# Assignment 4 Instructions

Linda Forrester, Rachel Schwartz. Markdown: Eren Ada

**BIO104 Lab A4: Plant Differentiation and Growth**

**Goals for Lab 4**

- Understand where the mass of the plants came from
- Understand variation in data and what that means
- Understand replication and pseudo-replication
- Represent variation in data on graphs
- Compare two treatments statistically, where the data varies among individuals

## 1. Getting Started

Variation in biology is generally used to describe the differences between individuals (or cells, or groups of organisms) in genetics, in phenotype, or in some combination. In data science, variation has a different meaning, as a measure of spread in data. Measures of variation (spread) describe the distribution of the data and may be calculated in many ways including range, variance, standard deviation, and interquartile range. Variance for example may be an indication of biological variation, but there can also be an additional source of variation that is included, that is chance variation, also known as noise or random error.

Chance variation can be a major challenge in analyzing data because practical considerations limit sample sizes and number of experiments. This random variability/noise appears at multiple levels, either because of changes in time, space, procedural changes, experimental bias, experimenter-bias, or chance events. What can we do to reduce this chance variation? Remember the 3 Rs - Randomization, Rigor, and Replication. Randomization reduces bias, and rigor means following the scientific method and use a methodical approach. Replication of treatments reduces random error, but we aim for true replicates and try to avoid pseudoreplication. How do you know whether you are pseudoreplicating? Think of the unit of your analysis.

Here you will be experimenting to determine the impact of light and dark on plant growth. To test for treatment effects, you must have (independent) replications of your treatments. In your experiment you may have 100 plants in each condition, light and dark, but no matter how many plants, you still only have one true replicate of the treatment of interest (light vs dark).

**1a. Avoiding pseudoreplication** Calculate the average height of your group's light and dark seedlings at each interval (2 days, 4 days, and 6 days) using Google Sheets. You should have 6 values total. Put your group's data into a Google Sheet in long form. Set it up with the column names exposure (light or dark), day (2,4,6), and height (what you measured).

By calculating the average of the 6 seedlings you measured, you are getting a good estimate of your group's sample. Each of the seedlings you measured are called pseudoreplicates, while the means for each of your seedling groups are called a true replicate. Calculating a single estimate from multiple plants minimizes the error in your estimate of the height of your seedlings (i.e. measurement error). However, it means you only have one height estimate per treatment/day.

**1b. Put class data in long form on Google Sheets** Go to the Lab Sakai Site and scroll down like you did in Labs 1 and 2. Select the link for your lab section. Put your group's data into the class' Google Sheet in long form using the same columns you did for your group data. Note that there is another column for group number. Copy the link to this Google Sheet to use in RStudio.

Now you have good estimates of multiple samples to work with. There will be some variation among samples due to the soil, the location the plants were grown, and how different people measured height. Having multiple samples allows you to estimate a mean height despite all this variation.

**1c. Load the packages you need**

```
# Store the Google sheets link as the variable url


# Load the data stored in the variable url using the gsheet2tbl function
# and store it as the variable plant_data
```

**2. Get your data into R** Remember, once we have our data in, it is always good to check to make sure the data was imported correctly. Use the commands you learned in Lab A2, Part 2.

```
#Check your data to make sure it looks right (first lines only and column names)
```

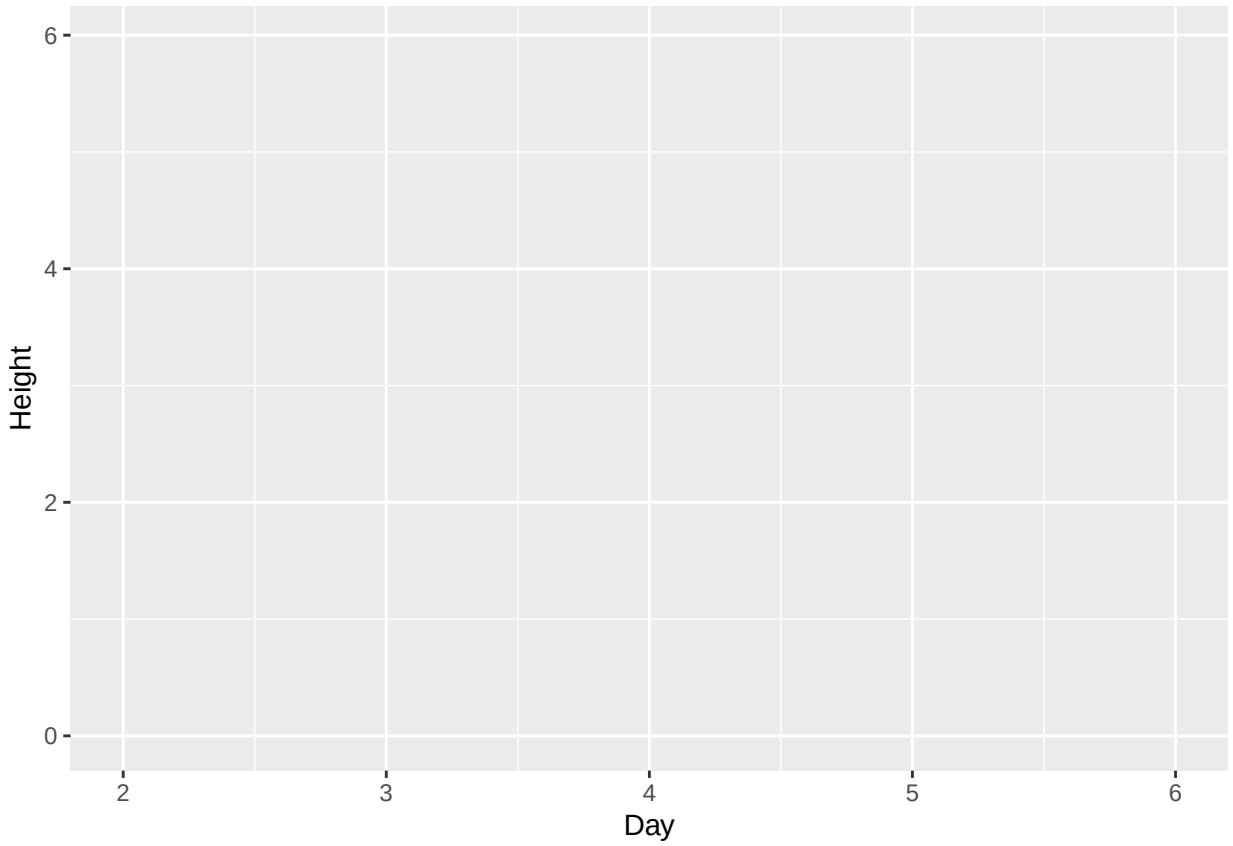
```
## # A tibble: 1 x 4
##   `Student Initials` Exposure   Day Height
##   <chr>              <chr>    <dbl>  <dbl>
## 1 RM                 light        2    0.5
```

**3. Graphing the Seedling Data** Because we want to compare the seedlings grown under light and dark conditions, you are going to graph both sets of data together. You will first graph the individual data as points. This way you can see how much variation is in your data.

**3a. Create the base layer of your plot** What is the independent variable (x axis)?

- What is the dependent variable (y axis)?
- How will you plot light and dark conditions separately? Refer to Lab A2, Part 4b if you need help.

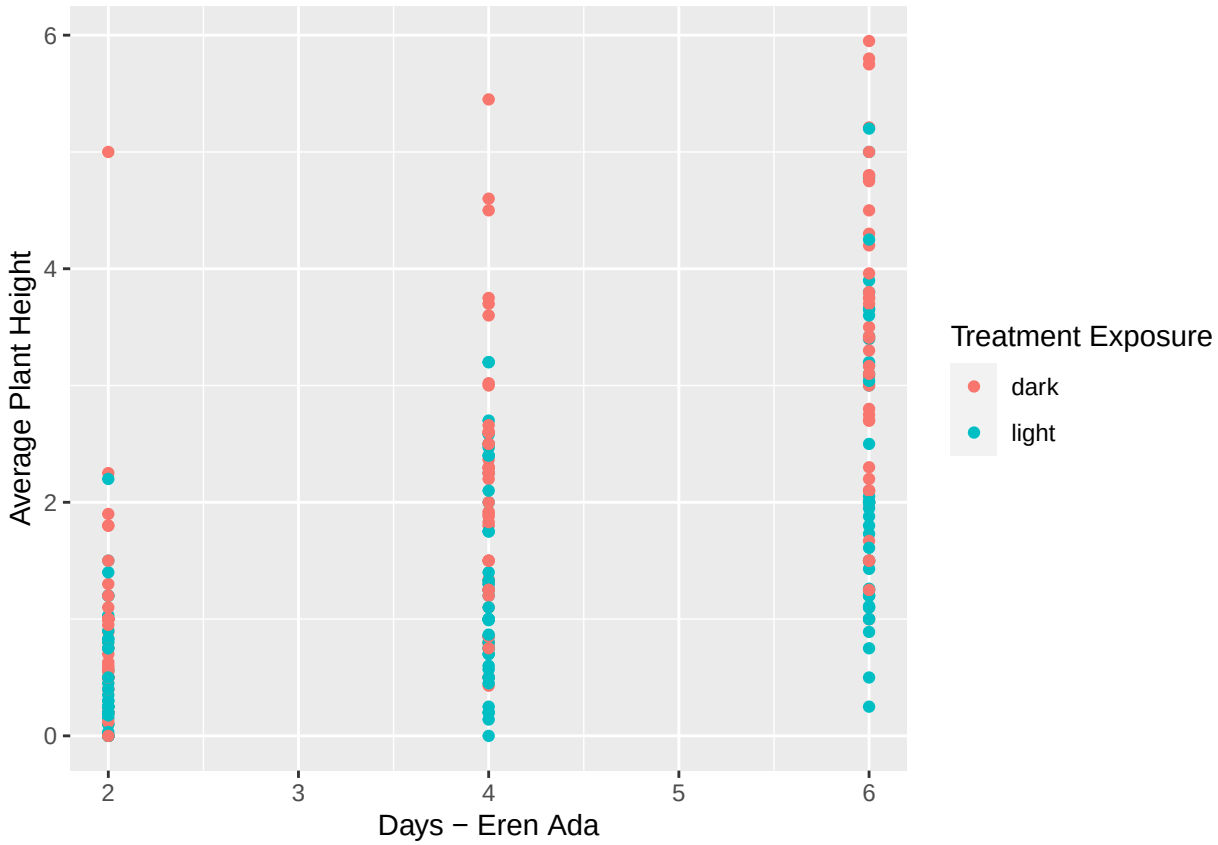
```
# Create the base layer of your plant_data plot using ggplot
```




**3b. Add the data and labels to the plot**

- Note: the variable day is an integer, which means it is on a continuous scale rather than in discrete categories; thus, the functions to label your axes are continuous (`scale_x_continuous`, not `scale_x_discrete`). We use a discrete scale when plotting variables like `SiteType` or `Color`, where each variable is its own category. Refer to Lab A1, Part 5b if you need help.

```
# Add datapoints and labels to the plot using geom_point
```




**Save this plot as Lab4_graph1 so that you can turn it in with your lab report.**

Does it look like there might be a difference between the two groups?

Based on just this plot, what can you say about the differences? How confident do you feel?

**4. Compare the Height of Seedlings in Different Conditions** We are interested in comparing the growth rates of seedlings in light and dark conditions. To make these comparisons we need to know the mean of each group of exposure data and how much variation there is within each group.

- Group the data by exposure and day. Refer to Lab A2, Part 3 if you need help.

```
# Group the plant_data variable by exposure and day using the group_by function
# and store it as the variable grouped_plant_data
```

Calculate the mean height of each group. Refer to Lab A2, Part 3 if you need help.

- Add `na.rm=TRUE` into the `mean()` function. This additional code tells R to ignore areas in the data set that had no values entered. For example, if no seedlings grew under your dark condition, you would not record any data under height for that experiment. By using this code, R can now ignore these blank spaces and calculate the mean correctly.

```
# Calculate the mean height of the seedlings at each interval using the summarise function and store it
```

```
## `summarise()` has grouped output by 'Exposure'. You can override using the `.groups` argument.
```

Check the first few lines of your mean data `(plant_data_means)` using `head()`

```
head(plant_data_means)
```

```
## # A tibble: 6 x 3
## # Groups:   Exposure [2]
```

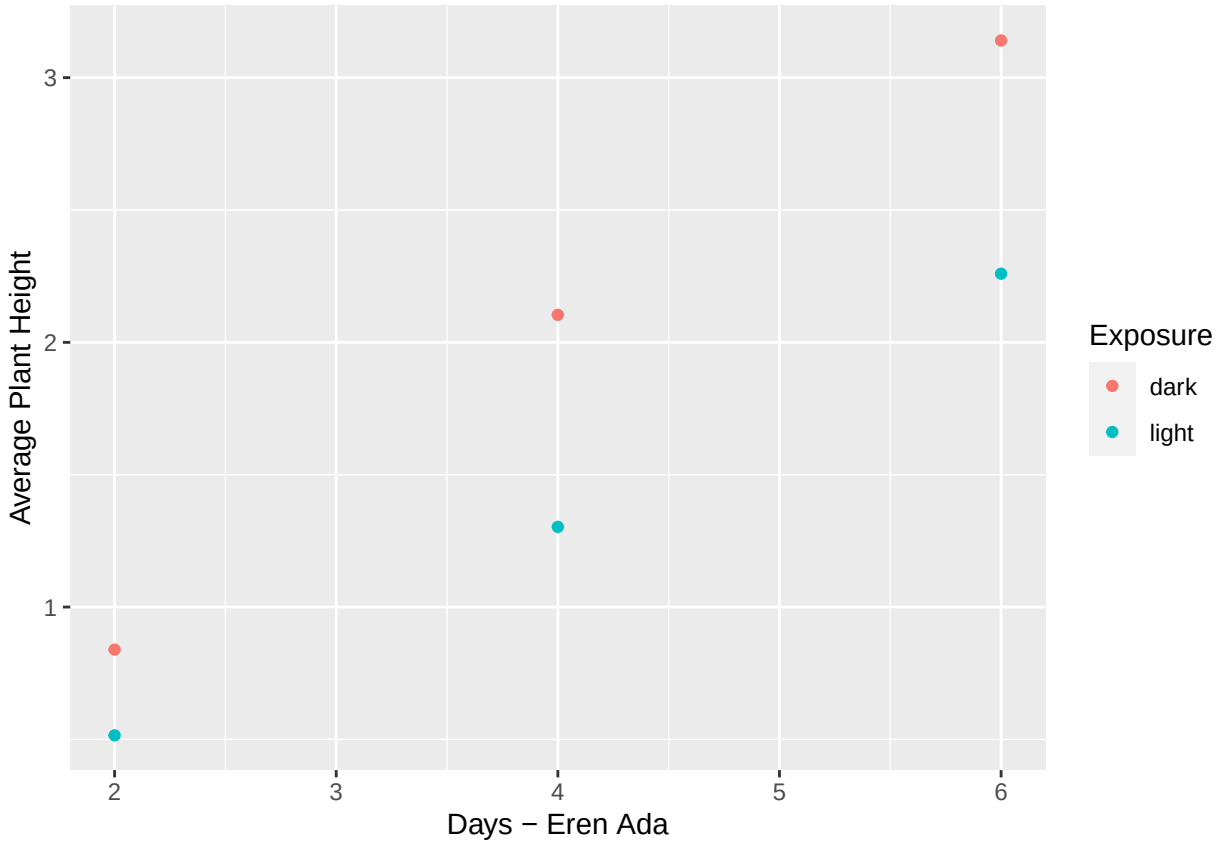
```
##   Exposure   Day  mean
##   <chr>    <dbl> <dbl>
## 1 dark         2 0.839
## 2 dark         4 2.10
## 3 dark         6 3.14
## 4 light        2 0.515
## 5 light        4 1.30
## 6 light        6 2.26
```

**5. Graphing the Mean Seedling Data** Because we want to compare the seedlings grown under light and dark conditions, we are going to graph both sets of data together.

**5a. Create the base layer of your plot and add points**

- What is the independent variable (x axis)?
- What is the dependent variable (y axis)?
- How will you plot light and dark conditions separately? Refer to Lab A2, Part 4b if you need help.

```
# Make your plot for the plant_data_means data using the ggplot command
```




**5b. Calculate the standard deviation of the means** It appears that there are differences in seedling heights between the two different growing conditions. However, if there is a lot of variation in our data those differences may not be statistically significant. We can check if the light and dark grown seedlings are statistically significantly different using a t-test.

- The standard deviation is how much the individual data points differ from the overall mean on average. Using the standard deviation gives a single value on how much variation there is in the data. For

example, if we saw seedling heights of 3, 2.9, 3, and 3.1, we would have a small standard deviation. However, if we had seedling heights of 0.5, 5.5, 2, and 4, the standard deviation would be high.

- Standard deviation is calculated in R using the function `sd`.

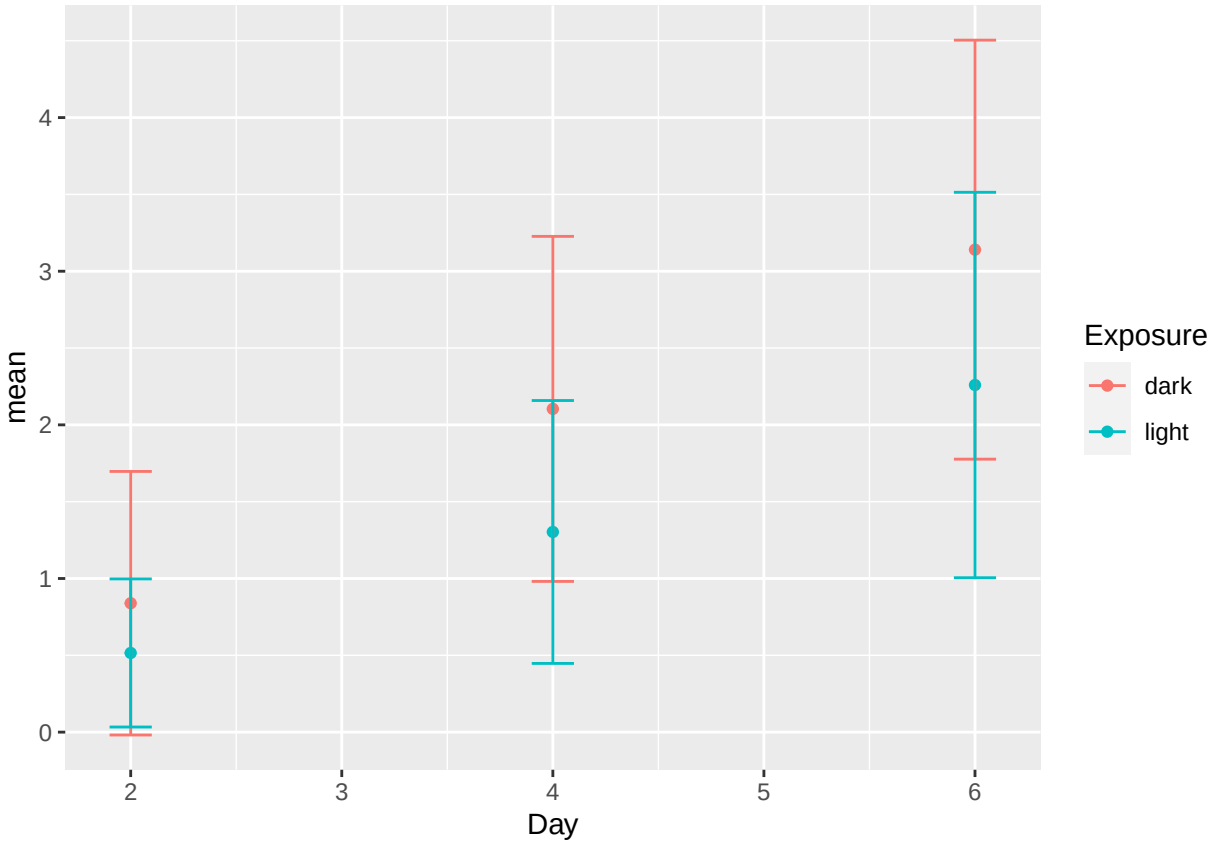
```
# Remake your table of means so it includes std deviation
# and store the additional data as the variable plant data means

plant_data_means <- summarise(grouped_plant_data,
                              mean = mean(Height, na.rm=TRUE),
                              stdev = sd(Height, na.rm=TRUE))
```

**5c. Add standard deviation bars to your plot** Next add the standard deviation bars to your graphs by adding a layer using `geom_errorbar`

- `geom_errorbar` draws an error bar that has an upper and lower value. In this case, the upper value is the mean + the standard deviation and the lower value is the mean - the standard deviation.
- Add + `geom_errorbar(aes(ymin=mean+stdev, ymax=mean-stdev))` to your ggplot command.

```
# Add the error bars to the plot using geom_errorbar
```



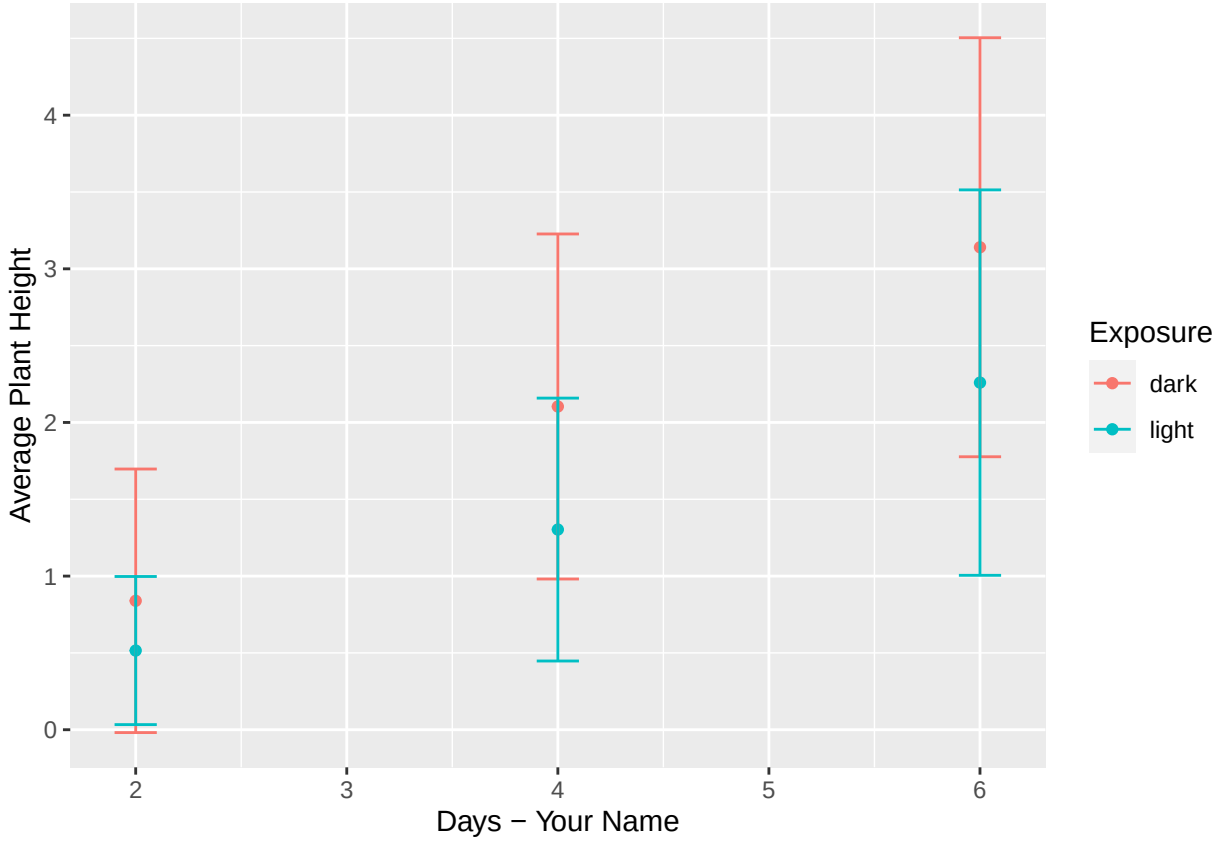

**5d. Add labels to your plot** Just like last class, you'll want to clean up your graph and make it look professional.

- Add labels to the axes and legend
- Change the width of the error bars by adding `width = 0.2` after `ymax` in `geom_errorbar`.

```
# Label your plot and change axes titles
# Edit the width of the error bars
```

```
# Label your plot and change axes titles
# Edit the width of the error bars

ggplot(plant_data_means, aes(x = Day, y= mean, color = Exposure)) +
  geom_point() +
  geom_errorbar(aes(ymin = mean+stdev, ymax=mean-stdev, width = 0.2)) +
  scale_x_continuous(name = "Days - Your Name") +
  scale_y_continuous(name = "Average Plant Height")
```

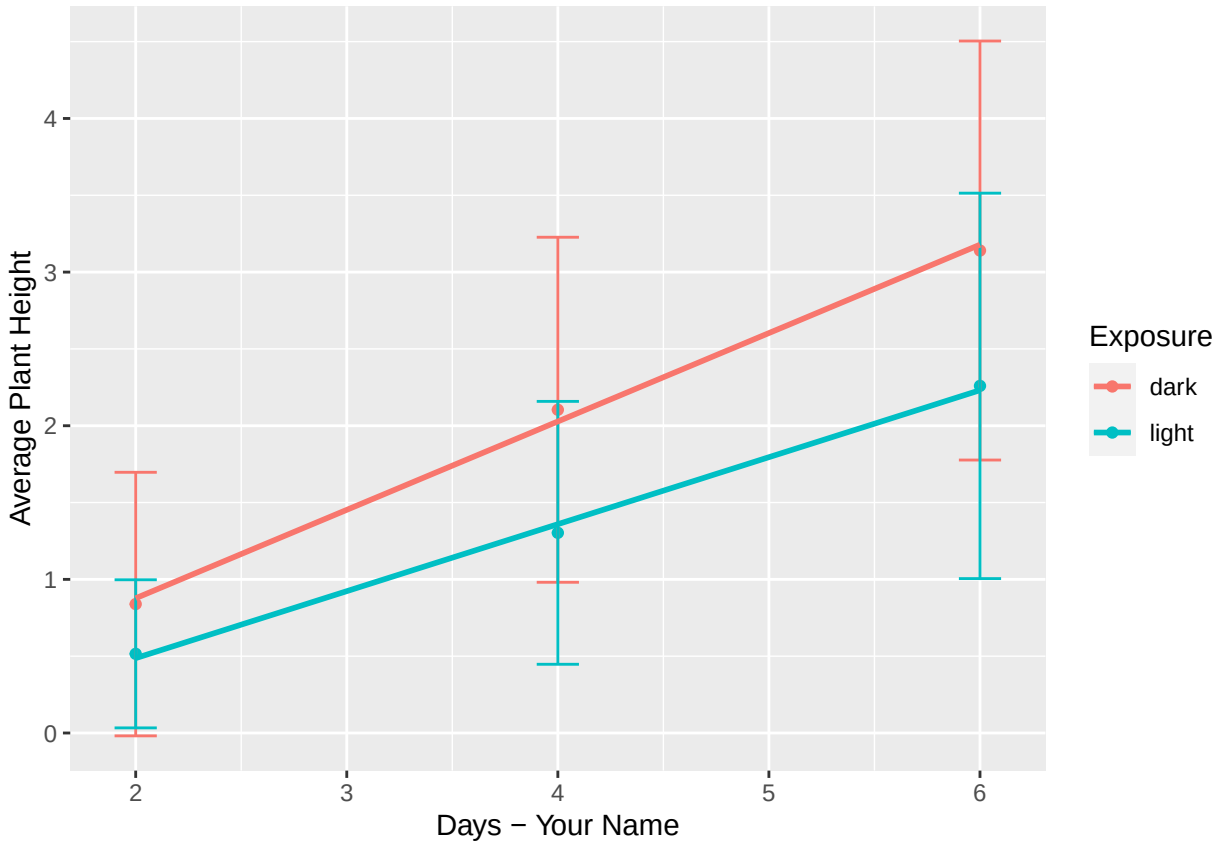



**5e. Draw lines representing a trend that fits your data** This model is a linear (straight) best fit line. It is referred to as a regression.

- Add the model lines to your graph by adding a layer using geom_smooth.
- There are two lines: one for each subset of data.
- Add `+ geom_smooth(method="lm", se = FALSE)` to your ggplot command.

```
# Add the regression to your plot using geom_smooth
```

Your plot should look something like this:




**Save this plot as Lab4_graph2 so that you can turn it in with your lab report.**

- Under which condition did the seedlings grow better? Light or dark?
- How confident do you feel? More or less than when you plotted all the data?

**6. Compare height under different conditions using statistics (a t-test)** Just like in Lab A2, we have two groups of data we want to compare to see if there is a statistical difference. In this case, we're interested in whether there is a difference between seedling heights in the 6 day old seedlings

- Filter light and dark grown seedlings for the day 6 time period. Refer to Lab A2. Part 6 for help.

```
# Subset plant data by exposure and growth time and store as the variables light6days and dark6days # R
```

- What is your null hypothesis for the 6 day old seedling data?

```
#t test to compare heights of 6 day old seedlings for light and dark conditions
```

```
##
##  Welch Two Sample t-test
##
## data:  dark6days$Height and light6days$Height
## t = 3.2057, df = 88.021, p-value = 0.001878
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.334854 1.427206
## sample estimates:
## mean of x mean of y
##  3.140378  2.259348
```

**Under which condition did the seedlings grow taller? Light or dark?**

**Lab 4 Report Submission** Turn in a hard copy of your lab report that includes the following:

1) Summary table for your section's data ( wide form, made in excel/sheets/or R). Do not include all the raw data, just means and standard deviations for each group.
2) Scatterplot of all data from Step 3b.
3) Graph with mean data points, error bars, and regression from Step 5e.
4) Your code. Code should be organized, include good comments, and include only code that works and does not produce errors.
5) Answers to the following questions (one full page, double-spaced):

- Under which condition did the seedlings grow taller? Light or dark?
- Describe the variance you see in the mean data.
  - How is this reflected in the error bars?
- How does your best fit line (regression) fit the mean data?
  - Does it help show the overall trend?
- What is your null hypothesis for the 6 day old seedling data?
- Explain why you rejected or failed to reject your null hypothesis based on your t-test results and explain the processes that may have influenced the experiment.

**More material to help you understand mean and standard deviation** Check out this interactive website for a better understanding of mean and standard deviation! http://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm

Click on the tutorial button to work though the example. Things to think about:

- When you sample multiple individuals do you see variation?
- When you calculate the mean of your samples what are you estimating?
- Why do you need to calculate the mean of many samples (each of which is the mean of multiple measurements)?