Worksheet 3: ANOVA

You will use one of the built-in R data sets called `crabs` in the `MASS` package ("Modern Applied Statistics with S" by Venables & Ripley).

```
      library(MASS)
      names(crabs)
      dim(crabs)
      crabs[1:10,]
      sapply(crabs,class)
      attach(crabs)
```

All set? There are two categorical variables, `sex` (gender) and `sp` (species). Let's use a function called `table()` to see how many samples are in each of the possible sub-groups:

```
      table(sp,sex)
```

This is a well-balanced factorial design, with the same number of samples within every possible combination of the two factors.

We want to investigate the influence of these two independent, categorical variables on one or more of the response variables (*i.e.*, `FL`, `RW`, `CL`, `CW`, `BD`). To find out what these variables actually are, type `help(crabs).`

So, there are two ways to look at an analysis like this. First, as scientists, we might just want to know whether gender and species are controlling factors on crab morphology. Second, more practically, we might want to have a statistical model that will allow us to estimate the size, length, weight, etc. of crabs, just by knowing the gender and/or species, without having to measure them (let's assume that the former is much easier and cheaper). *ANOVA will allow us to do both of those things*.

The R function that 'fits' an ANOVA model is called `aov().` There is an analogous function for linear regression called `lm()` that we saw briefly and will spend some time using in the coming weeks.

But first, let's check out the response variables (measurements).

```
      plot(data.frame(BD,CL,CW,FL,RW))
```

Note that this is exactly identical to the following expression:

```
      pairs(cbind(BD,CL,CW,FL,RW))
```

Which is not unusual in R, as we have seen. Anyway, it looks like all the measurements are highly correlated with one another, so it doesn't matter too much which one we pick to do this exercise.

By the way, I mentioned the `cor()` function in class. While we're looking at several columns of continuous data, let's just evaluate the correlation matrix.

```
     cor(cbind(BD,CL,CW,FL,RW))
```

This tells us the correlation coefficient for all the pairs of variables, and as we saw from the graph, they are all close to one, so they should all behave similarly.

Let's pick the first one, BD ('body depth') as our response variable of interest.

```
     plot(sp, BD)
     plot(sex, BD)
```

What do you think? Let's look at the values of the means by group using `tapply()`.

```
     tapply(BD, sex, mean)
     tapply(BD, sp, mean)
     tapply(BD, list(sex,sp), mean)
```

And, don't forget to check the 'grand mean'.

```
     mean(BD)
```

OK. So, in ANOVA, the predictive 'model' is just a mean value, the mean of the samples of a group. So, given a new observation of a sample belonging to a certain group (or groups), we estimate the response variable to be the mean of the samples in that group (or groups). The trick is deciding whether belonging to a group matters, *i.e.*, if there is a significant relationship between the groups and the response.

Specifically, for example if neither gender nor species seems to have an effect on body depth of crabs, then you would just use the overall ('grand') mean as the model, *i.e.*, every crab is the same, plus/minus uncertainty. Alternatively, if those things do matter, then we   use their group means, and the uncertainty will be smaller.

Let's estimate the models, starting with the one that includes both variables and interaction.

```
     m2i = aov(BD ~ sex*sp)    # two-way ANOVA with
interaction
     m2n = aov(BD ~ sex+sp)    # two-way ANOVA no interaction
     m1g = aov(BD ~ sex)       # one-way ANOVA on gender
     m1s = aov(BD ~ sp)        # one-way ANOVA on species
```

There's no particular reason to do these all at once, except that we know we will want to eventually choose the model that is "best", according to a number of constraints: (1) we prefer models that have lower error; but (2) the lower error, often due to more complexity, should be statistically significant; and (3) the underlying assumptions should be satisfied in order to interpret the model's significance.

Let's start with the one-way model `m1s`. This model assumes that species is the important factor, and that if you know species, then your estimate of BD will be the species mean. The generic R function for extracting model coefficients is called `coef()`. I think we saw this with a regression in first lab.

Is the `m1s` model valid, *i.e.*, are the differences in the means of groups `B` and `O` statistically significant?

```
summary(m1s)
```

Three stars. So, we "reject the null hypothesis that the groups in `sp` all have the same mean".

Now, compare `coef(m1s)` with `tapply(BD, sp, mean)`. See what's going on? Repeat this with model `m1g`.

The case of ANOVA with two variables, each with two levels, is best visualized using a 2x2 grid. We'll do this together in lab, but if you get there first, compare `coef(m2n)` with `tapply(BD, list(sex,sp), mean)`.

So we see what the ANOVA model is, and what is does. How do we pick the right one? The R function that helps us to do this is called, somewhat confusingly, `anova()`.

Let's use `anova()` to see if the difference in error is significant between the top two models.

```
anova(m2i, m2n)
```

Yes, slightly. This is evidence that we should keep the interaction term. I like to test all the models pairwise, and discard the ones that do not significantly reduce the error. Try all the combinations and make sure you understand what's happening before moving on.

So, you should have selected the full model `m2i` as the "best". Now we want to validate the assumptions. Note that we could have looked at assumptions first, before model selection. Going in this order, we want to check out the assumptions of `m2i`. If it looks bad, then we probably will want to revisit to the other models.

```
    plot(m2i)
```

What was that? These plots help us visualize the properties of the residuals. What we are looking for would be an obvious variation across groups in the mean or variance of the errors.

Finally, use `shapiro.test()` and `bartlett.test()` to check the residuals.

So, what's the answer? I have a female blue crab. What's the expected value of body depth?