# Lab #4, Part #2 – The t-Test and Experimental Data

The data we will use for this lab is from a Kaggle competition (see here: https://www.kaggle.com/lsind18/weight-vs-age-of-chicks-on-different-diets). The data include weights of baby chickens (chicks) at birth (`birthWeight`) and then 21 days later (`endWeight`). Weights are reported in grams (gm). The chicks are separated into 4 groups, each given a different diet. Since the diet is a manipulated variable – and chicks are assigned randomly to diet – this is an experimental design.

Throughout the lab, we'll use this data to test differences in weights across chicks on different diets. You can find the data at the Lab #4 page on Canvas.

*Please be sure to include all code and all visuals and tables in your submission on Canvas. Note that all responses to questions should be in complete sentences and provide details and explanations when asked.*

**Activity #1: Open R-Studio and Load in Data.**
Go ahead and download the files (`chickDat.csv`) from the Canvas Lab #4 page and save them to a new lab folder in your class folder directory (Lab 4 would be a good choice for your new folder) as .csv files. Then open up R-Studio, set your working directory with `setwd()`, and load the data. Name the data `chickdat`.

```
#set working directory and load data
setwd("c:/Users/…")
chickdat <- read.csv(file = "chickDat.csv", h = T)
```

Be sure that you've told R that the data has variable names (headers) with `h = T`, and confirm that they show up correctly. Take a look at your data and make sure it has 50 rows and column headers with the `head()` and `nrow()` functions. Also check for missing values (since it's a small data set you can probably just print the entire thing by typing `chickdat` and seeing if anything is missing).

**Activity #2. Explore Your Data and Make New Groups.**
Because today's lab involves t-tests – which require comparison of one or two groups – we are going to create a new `diet` variable called `diet2`. For this variable, we are going to group together `diet 1` and `diet 2`, and group together `diet 3` and `diet 4`. For our purposes for now, you can assume these diets are similar in nature such that the first two `diet` categories are a control group (a normal chick diet) and the other two `diet` categories are the treatment group (high fat or carb diets). You can create the variable with the code below. Note that we are indicated the control group as "0" and the treatment group as "1".

```
#consolidate diet groups into diet2
chickdat$diet2 <- ifelse(chickdat$diet == 1, 0, NA)
```

```
chickdat$diet2 <- ifelse(chickdat$diet == 2, 0, chickdat$diet2)
chickdat$diet2 <- ifelse(chickdat$diet == 3, 1, chickdat$diet2)
chickdat$diet2 <- ifelse(chickdat$diet == 4, 1, chickdat$diet2)
head(chickdat)
```

Now that we have our new variable, we want to be sure that we take a look at the data. Remember that if our group sizes are less than 30, we generally assume that our data distribution is normal. Luckily, we do have group sizes larger than 30 (one of the reasons we consolidated), but it's still a good idea to get a look at your data before continuing forward with analysis.

**2.1** Make a side-by-side histogram of the `birthWeight` and `endWeight` data with `hist()` and `par(mfrow = c(1,2))`. Be sure to include coherent axis labels (including unit of measure) and titles for your figures. Include these in your lab report and describe what you see across these figures with respect to skew and the distributions of each of the grouped histograms. Save your histograms as .png files with the `png()` function we've used in previous labs.

I've provided some code below to get you started (make sure to save as a .png as noted above).

```
#make histograms of weight by time
par(mfrow = c(1,2))
hist(chickdat$birthWeight, xlab = "Birth Weight (gm)",
     main = "Birth Weight of All Chicks", probability = T)
```

**2.2** Now, make 2x2 matrix of histograms of `birthWeight` and `endWeight` for each of the `diet2` groups with `par(mfrow = c(2,2))`. Include these in your lab report and describe what you see across these histograms with respect to skew, location, and spread.

I've again provided some code below to get you started (make sure to save as a .png as noted above).

```
#make histograms of weight by time and diet2 group
par(mfrow = c(2,2))
hist(chickdat$birthWeight[chickdat$diet2 == 0], xlab = "Birth
     Weight (gm)", main = "Birth Weight (gm) of Chicks in the
     Control Group", probability = T)
```

**2.3** Use the `aggregate()` function to look at the mean `weight` by `diet2` group with the code below. Modify the code and do the same for the `endWeight`, and then again for the standard deviation of each of the two weight measures. Report these in your lab report, and compare across groups. Make sure to put your values in a neat and tidy table.

```
#calculate mean weights by time and diet2
aggregate(chickdat$birthWeight ~ chickdat$diet2, FUN = mean)
```

**Activity #3. One Sample t-Test.**
Assume that we know that the average birth weight for chicks in general is 41 grams, and that this is the correct population parameter for birth weight (`birthWeight`). We want to test whether, at birth, our chicks are representative of the population of chicks. This allows us to generalize from our study on the effect of diets better. It also should help inform us that we've got a simple random sample of chicks from the population.

**3.1** State the null and alternative hypotheses for our t-test.

For this exercise, we'll introduce a new function in R-Studio: `t.test()`. This function does all the work for us. In the code below, I ask R-Studio to test the birthweight of chicks in `chickdat` at birth against the null hypothesis, `mu = 41`. Note that I've told the function that I want it to perform a two-sided hypothesis test. This is actually the default for `t.test()`, and you don't need to include it, but I wanted to make that clear here. We'll stick with two-sided hypothesis tests for today.

```
###test our samples weights against the population
t.test(chickdat$birthWeight, mu = 41, alternative = "two.sided")
```

**3.2** Report the results of the t-test. Don't just copy and paste your output, but report the test statistic, degrees of freedom, and reported p-value. Relate this back to your null and alternative hypotheses in words and in the context of the data (note that R-Studio reports the average difference between the observations and the population value). What does this say about our data and the population from which it was sampled?

**3.3** Now that you have this tool, let's make sure it is doing the calculation correctly. Calculate the test statistic by hand (include the work in your lab write-up – you can also type in the mathematical code manually and include it in your submitted R code) and see if it comes out the same as R calculated. Remember from the lectures that for the one-sample t-test, $t = \frac{\bar{x}-\mu}{s_{\bar{x}}}$ and $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$. You've already calculated $\bar{x}$ and $s_x$ in R-Studio in the previous activity (and $n = 50$ is given), so you just need to calculate the standard error and test statistic. Do you get the same test statistic and degrees of freedom that R-Studio spit out for you with `t.test()`?

**Activity #4. Independent Samples t-Test.**
Now that we've established our chicks are pretty representative of the population in terms of `birthWeight`, let's move toward testing our independent, randomly assigned `diet2` group final weights on day 21 (`endWeight`). Since our groups are independent of one another, and were randomly assigned, we'll use the independent samples t-test to evaluate whether there are differences in `endWeight` across the control `diet2` and treatment `diet2`.

Note that five chicks did not make it to Day 21 and died before their weights could be taken (you should have seen this when inspecting for missing data in Activity #1). Although it is unclear if diet had anything to do with this, we are going to assume that this happened at random. This means

that the sample size for this example will be smaller, with a total *n* of 45, whereas *n* = 50 at the start of the study (`birthWeight`).

We'll again make use of the `t.test()` function, but add some code to it. We want to tell R-Studio that we have two groups of interest, and to test the *difference* in these groups.

**4.1** For this test, write out the null and alternative hypotheses both with mathematical notation and in words.

We'll begin by assuming the variance between the two groups is equal. You can check this in your original table you made in Activity #2 with `aggregate()`. For this test, I've provided the code below. Notice that now, rather than providing a state population parameter (`mu`), we are asking R-Studio to evaluate `endWeight` by (`~`) `diet2` group.

```
###perform independent samples t-test for ending weight and diet
t.test(chickdat$endWeight ~ chickdat$diet2, var.equal = TRUE)
```

**4.2** Report the results of the t-test. Don't just copy and paste your output, but report the test statistic, degrees of freedom, and reported p-value. Relate this back to your null and alternative hypotheses in words and in the context of the data. What does this say about the weight of the two groups in the experiment? Do the degrees of freedom look correct? Do you think it was reasonable to assume equal variance (standard deviation) across the two groups? Why or why not?

**4.3** Repeat **4.2**, but remove the `var.equal = TRUE` from your code. Note that this will estimate what is called Welch's Test, which is just the t-test that adjusts for unequal variance across groups. Do the results change? What changes? Does this change impact whether you reject your null hypothesis or not?

Although we've tested whether all the data seems representative of the chick population at birth (Day 0), and we've tested whether weight was different for the two diet groups on Day 21, it's possible that the two groups were different to start. Given this, let's do an additional t-test that tests the difference in `birthWeight` across our two `diet2` groups. If we randomized correctly, hopefully we will see that `weight` is quite similar across the two groups. This could give us more confidence in the comparison of the `endWeight` on Day 21 in the previous problem.

**4.4** Write out your null and alternative hypotheses and go ahead and do the same test in R-Studio, but this time for `birthWeight` and assume equal variance. Report the results of the t-test. Don't just copy and paste your output, but report the test statistic, degrees of freedom, and reported p-value. Relate this back to your null and alternative hypotheses in words and in the context of the data. What does this say about the weight of the two groups at the start of the experiment?

**Activity #5. Paired Samples t-Test.**
In the previous activity, we largely ignored the fact that our data is paired! We tested whether there were differences in weight across the diet groups after 21 days of being on the diet. And we tested whether there were differences in starting weight for the two groups to be more confident that they were comparable at the start. But usually in these situations, we'd prefer to make use of the valuable information that these data are paired together. In other words, measurements were taken twice for each chick. This sort of question is better suited for the next lab, where we cover ANOVA. For now, we'll ask a very simple question for the context of a paired t-test.

Here, we simply want to know if weight is different at birth (`birthWeight`) relative to at the end of our experiment (`endWeight`). In other words, are our chicks growing?

**5.1** Write out your null and alternative hypotheses with mathematical notation and then in words.

Remember that we lost 5 chicks before they could be weight on Day 21. Let's just go ahead and remove these chicks' rows from our data with the code below.

```
###remove missing Day 21 chicks
chickdat <- subset(chickdat, is.na(chickdat$endWeight) == FALSE)
nrow(chickdat)
chickdat
```

You should be able to use this code to see there are now 45 rows in the data, and that there are no more missing `endWeight` observations. But note that our $n = 45$ now, rather than our previous $n = 50$.

**5.2** We'll use the `t.test()` function again, this time with another small adjustment so that R-Studio knows the data are paired together. The code below should get you there. We'll assume variance is not equal, so this will be a variant of Welch's test. As with previous tests, report the test statistic, p-value, and degrees of freedom. Relate this back to your null and alternative hypotheses in words and in the context of the data. What does this say about the weight of chicks at birth and the weight of chicks after 21 days?

```
###test starting and ending weight
t.test(chickdat$endWeight, chickdat$birthWeight, paired = TRUE)
```

**5.3** Now repeat 5.2, in its entirety but do so separately for the control group (`diet2 == 0`) and then the treatment group (`diet2 == 1`). It may be that one group got larger over this period while the other didn't.

**5.4** Take a look at the mean differences reported for each of the two paired t-tests you've done here. Do they seem different? Relate this back to Activity #4 and describe the overall pattern you see. Note that while we can test these separately, as noted at the start of Activity #5, the preferred way to do this analysis would be an ANOVA. We'll get to that next week, but for now, we want to start thinking about more complex questions related to these differences.