Week 2 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

21 March 2016

Contents

Problem based learning workshop

Sample size, errors, and power

- *How does sample size affect your type 1 error?*
  - It doesn't.
- *How does sample size affect your type 2 error?*
  - Increases in sample size reduce type 2 error.
- *Discuss some key elements that will maximise the power of your experiment*
  - Effect size - the bigger the expected difference the more power you have.
  - Sample size - the more replicates you have; the more power you have.
- *If you had limited resources (ie. small N), could you still increase the power of your experiment?*
  - Yes, select treatments that will maximimise the expected effect size.

Sample size, errors, and power

Activity 1

- *Write down the equation that describes the least square regression for the data in example 1 below.*
  - ntc = 54.382 + (cpd * 7.977)
- *Why are there two t-tests in the table above? How are the t-values calculated? Calculate them and their associated p-values.*

  - The t-tests are testing if the model coefficients are significantly different to zero. The first test is for the intercept and the second for the slope. The t-values are calculated by dividing the term estimate ("Estimate") by the the uncerainty ("Std. Error) surrounding this estimate. The t-values are then converted to p-values using the table in the manual. Alternatively, using the R code `2 * pt(abs(x), df=y, lower=FALSE)` where `x` is the t-statistic and `y` is the residual degrees of freedom in the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	54.382	42.659	1.275	0.214
cpd	7.977	1.807	4.414	0.0001

  -

- *Calculate the F value for the ANOVA table above and its associated p-value?*

  - For each term (row in the table), the mean sums of squares ("Mean Sq") are calculated by dividing the sum of squares ("Sum Sq") by the degrees of freedom ("Df"). The F-statistic ("F value") is then calculated by diving the mean sum of squares for the slope (512522) by the residual mean sum of squares for the model (26288.85). The p-value is then calculated using `pf(abs(x), df1, df2, lower=F)` where `x` is the F-value (19.498), `df1` is the degrees of freedom consumed by the slope (1) and `df2` is the residual degrees of freedom (26).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cpd	1	512522	512522	19.498	0.0001
Residuals	26	683510	26288.85		

  -

- *What can you say about the relationship between smoking cancer according to the results above? Please comment on both the null hypothesis being tested and also on the fit of the model (hint: also calculate R^2)*

  There seems to be a positive linear relationship between the number of tumor cells and the number of cigarettes smoked per day. The intercept was not found to be significantly different to zero. This suggests that individuals who did not smoke any cigarettes were not associated with any cancer cells. The slope of the regression was, however, found to be significantly different to zero. This suggests that individuals who smoke cigarettes had more tumor cells. This model explained a large amount of variation (R^2 = 512522 / (512522 + 683510) = 0.95). *However, we should also note that this model is totally inappropriate for analyzing this data since we are using count data. Stay tuned for later in the course when we learn how to deal with this.*

Activity 2

- *Write down the equation that describes the least square regression for the data in example 2 on the next page.*
  - Rate = 0.3800 + (0.33857 * X_1) + (-0.01667 * X_2) + (-0.03524 * X_3)
- *What can you conclude about the effect of team and Experiment on the Rate response?* Looking at the boxplots, there doesn't seem to be much difference between Team A or Team B. The treatment group has slightly higher values than the control group.

Activity 3

- *Is there a significant interactions between TEAM and EXPERIMENT?*

  Test significance of each term

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.38	0.045	8.444	<0.001
ExperimentTreatment	0.339	0.064	5.297	<0.001
TeamB	-0.017	0.067	-0.254	0.802
ExperimentTreatment:TeamB	-0.035	0.095	-0.368	0.716

ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Experiment	1	0.675	0.675	48.214	<0.001
Team	1	0.008	0.008	0.571	0.458
Experiment:Team	1	0.002	0.002	0.143	0.709
Residuals	22	0.318	0.014		

- *Calculate the mean for each group using the information above.*

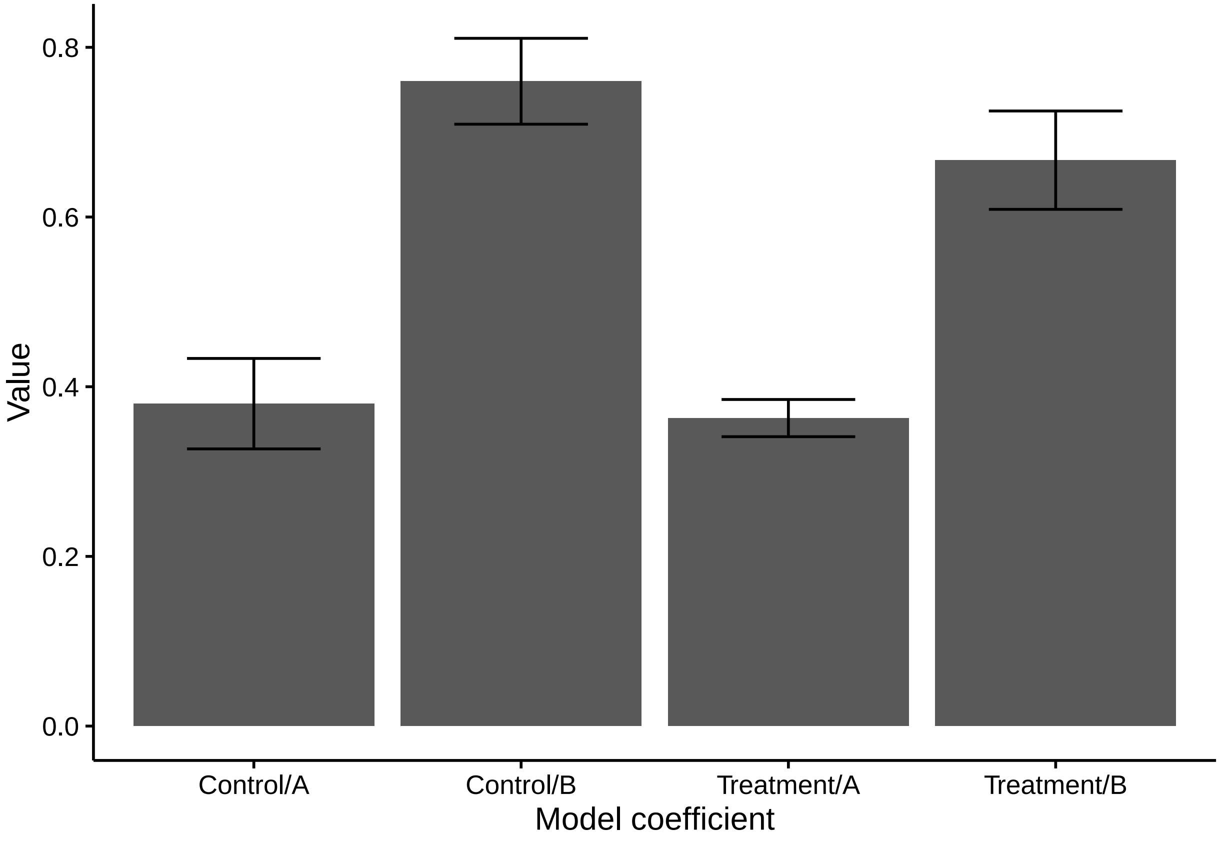
Group	Intercept (0.38)	X_1 (0.34)	X_2 (-0.02)	X_3 (-0.04)	Prediction
Control, A	1	0	0	0	0.38
Control, B	1	1	0	0	0.76
Treatment, A	1	0	1	0	0.363
Treatment, B	1	1	1	1	0.667

The predictions are calculated by multiplying the values in each column by the coefficient for that column, and then summing all those values togeather. For example, the prediction for

‘Control, A’ in the first row is derived using: (1 * 0.38) + (0 * 0.34) + (0 * -0.02) + (0 * 0.04).

- *Draw a single Bar Graph the relationship amongst the four group, and add approximate standard errors to each mean.*

```
# init
library(ggplot2)
# create data.frame with means and sds
df1 <- data.frame(
    group=c('Control/A','Control/B', 'Treatment/A', 'Treatment/B'),
    mean=c(0.38,0.76,0.363,0.667),
    sd=c(0.141,0.124,0.058,0.142),
    n=c(7,6,7,6)
)
# calculate approx errors
df1$se <- df1$sd / sqrt(df1$n)
# make plot
ggplot(data=df1) +
    geom_bar(aes(x=group,y=mean), stat='identity') +
    geom_errorbar(aes(x=group, ymin=mean-se, ymax=mean+se), width=0.5) +
    theme_classic() +
    xlab('Model coefficient') +
    ylab('Value')
```



`R` practical session