Basic Statistical Analyses

The purpose of this handout is to introduce you to basic statistical analyses using R. You will learn how to do two types of statistical analyses that are very common in Biology, 1) Student's t-test and 2) Analysis of Variance, commonly abbreviated as ANOVA.

To ensure that everything in this handout works, please make sure you have completed the second handout in this series titled "Exploratory Data Analysis and Plotting." If you are using the Vassar RStudio server, the dataset you worked with before should be loaded in as soon as you log in. Also, remember to type your code into a script window, which easily allows you to save and modify your code for future use.

1 Determining What Type of Analysis to Do

There are several questions to ask yourself as you prepare to analyze your data.

- What kind of data do you have?
  - Normal or non-normal response?
  - Continuous or discrete response?
  - Continuous or discrete predictor?
  - Only one predictor or multiple predictors?

You probably do not know what all of that means, and at this point that is okay. Here we will be concerned only with "normal" data. By normal data, we are referring to data that fit a normal distribution, which means that there is an approximately even distribution of values around the mean. How do you know if your data are normally distributed? That is a tough question. In a simple sense, if you take a set of numbers (i.e., your data) and calculate the average value (i.e., the mean), about half of the raw values should be larger then the mean, and half should be smaller. Lastly, most of the raw values should be fairly close to the mean whereas fewer should fall far away from the mean.

Many standard statistics rely on data being approximately normally distributed. Statistics that require "normal" data are also called parametric statistics. For the most part, all of these models follow the same basic design in R.

- `lm(Response ~ Predictor1 + Predictor2 +..., data)`

The code above indicates 1) we have a function called `lm` (for "linear model") that specifies what type of analysis we are going to do, 2) we use a ~ to separate the response (AKA dependent variable) written on the left side of the equation from one or multiple predictors (AKA independent variables) written on the right side of the equation, and 3) we specify where R will find the data. Table 1 shows how different combinations of continuous and discrete predictor variables determine which type of model you might do.

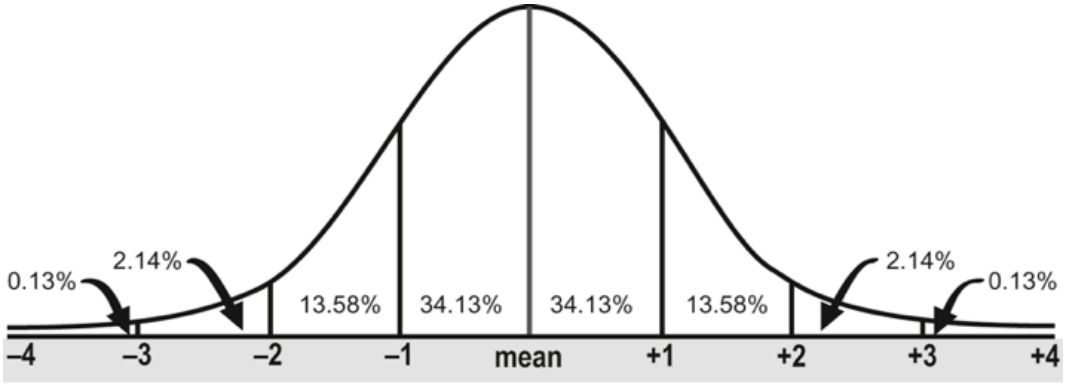

Figure 1: A normal distribution (also called a Gaussian distribution) showing the even spread of values around the mean

Table 1: Linear models and how they relate to different types of response and predictor variables. The response variable in all models described below is normally distributed.

Response variable	Predictor variable	# Predictors	Test
continuous	discrete	1 (2 categories)	t-test
continuous	discrete	1 (>2 categories)	one-way ANOVA
continuous	discrete	>1	multi-way ANOVA
continuous	continuous	1	linear regression
continuous	continuous	>1	multiple regression
continuous	discrete & continuous	>1	ANCOVA

2 The Data

This section serves as a reminder of the tadpole dataset, which you have worked with previously. These data come from a hypothetical 3 X 2 factorial experiment investigating interacting effects of predation risk and resources on tadpole growth. There are 8 variables, viewable with the `str()` command.

Which variables do you think are response variables and which are predictors?

- **Response variables**
  - Size (measured once during larval period)
  - Spots (measured number of spots on tadpole tail)
  - MetSex (at metamorphosis, recorded sex of juvenile froglets)
  - AgeAtMet (time from hatching to metamorphosis)
- **Predictor variables**
  - Pond (one of five ponds where eggs were originally collected)
  - Food: (food tretament: control or protein supplemented)
  - Pred: (predator density: low, medium or high density)

In addition, there are some response variables that might be useful as predictors of other response variables. For example, perhaps age at metamorphosis is actually a predictor of sex at metamorphosis.

3 Student's t-test

One of the simplest statistical analyses to conduct is Student's t-test, and there are multiple uses for the t-test. (For an interesting and brief read about why this analysis is called what it is, read The History of Student's t-test.) The t-test is particularly useful for small sample sizes (N $<$ 30). With a large sample size, the t-test is equivalent to a linear model (more on the below).

1. Compare the means of two independent groups of normally distributed data.
2. Compare the means of two groups of paired data (e.g., before and after measurements).
3. Compare the mean of a single set of data with a hypothetical mean.

The function to conduct a t-test in R is simply `t.test()`. Although there are several ways to code the model, the most straightforward is how we presented above `t.test(response~predictor)`. For example, maybe we would like to know if there is a significant difference in the final size of metamorphs based on the food treatment they experienced as tadpoles. One might expect that tadpoles fed the protein enhanced diet would grow to a larger size. In the code below, notice that we no longer call each variable directly from the data frame using the $ operator, but instead we use `data=` to specify where the model should look for any variables we give it. We will also add one additional argument, `var.equal=TRUE` which tells the model that we have equal variance in our two groups.

```
> t.test(Size~Food, data=tadpole, var.equal=TRUE)

 Two Sample t-test

data:  Size by Food
t = -18.2163, df = 298, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.825630 -3.884637
sample estimates:
mean in group Cont mean in group Prot
          13.01187           17.36700
```

What does that tell us? First, R tells us what we did (a two sample t-test) and reminds us of what we are analyzing, the Size variable by the Food treatment. Next, we get the t-statistic and the estimated degrees of freedom, and the p-value for the test. This tells us that the mean SVL of metamorphs differed if the rearing tank received Control or Protein enhanced food. The last two lines of the output tell us the means of each group. We can clearly see that the means are quite different; metamorphs from the Protein enhanced diet group were a little more than 4mm longer than individuals fed a Control diet.

4 Linear Models

The model commonly referred to as a linear model, or `lm()`, is one of the most flexible and useful in all of statistics. The constraint on a linear model is that the response variable must be normally distributed, but the predictor variable (or variables) can be either continuous or discrete (i.e., categorical). See Table 1 for the lay terms which are commonly applied to linear models with various combinations of predictor variables.

4.1 Useful functions for linear models

There are several functions that are very useful for any kind of linear model. These include the following:

- `summary()` - Summarizes your model and gives you important information such as adjusted R-squared and treatment means, as well as the F-statistic and p-value for the model. For linear regression and ANCOVA, provides slope and intercept of best fit regressions.
- `anova()` - Provides a very brief summary of the overall model but does not provide information on individual levels within factors. Statistical significance of each predictor is calculated in a stepwise manner, adding each factor one-by-one.
- `Anova()` - Provides similar summary information to the `anova()` function, but statistical significance of each predictor is calculated assuming all other factors are in the model. Found within the `car` package.
- `plot()` - Provides diagnostic plots of the residuals of the model. Useful for assessing model fit and balance.

4.2 One-way analysis of variance - ANOVA

For the purposes of this tutorial, we will only concern ourselve with one of the most common types of statistical analyses is the Analysis of Variance, called ANOVA for short. In its simplest form, an ANOVA is a statistical test of whether or not the means of multiple (more than two) groups are equal. Thus, it is essentially the same as the t-test but for more than two groups. The resulting p-value and test statistic give us some idea of the probability that those group means are indeed different. For example, perhaps we want to know if Size differs across our three Predator treatments (Low, Medium and High predator density). Here, we will actually make our model into an `R` object and then examine it using the `summary()` function. Note that in this example, the order of the treatments in the Pred variable have been made Low, Med, and High, which is different from the default of High, Low, Med (which is alphabetical). See tutorial #2 for instructions on how to change the order of the treatment levels.

NOTE: The information below is being provided to help explain what `R` is doing and will be useful for you to understand in the future, but it is not crucial to understand right now. Although you are encouraged to read through this section, you can also feel free to skip to section 4.2.1.

```
> lm1<-lm(Size~Pred, data=tadpole)
> summary(lm1)

Call:
lm(formula = Size ~ Pred, data = tadpole)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.8414 -2.1285  0.0826  2.0741  4.5086

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.4835     0.2419  72.289  < 2e-16 ***
PredMed      -2.5101     0.3420  -7.339 2.07e-12 ***
PredHigh     -4.3721     0.3420 -12.783  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.419 on 297 degrees of freedom
Multiple R-squared:  0.3566,  Adjusted R-squared:  0.3522
F-statistic:  82.3 on 2 and 297 DF,  p-value: < 2.2e-16
```

So what all does that mean? We created an object called `lm1` and then used the `summary()` function to look at it. The top of the summary output tells us the model we made and gives us information about the distribution of the residuals around the mean. The section entitled `Coefficients` provides information about the estimated mean for each level (Low, Med, and High) within the factor (Pred). The way `R` reports estimates of means is that a baseline is established (the first treatment group, which we designated as Low in tutorial 2), and everything else is in relation to that baseline. Thus, in the output for `lm1`, what is labeled as `(Intercept)` is the mean for the Low treatment. Thus, the SVL of metamorphs from the Low predator density tanks is 17.48 mm.

The subsequent `Estimates` are how much those treatments differ from the baseline. Thus, the `PredMed` level is 2.5101 larger than the Low-Predator treatment, or

```
> 17.4835-2.5101
```

or 14.97 mm. Lastly, the SVL of metamorphs from the High treatments are the last line, and are

```
> 17.4835-4.3721
```

or 13.11 mm. The bottom three lines of the output provide the summary statistics that you might include in a manuscript. The adjusted R-squared, the F-statistic, the degrees of freedom, and the p-value for the ANOVA. As you can see, the Predator treatment had a very signficant effect on metamorph SVL `p-value: < 2.2e-16`. We can confirm the treatment means with the `tapply()` function.

```
> tapply(tadpole$Size, tadpole$Pred, mean)
    Low     Med    High
17.4835 14.9734 13.1114
```

4.2.1 Using the `anova()` function

Another way to view the summary statistics for the model as a whole is with the `anova()` function, which gives just the necessary info that you might include in a manuscript: degrees of freedom (both numerator and denominator), F-statistic and p-value. *This is the most important thing for you to understand right now.*

```
> anova(lm1)
Analysis of Variance Table

Response: Size
           Df  Sum Sq Mean Sq F value    Pr(>F)
Pred        2  962.76  481.38  82.296 < 2.2e-16 ***
Residuals 297 1737.27    5.85
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The `anova()` function gives you the most important information about your statistical test. 1) The degrees of freedom (here, 2 and 297), 2) the F-statistic (here, 82.296) and 3) the p-value (here, $P < 2.2 \times 10^{-16}$).

4.2.2 Diagnostic plots

It is always a good idea to look at the diagnostic plots for your model. All statistical analyses have underlying assumptions, such as the response variable being normal, as we have already discussed. However, other assumptions include having relatively even amounts of variation in your different treatment groups (e.g., the Prot vs Cont food groups). By default, the `plot()` function will give you four diagnostic plots of your model.

- The residuals plotted against the fitted values (treatment means). This is used to check for constance of variance across your treatments.
- A quantile-quantile normal distribution plot (called a Q-Q plot for short), plotting standardized residuals against the theoretical quantiles from a normal distribution. This is used to check for normality of errors.
- The square root of the absolute value of the standardized residuals plotted against the fitted values (treatment means). Similar to the first plot, this is used to check for constancy of errors across treatments.
- "Cook's distance", which is a measure of the influence of each observation on the parameter estimates.

These plots are more useful in advanced statistics, but for now, the first plot is really the important one to look at. We want to make sure that the amount of variation in our data is roughly similar across the different treatment groups.

```
> plot(lm1)
```

These plots look okay. The first and third show a relatively even spread of points across our three threatment means, which is good. The second plot shows a modest similarlity between our standardized residuals and the straight line derived from a normal distribution. This indicates good model fit. If the points in the Q-Q plot deviate considerably from the line, your data do not fit a normal distribution well. The last plot does not show us much in this case, but that is good. If there were points with too much leverage, it would be obvious.

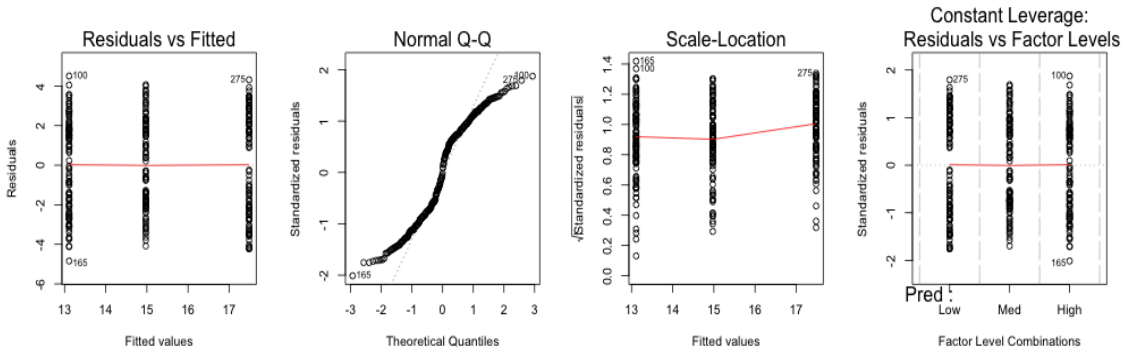

Figure 2: Four diagnostic plots of model `lm1`

4.2.3 Post-hoc comparisons

It is often useful (and necessary) to be able to compare the different levels within a single categorical variable. For example, in the example here we might want to know which of our three Predator treatments are different from one another. This is accomplished with the `glht()` function in the `multcomp` package. `multcomp` is short for "multiple comparisons" and `glht()` stands for "general linear hypothesis test". **NOTE:** You will need to install and load the `multcomp` package. To do this, either 1) type `install.packages("multcomp")` at the prompt or 2) click on the Packages tab in the lower right corner of the screen, then click the "install" button and type the name of the package (multcomp). Once the package is installed, you will need to load it by clicking on it in the Packages window or by typing `library(multcomp)` at the prompt.

```
> ph1<-glht(lm1, linfct=mcp(Pred="Tukey"))
```

What does all that code mean? We are using the function `glht` to analyze a model we have already made (`lm1`), and we have defined a linear function (the `linfct` part) which will do multiple comparisons (the `mcp` part). We have specified which type of multiple comparisons to do by saying to conduct a Tukey test on all the levels of the factor `Pred` in our model. By specifying Tukey comparisons, we have told `R` to do all possible pairwise comparisons (in this case, low vs medium, low vs high and medium vs high). We can view the output of the post-hoc tests with the `summary()` function.

```
> summary(ph1)

   Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: lm(formula = Size ~ Pred, data = tadpole)

Linear Hypotheses:
               Estimate Std. Error t value Pr(>|t|)
Med - Low == 0   -2.510      0.342  -7.339  < 1e-07 ***
High - Low == 0  -4.372      0.342 -12.783  < 1e-07 ***
```

```
High - Med == 0   -1.862      0.342  -5.444 2.49e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Adjusted p values reported -- single-step method)
```

The main thing to look at in the output above is 1) which treatments are being compared with one another and 2) the p-values for those comparisons on the right hand side. The post-hoc comparisons here allow us to conclude that size at metamorphosis differs significantly for all three predator densities. You can see that the reported t-statistics for the comparisons between Med and Low treatments, or High and Low treatments, are the same as those listed in the summary output of the linear model (-7.339 and -12.783, respecively). However, here the p-values are adjusted for the multiple tests being run, and so are more appropriate to use. A handy function (particularly if you have many treatments) is `cld()`, which stands for "compact letter display". It gives you a very simple depiction of which treatments are statisically different from one another.

```
> cld(ph1)
 Low  Med High
 "c"  "b"  "a"
```

Treatments with the same letter are not different. Thus, in this case, since all three treatments have different letters, we can see they are all different from one another.

4.3 Student's t-test as a linear model

As mentioned above, when the size of the dataset being analyzed is > 30, the t distribution is identical to the normal distribution. Thus, there is really no practical difference between a t-test and a one-way ANOVA except the number of categories in the predictor (see Table 1 above). We can run our t-test from above using the `lm()` function with the following code.

```
> lm2<-lm(Size~Food, data=tadpole)
> anova(lm2)
Analysis of Variance Table

Response: Size
           Df Sum Sq Mean Sq F value    Pr(>F)
Food        1 1422.5 1422.54  331.83 < 2.2e-16 ***
Residuals 298 1277.5    4.29
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

5 BIOL 106 Pre-lab assignment

This pre-lab for Module 2 contains three short assignments. Your instructor will tell you when it is due for your section and how you should turn it in. For each part, paste the output into a Word document, and write 2-3 sentences answering the questions below.

1. Conduct a t-test analyzing the effect of Food treatment on Age at metamorphosis. Did the predictor variable significantly affect Age at metamorphosis? How do you know?
2. Conduct a one-way ANOVA analyzing the effect of Predator treatment on Age at metamorphosis. Did the predictor variable significantly affect Age at metamorphosis? How do you know?
3. Using the model you made in #2, conduct a post-hoc test comparing the three Predator treatments. Were any of the three treatments significantly different from one another? Were any treatments not significantly different from one another?