1. Introduction to the Project

In this project, we are going to analyze data (simulated by me) that supposedly comes from a genetic cross. However, some of these data may be too good to be true!

This is inspired by a discovery made by R.A. Fisher, who analyzed some of Mendel's data and found that some of these results were "too good to be true". It has been suggested that a helper knew ahead of time what the expected values for the ratios would be. In this case, there might be less error that you would expect if the helper was biased toward the results that Mendel might have expected.

In this project, I will send you a simulated data set for a particular cross, generated by two "students" of Mendel seeking to determine the pattern of segregation for a cross. Purple pea flowers are proposed to be either the genotype *PP* or *Pp* (The *P* allele which causes purple color is dominant). White flowers are proposed to by *pp*. A cross is made between a true breeding purple strain (*PP)* and a true breeding white strain (*pp*). The F1 progeny are selfed and the total number of purple and white F2 progeny are counted. Each student repeated the experiment 65 times, with 100 individual F2 progeny per experiment.

Your job is to determine whether the data from either (or both) of the students are too good to be true. Is the variance among the 65 experiments reasonable? If the data are too good to be true, there might be much less variance that expected. ***Note: The project is due on the date provided by the syllabus. However, after handing it in, I will give you one more chance to meet with me and fix the project if you want. This way, everyone should be able to get 25 points.***

2. R commands you will find useful

<-

Assign a variable.

Example:

```
x <- 5
x
[1] 5
```

c(x,y,z,....)

c stands for concatenate.

Example:

```
x <- c(1,10,30)
x
[1] 1 10 30
```

rbinom(x,y,z)

Generate **x** number of random binomial numbers, with sample size **y** and probability **z** (**z** is between 0 and 1)

Example:

```
x<-rbinom(10,20,0.5)
x
[1]  9 12 14 13  9 12 10  6 10  11
```

Calculations can be performed on a vector.

Example

```
y <- x/20
y
[1] 0.45 0.60 0.70 0.65 0.45 0.60 0.50 0.30 0.50 0.55
```

Summarizing data

hist(x): histogram
note: limits of X on histogram can be set with xlim. For example, hist(x,xlim = c(0,200)) sets the x values between 0 and 200

plot(x): scatter plot, where x is plotted as Y axis and x-axis is an index
mean(x): mean
var(x):variance

Note: Variance (s^2) is a measure of the spread of a distribution.

$$s^2 = \sum(x_i - \bar{x})^2/(n-1)$$

s^2 = *sample variance*
x_i = *ith element of the sample*
$\bar{x}$ = *mean of the sample*
n = *sample size*

Comparing distributions

Test of mean:
t.test(x,y)

Kolmorgov-Smironov test for equal distribution:
ks.text(x,y)

F-test for difference in variance:
var.text(x,y)

Control Commands

Adding something to a vector

```
x <- c(x,10)
```

Example:

```
x <- c() establishing a null vector
x <- c(x,10)
x
[1] 10
x <- c(x,10,20,30)
x
[1] 10 10 20 30
```

Loops

```
for (i in 1:10) {...
}
```

Example:

```
for (i in 1:10) {
x <- c(x,2*i)
print(x)
}
```

Note: To run a program, you will need to run it above the single line, in the text editor above, then click Run code.

3. Assignment

You will be given a set of frequencies for the purple plant from two different "students" of Mendel. Each student supposedly performed 65 different experiments with **100 F2 progeny per experiment**.

The results from student A are on the first line after your ID. The results from student B are on the second line after your ID. Perform each of the steps and provide the answers in a report. **ALSO: Please provide a text document including the code you used for the project. You will need to do screen grabs to obtain the figures and plots. In Mac, this is with the Grab tool. In PC, this is with the Snipping tool.**

20 Points.

a. Plot histograms of purple frequencies (with x ranging from 0 to 1) for the 65 different experiments that I send you from student A and student B.

b. Simulate a data set of 65 experiments (100 plants each) and plot the histogram of purple frequencies from this new "experiment".

c. Describe the differences between your simulated data set and the two histograms from student A and B. Provide the mean frequency of each and the variance of each. What do you notice? Do the distributions appear equal? If not, how so?

d. Perform the three statistical tests described above comparing your data set to the data set from each of the students. Provide the outputs in the report. Provide a paragraph of text interpreting these results. Do you conclude that the data generated by the "students of Mendel" were or were not *too good to be true?*

5 Points.

In this treatment, we assume that 100 plants were used in each experiment. Will the variance in the frequency of purple plants (across the 65 experiments) depend on the number of plants used in each experiment?

a. Using simulation, plot the variance in frequency of purple plants (for 65 experiments) with increasing number of plants per experiment. Start with 2 plants, and plot the variance up to 300 plants per each of the 65 experiments. *Provide the plot.*

b. How do you interpret this result? Provide a few sentences.