting standards for communicating the results of common statistical tests.
2. To construct tables appropriate for publication in a scientific journal.
3. To understand and apply the key elements of effective visual display of data.
4. To write accurate and effective figure legends appropriate for publication in a scientific journal.

PART 1. Reporting analyses in text

Exercise - Chickadee calls in relation to predator mass

The problem: chikadees are a small bird native to North America. They have a characteristic call when there is an avian predator around. It goes like chik-a-dee-dee-dee. Individuals tend to differ in the number of "dees" they add to the end of the call. One hypothesis for this behaviour is that the number of "dees" is an indicator of the size of the particular avian predator that is around. Researchers collected data on the average number of dees and the body mass of the predator species. The goal of the study was to tell whether predator body mass could explain call variation.

Often statistical analysis packages give you a lot more information than you need. The goal of inline reporting of statistical analyses is to simplify as much as possible but without making the reader have to guess what you actually did.

Activity 1: The following R code was used to analyse the data using linear regression. Read the output and write a few sentences reporting the statistical results.

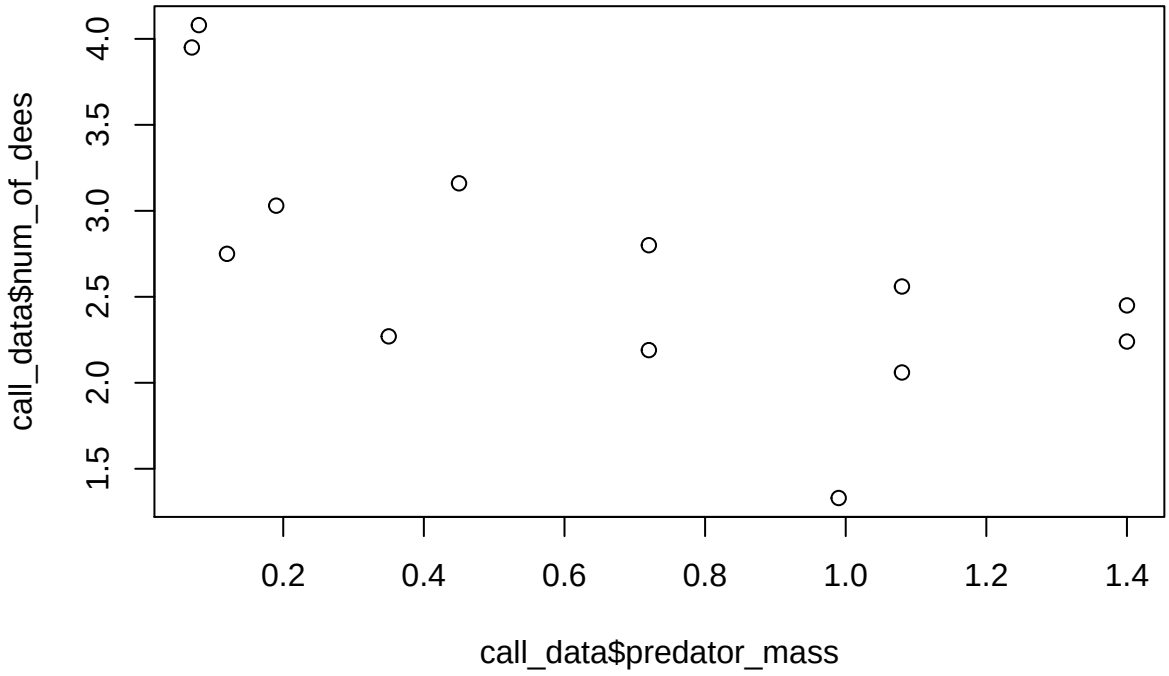
```
# create data
predator_mass <-c(0.07,0.08,0.12,0.19,0.35,0.45,0.72,0.72,1.40,0.99,1.4,1.08,1.08)
num_of_dees <- c(3.95,4.08,2.75,3.03,2.27,3.16,2.19,2.80,2.45,1.33,2.24,2.56,2.06)
call_data <- as.data.frame(cbind(predator_mass,num_of_dees))

# fit model
lmfit <- lm(num_of_dees ~ predator_mass, data=call_data)

# assess significance of relationship
anova(lmfit)
```

```
## Analysis of Variance Table
##
## Response: num_of_dees
##               Df Sum Sq Mean Sq F value  Pr(>F)
## predator_mass  1 3.1268 3.12683  9.3106 0.01102 *
## Residuals     11 3.6942 0.33584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# print model summary
summary(lmfit)
```

```
##
## Call:
## lm(formula = num_of_dees ~ predator_mass, data = call_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0153 -0.4356  0.1744  0.3204  0.7899
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.3731     0.2776  12.149 1.02e-07 ***
## predator_mass  -1.0382     0.3402  -3.051    0.011 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5795 on 11 degrees of freedom
## Multiple R-squared:  0.4584, Adjusted R-squared:  0.4092
## F-statistic: 9.311 on 1 and 11 DF,  p-value: 0.01102
```

```r
# you can also plot the relationship if you like
plot(call_data$predator_mass, call_data$num_of_dees)
```




```r
anova(lmfit)
```

```
## Analysis of Variance Table
##
## Response: num_of_dees
##               Df Sum Sq Mean Sq F value  Pr(>F)
## predator_mass  1 3.1268 3.12683  9.3106 0.01102 *
## Residuals     11 3.6942 0.33584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(lmfit)
```

```
##
## Call:
## lm(formula = num_of_dees ~ predator_mass, data = call_data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.0153 -0.4356  0.1744  0.3204  0.7899
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.3731     0.2776  12.149 1.02e-07 ***
## predator_mass  -1.0382     0.3402  -3.051    0.011 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5795 on 11 degrees of freedom
## Multiple R-squared:  0.4584, Adjusted R-squared:  0.4092
## F-statistic: 9.311 on 1 and 11 DF,  p-value: 0.01102
```

Example answer: Predator mass explained 45% of the variation in the numbers of dees at the end of the alarm calls. The number of dees declined with increasing predator mass (equation: *number of dees* = 3.37 - 1.04 × predator mass, $F_{1,1}$, P=0.01, R^2=0.458).

Want to know more? We won't be taking this specific analysis further here but in case you are interested in the more complete analysis of this story see Whitlock and Schluter Section 17.3

PART 2 Displaying Results in Tables

For the next few exercises we will work on a small dataset from a marine invertebrate. The data are from an experiment designed to test the effects of copper and sperm concentration on fertilisation success in the free-spawning marine invertebrate, *Galeolaria caespitosa*. Raw data are the number of eggs fertilized per 100. *Note*: The Copper concentrations are in units of milligrams per Litre)

Activity 2: Here is a list of means and standard errors from the different treatments in the experiment. Try and create your own table displaying these results. In a few minutes, I will place a table on the monitors. Write down what's wrong with it and then attempt to construct a better version. You don't need a word processor to do this.

Put the data into a table.

```
# load data
data1 <- read.table('data/week-8/data1.csv', sep=',', header=T)
control <- subset(data1, copper_concentration=="0")
copper <- subset(data1, copper_concentration=="50")

#means
mean(control$fertilisation_success)
```

```
## [1] 65.94444
```

```
mean(copper$fertilisation_success)
```

```
## [1] 56.36111
```

```r
# Std errors
sd(control$fertilisation_success)/sqrt(nrow(control))
```

```
## [1] 2.784009
```

```r
sd(copper$fertilisation_success)/sqrt(nrow(copper))
```

```
## [1] 3.667349
```

Example answer:

Table guide lines *Style rules*: no lines in table, column names are bold. *General rules*: need caption, need appropriate/consistent significant figures, show units.

Example table

```r
# make table
curr.DF <- data.frame(
    control=paste0(round(mean(control$fertilisation_success),2),' $\\pm$ ',
        round(sd(control$fertilisation_success)/sqrt(nrow(control)),2)),
    copper=paste0(round(mean(copper$fertilisation_success),2),' $\\pm$ ',
        round(sd(copper$fertilisation_success)/sqrt(nrow(copper)),2))
)

# render table
knitr::kable(
    curr.DF,
    caption='Number of eggs fertilized per 100 (mg per L). Data represent mean $\\pm$ standard
    col.names=c('\\textbf{Control}', '\\textbf{Copper treatment}'),

)
```

Control	**Copper treatment**
65.94 ± 2.78	56.36 ± 3.67

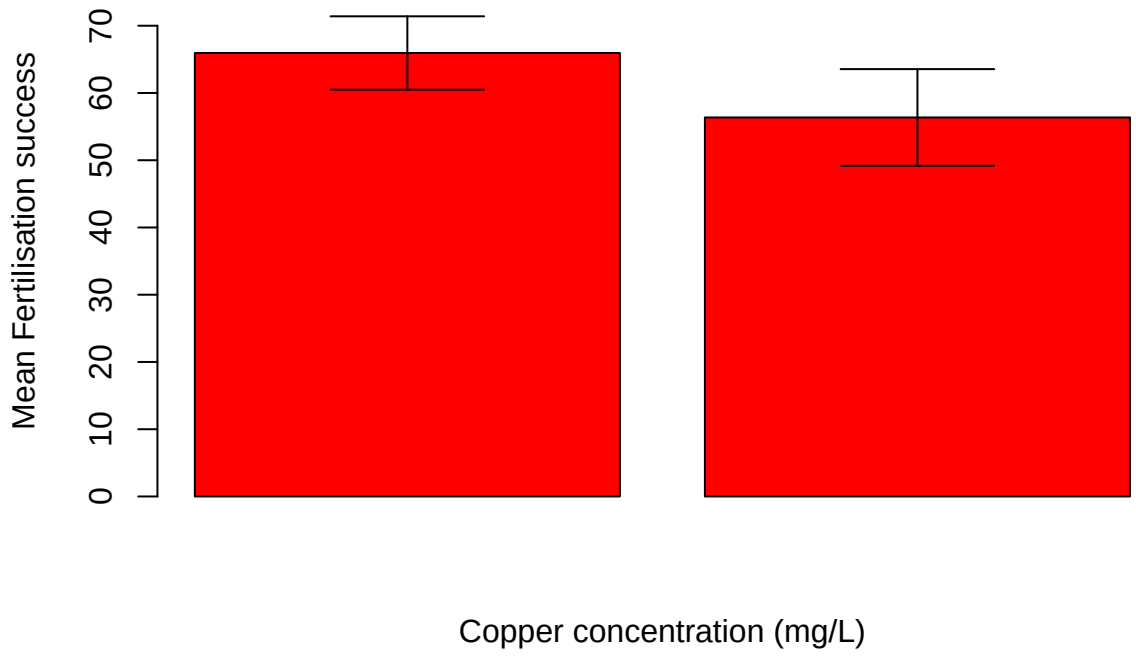
Table 1 Number of eggs fertilized per 100 (mg per L). Data represent mean ± standard errors.

PART 3 Graphical representation of data

Activity 4: Create a bar chart by pasting this code into R

```r
# Create a bar graph
groups <- data.frame(group=c("0 mg/L", "50 mg/L"))
groups$means =  c(  65.94444 , 56.36111    )
```

```r
groups$se = c(  2.784009, 3.667349)
attach(groups)

# Create a bar chart
myplot <- barplot(groups$means, xlab="Copper concentration (mg/L)", ylab="Mean Fertilisation su

# add 95% Confidence Intervals
arrows(myplot, means-1.96*se, myplot, means+1.96*se, length=0.4, angle=90, code=3)
```



Question: Is there a strong difference between your treatments? Explain. (Remember: error bars and confidence intervals help us assess whether the two means are very different or not).

Critical issue: Overlap of confidence intervals ? what it can and cannot tell you (I will review on the monitors)

Example answer: Difficult to tell confidence intervals are larger than standard errors. A statistical test is required.

Challenge: Using R studio, try and perform a statistical test for a difference between the copper treatments. You need to download and import the raw data int R to do this. *Ask your tutor to help with the data import if necessary.*

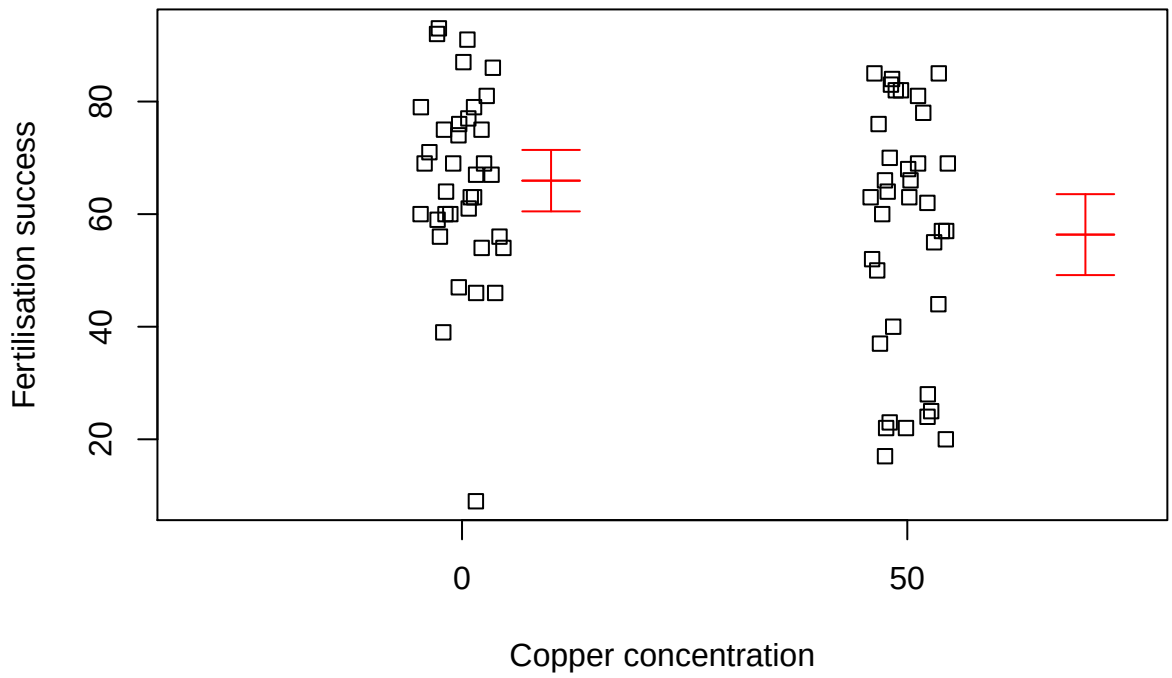
```r
# perform a t-test
t.test(fertilisation_success ~ factor(copper_concentration), data=data1)
```

```
##
##  Welch Two Sample t-test
##
## data:  fertilisation_success by factor(copper_concentration)
## t = 2.0814, df = 65.283, p-value = 0.04133
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.3885416 18.7781250
## sample estimates:
##  mean in group 0 mean in group 50
##         65.94444         56.36111
```

Rule for Bar Graphs: The y-axis scale must always include zero.

Activity 5: The code below presents the same information as in your first barplot but in a different way. Import the raw data file called *data1.csv* and execute this code in RStudio.

```
data1 <- read.table('data/week-8/data1.csv', sep=',', header=T)

s_low <- subset(data1, copper_concentration=="0")
s_high <- subset(data1, copper_concentration=="50")

means <- c(mean(s_low$fertilisation_success),
    mean(s_high$fertilisation_success))
se <- c(sd(s_low$fertilisation_success)/sqrt(nrow(s_low)),
    sd(s_high$fertilisation_success)/sqrt(nrow(s_high)))
stripchart(data1$fertilisation_success ~data1$copper_concentration,
    method="jitter",vertical=TRUE, ylab="Fertilisation success",
    xlab="Copper concentration",xlim=c(0.4,2.5))

arrows(myplot+0.5, means-1.96*se, myplot+0.5, means, length=0.15,
    angle=90, code=3,col="red")
arrows(myplot+0.5, means, myplot+0.5, means+1.96*se, length=0.15,
    angle=90, code=3,col="red")
```



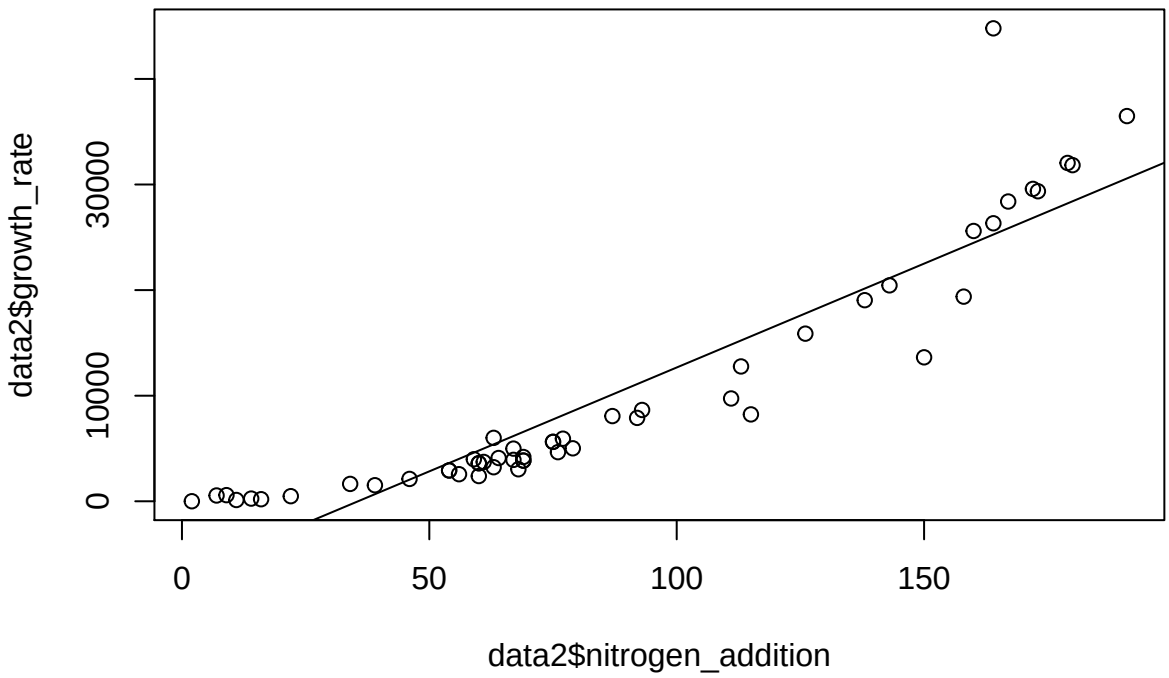
Question: Do you think this is an improvement over the bar chart you created earlier? If so, in what ways?. *Example answer*: This plot shows the range and distribution of the data.

Scatter plots

These are used to examine the relationship between two continuous variables.

Open the data file `data2.csv` which is a collection of plant growth rate measurements grown at different nitrogen levels. Produce a scatterplot of ?nitrogen addition? (x-axis) versus ?growth rate? (y-axis).

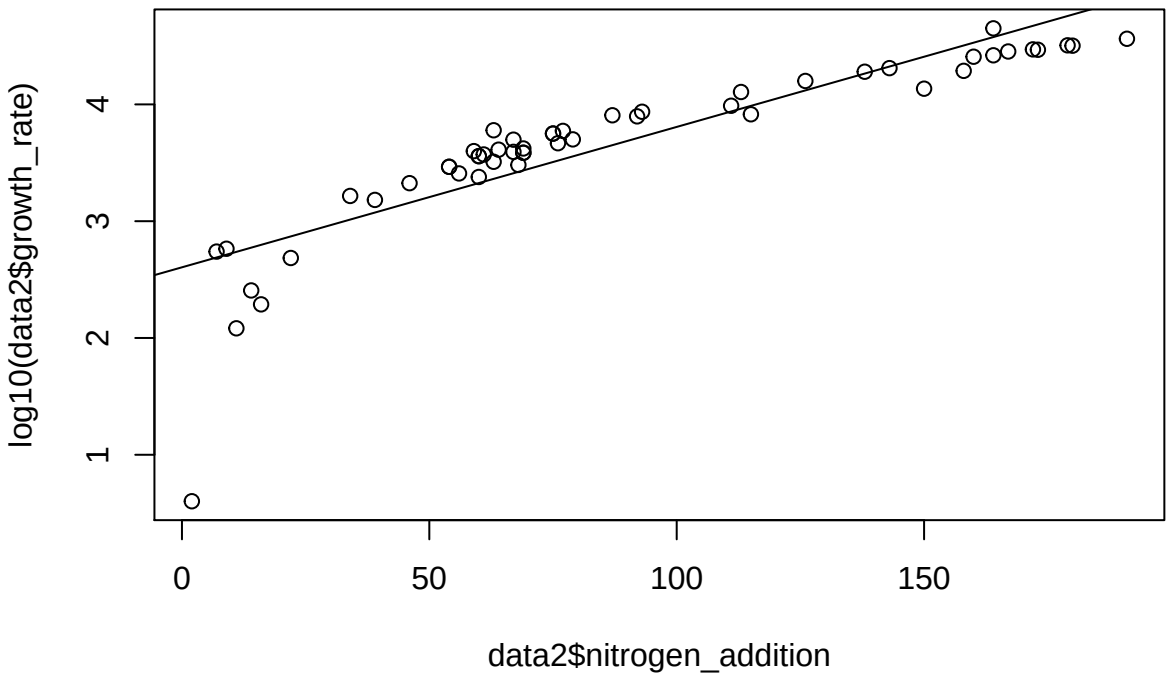
```
# Load the data OR use the import button in RStudio
data2 <- read.table('data/week-8/data2.csv', sep=',' , header=T)

# Make the scatter plot
plot(data2$nitrogen_addition, data2$growth_rate)

# Add a linear trendline to the figure.
regression <- lm(data2$growth_rate ~ data2$nitrogen_addition)
abline(regression)
```



Some Questions

1. Is there a relationship between nitrogen addition and growth rates? Is it linear? Yes there seems to be a relationship. But its not linear.
2. What effect does nitrogen have on growth rate? A positive effect.
3. What would you expect to happen if 500 units of nitrogen were added?
4. Try log transforming the data and then re-examining the relationship. Discuss the effect of the transformation.

```
plot(data2$nitrogen_addition, log10(data2$growth_rate))
regression <- lm(log10(data2$growth_rate) ~ data2$nitrogen_addition)
abline(regression)
```



PART 4: Figure captions

Figures must be able to be interpreted independently of the main text of a paper or report. This means that the figure captions must describe what is in the figure as well what data is actually being presented here. This is one of the major things that people get wrong throughout their student and professional lives. During the PBL, I will supply an excerpt of a study, and a figure from that study.

Activity 7: Write a figure caption for the figure and text supplied on the monitors.

Example answer: The effect of ambient temperature on the walking speed of male and female *Drosophila melanogaster* adults. Shown are mean ± 95% confidence intervals walking speeds of n=80 flies subject to two different temperature treatments.

Remember: The job of a figure caption is not to give you all the results, but to simply describe what?s in the figure. The in-text results section usually provides the details of specific tests.

Activity 8: Try again using the chart you just created in *Activity 5*.