Assignment 6

1. When using analysis of variance, what are the main advantages of a
a. large sample size?
b. balanced design?

2. An observational study gathered data on the rate of progression of multiple sclerosis in patients diagnosed with the disease at different ages. Differences in the mean rate of progression were tested among several groups that differed by age-of-diagnosis using ANOVA. The results gave $P = 0.12$. From the following list, choose all of the correct conclusions that follow from this result (Borenstein 1997). Explain the basis of your answer.
a. The mean rate of progression does not differ among age groups.
b. The study has failed to show a difference among means of age groups, but the existence of a difference cannot be ruled out.
c. If a difference among age groups exists, then it is probably small.
d. If the study had included a larger sample size it probably would have detected a significant difference among age groups.

3. The parasitoid wasp, *Leptopilina heterotoma*, injects eggs into young larvae of fruit flies, *Drosophila melanogaster*. One reaction by the flies is to self-medicate by consuming alcohol (ethanol), which is naturally present in the decaying fruits where they live. The ethanol reduces oviposition by wasps, and it increases death rates of wasp larvae within parasitized flies. Kacsoh et al. (2013) investigated whether the presence of the wasp influences where female fruit flies prefer to lay their eggs. They presented female flies in cages with two dishes of fly food, one having 6% ethanol and the other with 0% ethanol. They recorded the proportion of eggs laid in the 6% ethanol dish when females were placed with female wasps, with male wasps, or with no wasps. The data below give the proportion of eggs laid in the ethanol dish for multiple replicates of each wasp treatment.

No wasp: 0.25, 0.40, 0.46, 0.44
Male wasp: 0.42, 0.47, 0.31, 0.52
Female wasp: 0.89, 0.83, 0.92, 0.93

a. Proportion data often show differences in standard deviations between groups that differ in the mean proportion (tending to be smaller in groups whose means are close to 0 or close to 1). Do these data show such a trend? To answer, make a table of the means and standard deviations of groups (include sample sizes).
b. Carry out a transformation suitable for proportion data and make a new table. Does the transformation reduce heterogeneity among groups in the standard deviation?
c. Test the null hypothesis of no differences among group means using the transformed data.
d. What fraction of the variation in the response variable is explained by treatment?

R exercise

e. Guess what: let’s load the data file *FlySelfMedication.csv* into R. Try to make a table using *aggregate*() to get the mean, sd and n and use *cbind*() to combine the three *aggregate* outputs into a nice table. With *colnames*() you can give each column a name (TIP: use *str*() to check the dimension of the new cbind-ed object). If *aggregate* does not work for you, try a method you already know

f. Perform the transformation in b to the data and remake the table. See hint[3] if you don’t know which transformation to do.

g. There are different ways to do an ANOVA; *lm*() or *aov*(). Underneath the hood they do the same, the output is different. Use them both to see if you can find the same answers in your R summary as you found in c. HINT: it is often useful to save the output of a model to a new object and use *summary*() on the new object. (model.1 <- lm (respo.var ~ expl.var); summary(model.1))

h. Compare the output of your model estimates, more precisely the *Coefficients* (model.1$coefficients), with the means of the three groups. What did R estimate with the coefficients?

i. you can check the residuals of the model. These should always sum to 0 (or a number very close to 0, as due to rounding errors this can deviate a tiny bit). Check this. Hint: you can access the residuals in a similar way as the coefficients.

Correlation and regression

4. The following data are from a laboratory experiment by Smallwood et al. (1998) in which liver preparations from five rats were used to measure the relationship between the administered concentration of taurocholate (a salt normally occurring in liver bile) and the unbound fraction of taurocholate in the liver.

Rat	Concentration (µM)	Unbound fraction
1	3	0.63
2	6	0.44
3	12	0.31
4	24	0.19
5	48	0.13

a. Calculate the correlation coefficient between the taurocholate unbound fraction and the concentration.

b. Plot the relationship between the two variables in a graph.

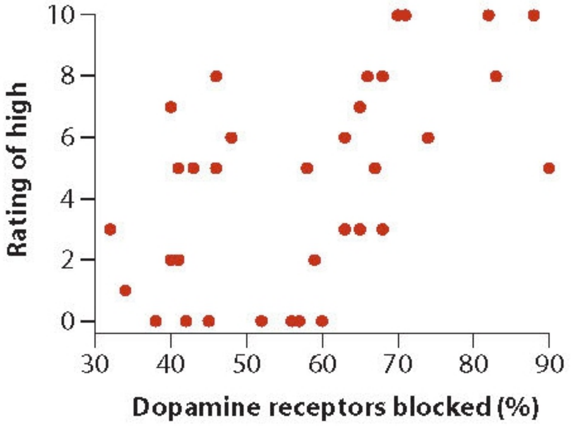
c. Examine the plot in part (b). The relationship appears to be maximally strong, yet the correlation coefficient you calculated in part (a) is not near the maximum possible value. Why not?

d. What steps would you take with these data to meet the assumptions of correlation analysis?

R exercise

e. Read the data *LiverPreparation.csv* into R and make a scatter plot. Give the axes proper labels and change the symbol type (*pch*()) and colour (to blue).
f. Calculate the correlation coefficient using R (*cor*(), note that you separate the two variable by ‘,’, not ‘ ~ ’ (there are not response and explanatory variables)).

5. Cocaine is thought to affect the brain by blocking the dopamine transporter, increasing the amount of dopamine in the nerve synapse. To investigate this, Volkow et al. (1997) administered intravenous doses of 0.3 to 0.6 mg/kg of cocaine to volunteers. They used PET scans to compare the magnitude of the perceived “high” of regular cocaine users with the percentage of dopamine transporters blocked. The results for 34 subjects are illustrated below. The full data are available at the book’s website.

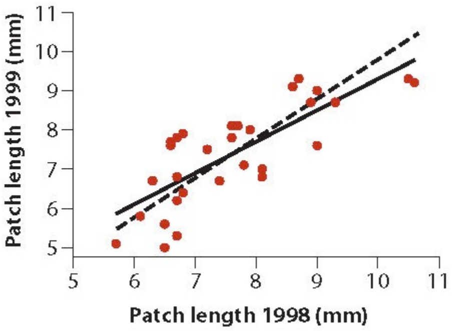


a. Using the following quantities, calculated from these data, estimate the correlation between the percentage of dopamine transporters blocked and subjects’ ratings of the cocaine “high.” Provide a standard error with your estimate.

Sum of products: 957.5
Sum of squares (transporters blocked): 8145.441
Sum of squares (rating of high): 372.5

b. Calculate a 99% confidence interval for the correlation in the population.
c. What are your assumptions in part (b)?
d. Imagine the following scenario: A second team of researchers carried out a similar study using the same population and sample size. They used a narrower range of intravenous doses of cocaine in their experiment, which led to a smaller range of values than in the above study for the percentage of dopamine transporters blocked. When they analyzed their results, they found only a low correlation between percentage dopamine transporters blocked and perceived “high.” In their published report, they concluded that the true correlation between these variables is much lower than estimated in the Volkow et al. study. Who is right? Explain.

6. The white forehead patch of the male collared flycatcher is important in mate attraction. Griffith and Sheldon (2001) found that the length of the patch varied from year to year. They measured the forehead patch on a sample of 30 males in two consecutive years, 1998 and 1999, on the Swedish island of Gotland. The scatter plot provided gives the pair of measurements for each male. The solid regression line predicts the 1999 measurement from the 1998 measurement. The dashed line is drawn through the means for 1998 and 1999, but it has a slope of 1. The difference between the two lines indicates that males with the longest patches in 1998 had smaller patches in 1999, relative to the other birds. Similarly, the males with the smallest patches in 1998 had larger patches, on average, in 1999, relative to other birds.



a. The following table summarizes the data. Use these numbers to calculate the regression slope.

		Sum of	Sum of
	Mean	**squares**	**products**
Patch length 1998	7.62	45.43	
Patch length 1999	7.40	47.03	36.26

b. Now let the patch length in 1998 be the response variable (*Y*). Use the patch length in 1999 to predict patch length in 1998. What is the slope of this new regression? Does this choice of response and explanatory variable make biological sense to you?

c. What is the most likely reason why the slope is less than one in both regressions?

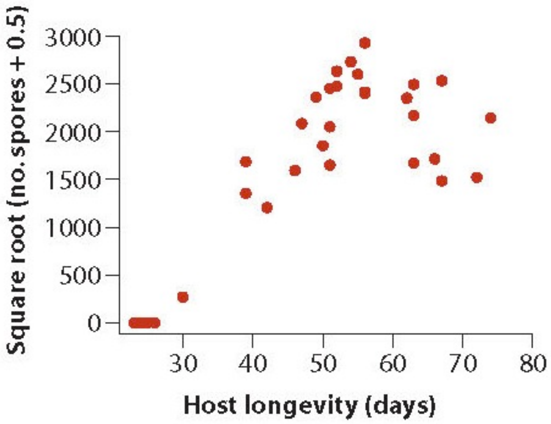
R exercise

d. Load the data *FlycatcherPatch.csv* into R. Make a scatter plot with the patch in 1998 as response variable and 1999 as explanatory. Switch the orientation of the numbers on the y-axis 90° CW by adding , las = 1 to the *plot*() formula).

e. Do a regression analyses and add the predicted line to the graph. Also draw a line with slope of one and intercept of c(0,0) to the plot. HINT: you can use *abline*() and enter the model output and it will draw the fit, but this extends beyond the range of x values. You can use *lines*() instead, but more tricky but great control and widely applicable.

f. Advanced: Extract the residuals from the model output (first store the regression model in an object). What do the residuals have to sum to? Hint[6]. Each residual is the difference between the expected (line) and observed value for a specific response variable. Check this by looking at the vector of residuals and the vector of predictor (1998) values. Think of a smart way to plot the residuals into the graph. Using a *for*() loop for all 30 observation is wise!

7. The parasitic bacterium *Pasteuria ramose* castrates and later kills its host, the crustacean *Daphnia magna*. The length of time between infection and host death affects the number of spores eventually produced and released by the parasite, as the following scatter plot reveals. The *X*-axis measures age at death for 32 infected host individuals, and the response variable is the square-root transformed number of spores produced by the infecting parasite (Jensen *et al*. 2006).



a. Describe the shape of the relationship between the number of spores and host longevity.
b. What equation would be best to try first if you wanted to carry out a nonlinear regression of *Y* on *X*?

R exercise
c. Load the data DaphniaParasiteLongevity.csv into R. Plot the data.
d. Performa regression analysis and plot the fitted line. Look at the output of regression (especially R^2) and the fit of the line. Is this a good fit?
e. Let's fit a quadratic. The most basic way is to make a new variable in which to store the quadratic values and include it into the model as a second term. Check the output. What does the R^2 value tells you about the fit of the model to the data?
f. Advanced: Try to plot the predicted values from the quadratic model using *predict*(). This take two argument, an object containing the model , lets say *m.quadratic*, and a list of x values which have to be predicted. The latter is both for *longevity* and *longevity^2*. Using *seq*() make a range of x values you want predicted values for. One other tweak needed is that you use 'data = …' in your *lm*(). So that you don't need to call the data.frame.name$.... (e.g. data frame called daphe: *lm*(daphne$Y ~ daphne$X) is same as : *lm*(Y ~ X, data = daphne). You call the name of the variables now in the *predict*(). If this is too cryptic, see hint[7].
g. Advanced: Comparing the different models using the R^2 value is a good first pass. But typically the variance explained will increase with more variables, the real question is if it is worth to add the extra variable, estimate its parameters (e.g. slope) and loose degrees of

freedom. To evaluate this one can perform formal model selection test. Try out the *drop1*() function and look at the output.

Hints:
[3] Arcsine transformation and don't forget the square root step. Arcsine is *asin*() function.
[6] The residuals always have to sum to 0. This is always good to check and small deviations should only be due to rounding errors and in the range of 10^{-16} range).
[7] data frame name is daph. Run the model: daph.m2 <- lm(sqrtSpores ~ longevity + longevitySq, data = daph). To check the range of X values; range(daph$longevity) and choose your own range; range.x <- seq(22, 74, 1). Get the predicted values for the range: daph.pred2 <- predict(daph.m2, list(longevity = range.x, longevitySq = range.x^2)). You can use the predicted (Y) values together with the X values you provided (range.x) to plot using lines().