Lab A4: Plant Differentiation and Growth

TO DO:

- Take out plots
- Fix data input

Goals for Lab 4

Be able to…

- understand where the mass of the plants came from
- understand variation in data and what that means
- understand replication and pseudoreplication
- represent variation in data on graphs
- compare two treatments statistically, where the data varies among individuals
- model the relationship between two variables

1. Getting Started

1a. Avoiding pseudoreplication

Before we begin to use R today, calculate the average height of your light and dark seedlings at each interval (2 days, 4 days, and 6 days) using Google Sheets.

By calculating the average of each data point you are getting a pretty good estimate of your single sample. 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 data in long form on Google Sheets

Put your group's data into a Google Sheet in long form. Set it up with the column names

TAs - fix this

Once you have an estimate for your sample, enter this into the class data sheet (your TA will provide this). This data sheet is what you'll use during this lab.

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

Make a new project in a new directory and open a new script. *Refer to Lab 1, Part 2 if you need help.*

- Copy the comment lines into your script
- Enter the code based on the previous labs and your R scripts from Labs A1 and A2.

```
# Load the packages ggplot2 and gsheet and dplyr
```

2. Get your data into R

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

TAs fix this

Next, load in the data. Today's data is found at:

https://docs.google.com/spreadsheets/d/1MOkh3SNsuuTixjRFi4ptYmjimUJ9rGXIta1iJDjre9k

```
# Assign website address for the data to a variable
```

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

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 2, Part 2.*

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

3. Graphing the Seedling Data

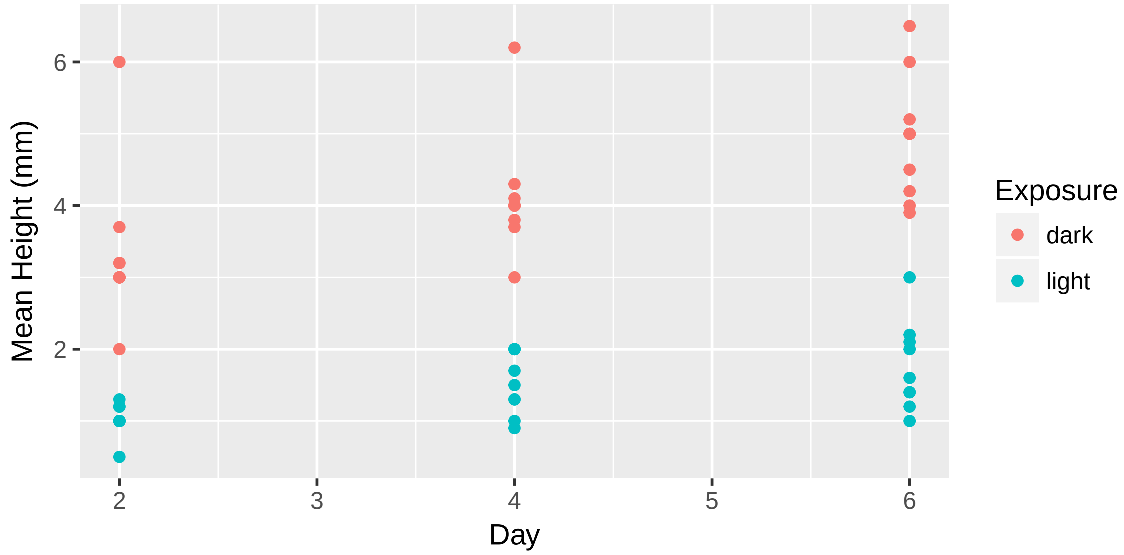
Because we want to compare the seedlings grown under light and dark conditions, we are going to graph both sets of data together. We are first going to graph the data as points. This way you can see how much variation is in your data.

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?

3b. Add the data and labels to the plot

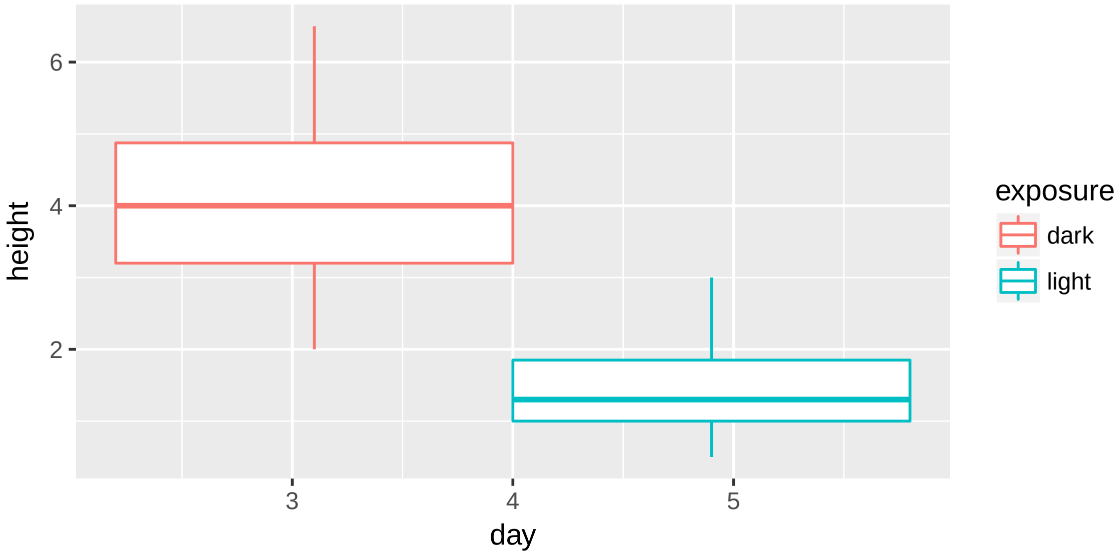
- note: the variable `day` is an integer, which means it is on a continous scale rather than in discrete categories; thus, the functions to label your axes are *continuous* (`scale_x_continous` not `scale_x_discrete`)



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

What can you say at this point about the difference? How confident do you feel?

3c. Visualize the variation in the data without plotting every point



4. Compare the Height of Seedlings in Different Conditions

We are interested in comparing the growth rate of seedlings in light and dark conditions. To make these comparisons we need to know the mean of each group of data and how much variation there is in that group.

- Group the data by exposure and day.

```
# Group data by exposure and day
```

- Calculate the mean of each group

This code is just like you used last time, with the addition of `na.rm=TRUE`. 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.

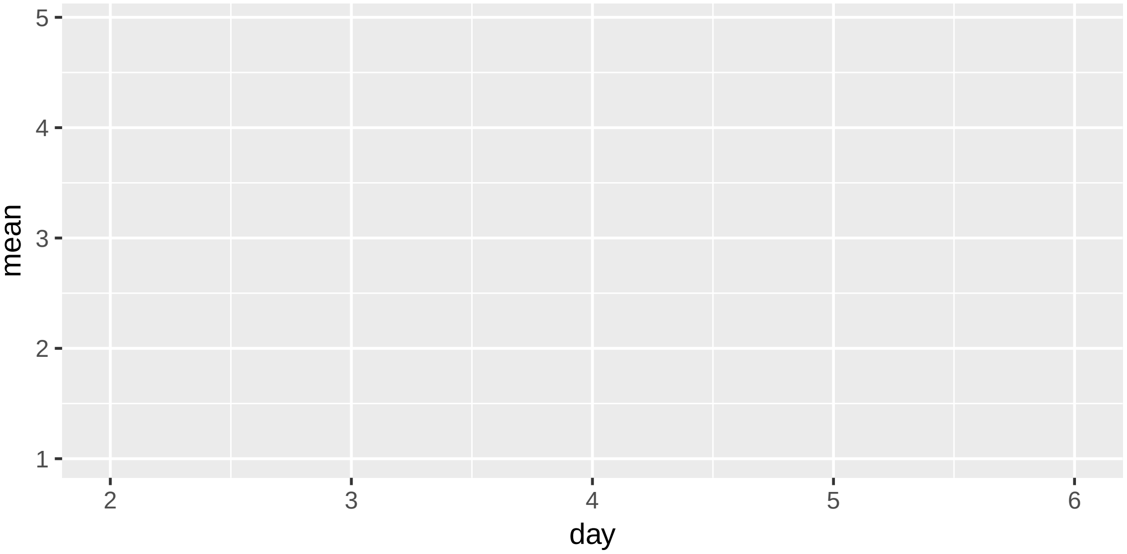
Check the first few lines of your data.

4a. Graph the Means

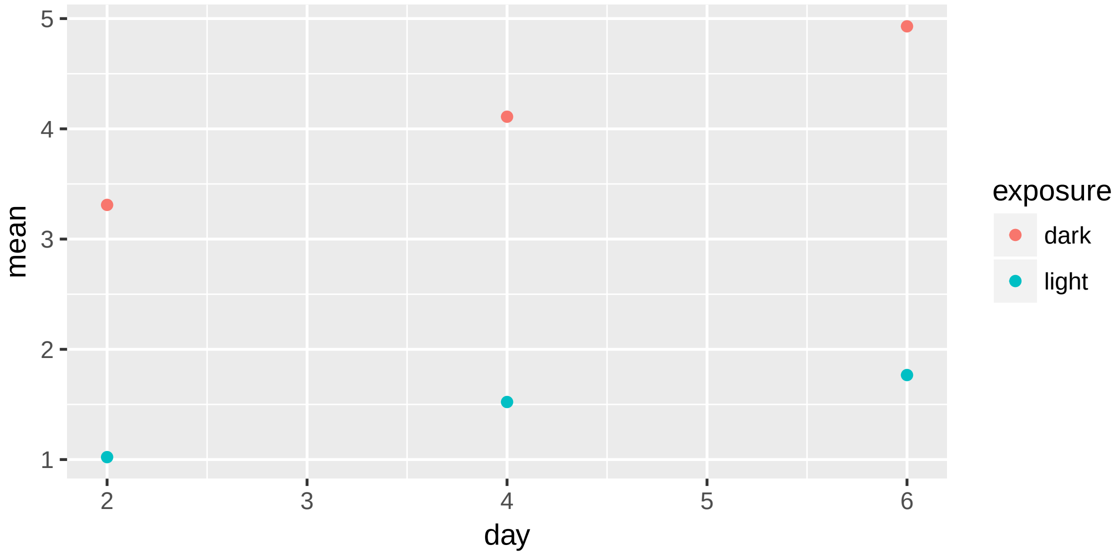
Because we want to compare the seedlings grown under light and dark conditions, we are going to graph both sets of data together.

4a1. 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?



4a2. Add the data to the plot (using `geom_point`)



4a3. Add standard deviation bars to your plot

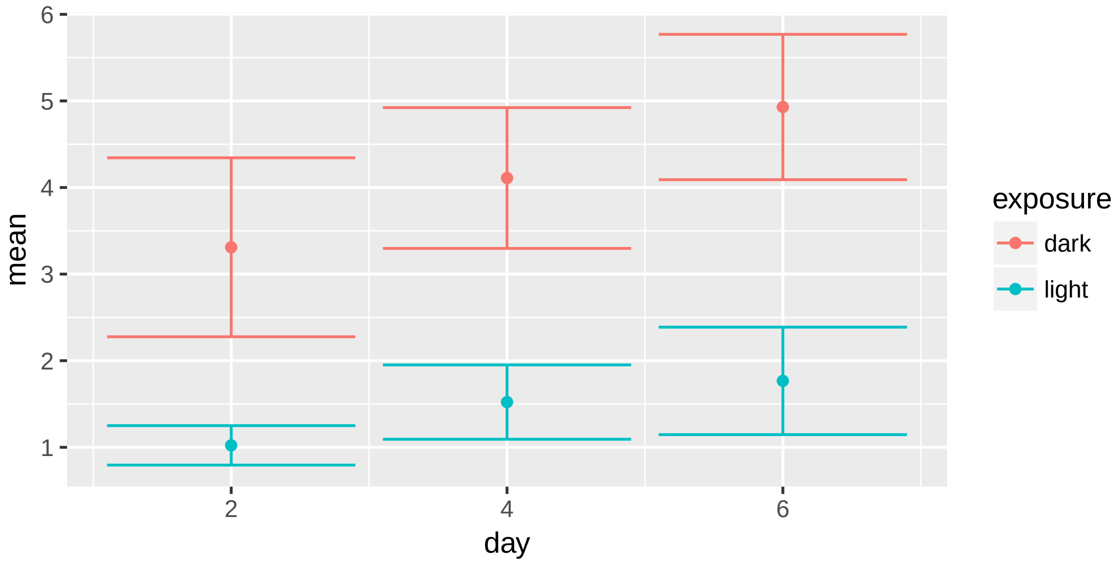
- It appears that there are differences in seedling heights between the two different growing conditions. However, if there is a lot of variation in our data those differences may not be significant (remember the t-test).
- The *standard deviation* is how much the individual data points differ from the 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
plant_data_means <- summarise(grouped_plant_data,
                              mean = mean(height, na.rm=TRUE),
                              stdev = sd(height,na.rm=TRUE))
```

Next add the standard deviation bars to your graphs by adding a layer using `geom_errorbar`

- `geom_errorbar` draws an errorbar 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 the following layer to your plot
  geom_errorbar(aes(ymin=mean+stdev, ymax=mean-stdev))
```

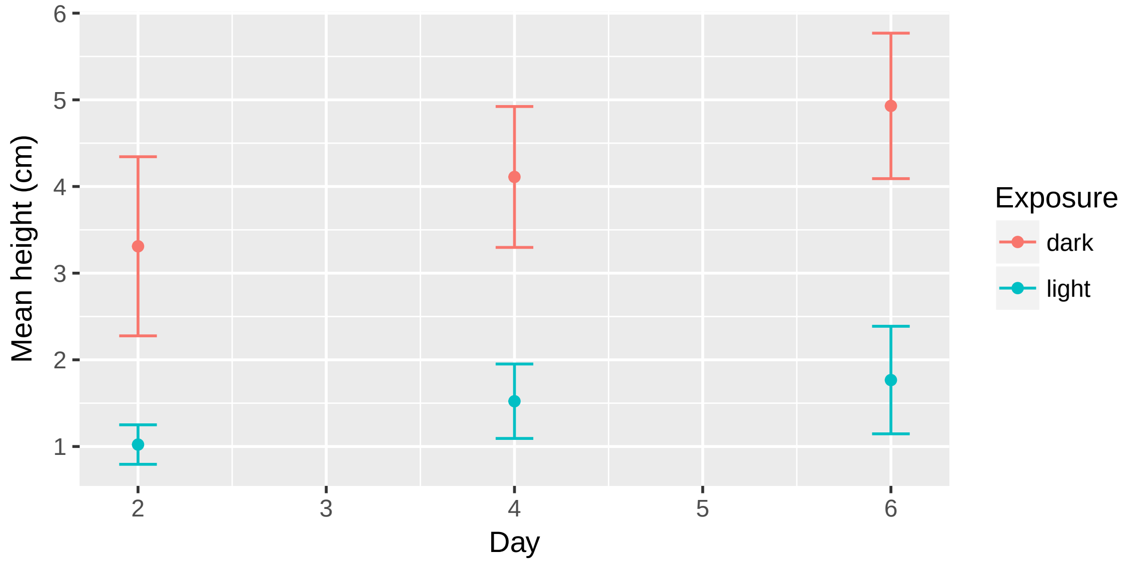


4a4. Add labels to your graph

Just like last class, you'll want to clean up your graph and make it look professional.

- Add labels
- Change the width of the error bars by adding `width = 0.2` after `ymax` in your error bar layer.

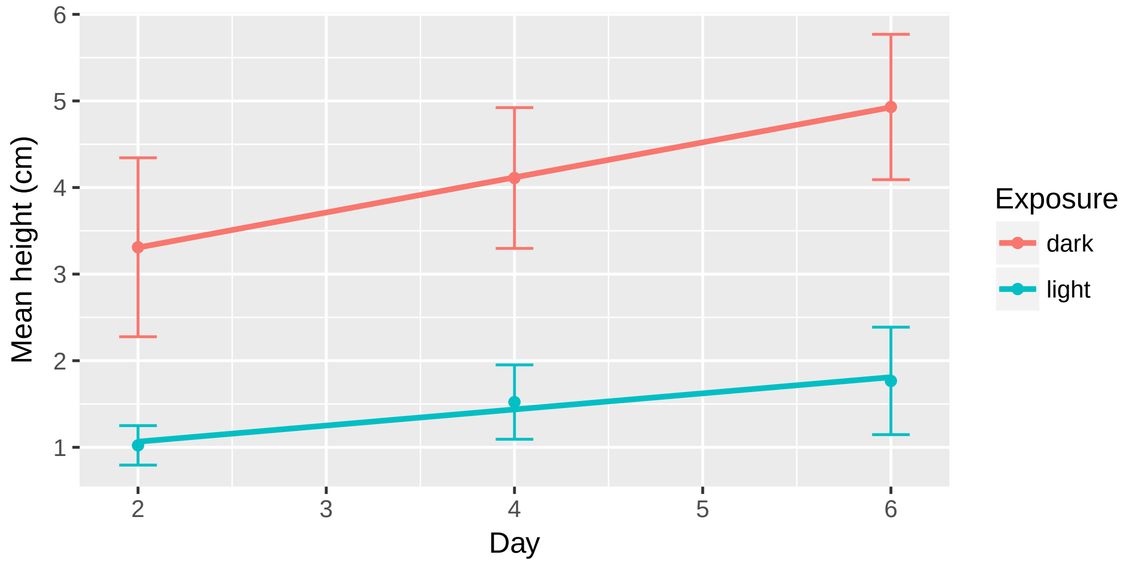
```
#Label your plot and change axes titles
```



4a4. Draw lines representing a model 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 the following layer to your plot
  geom_smooth(method="lm", se = FALSE)
```



Answer the following questions

- 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?

5. Fit a model to represent the change in seedling height over time

- This is the model shown in the lines on your plot
- Use the function `lm` (linear model)
- Provide variables as y~x (y as a function of x, i.e. y is dependent on x)
- Set `data = plant_data`

```
# Build a model
plant_regression <- lm(height~day, data = plant_data)
```

- Print the output of the model
- Look for the R-squared. This value measures the fit of the line to the data. R-squared explains the fraction of variation in the plant height that is explained by the time they have had to grow. An R-squared value of 0 means that the model does not explain why the measurements vary. An R-squared value of 1 means that the model perfectly explains changes in the measurements.

```
# Print the output of the model
summary(plant_regression)
```

```
##
## Call:
## lm(formula = height ~ day, data = plant_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4500 -1.2447  0.1526  1.1526  3.7553
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6421     0.5405   3.038  0.00364 **
## day           0.3013     0.1251   2.409  0.01939 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.542 on 55 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.09542,    Adjusted R-squared:  0.07898
## F-statistic: 5.802 on 1 and 55 DF,  p-value: 0.01939
```

What is the value for R-squared?

Notice the regression doesn't fit well. In our plot we had two separate lines - one for each exposure because the general trend is the same but the values are quite different. We need to build a model that incorporates both the effect of time (day) and exposure (light/dark) on height (mm).

- Provide additional independent variables with +

```
# Build a model by fitting each exposure separately
plant_regression2 <- lm(height~day+exposure, data = plant_data)
summary(plant_regression2)
```

```
##
## Call:
## lm(formula = height ~ day + exposure, data = plant_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5140 -0.5140 -0.1167  0.2807  2.4860
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.91140    0.27181  10.711 5.68e-15 ***
## day            0.30132    0.05922   5.088 4.69e-06 ***
## exposurelight -2.67963    0.19368 -13.836  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7301 on 54 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.801,  Adjusted R-squared:  0.7936
## F-statistic: 108.7 on 2 and 54 DF,  p-value: < 2.2e-16
```

What is the value for R-squared for this two-component model? How well does this model fit the data.

6. Compare height under different conditions using statistics (a t-test)

Just like in the Mechanisms of Evolution lab, we now have data for which we want to see if there is a statistical difference between two groups. 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.
- What is your null hypothesis for the 6 day old seedling data?
- Under which condition did the seedlings grow better? Light or dark?

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.

- 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)?