

Laboratory 2: Artificial Selection Experiment. **Statistical Analysis and Graphical Presentation**

Introduction

Now that you have completed the artificial selection for fast *C. elegans*, it is time to assess whether your selection procedure had an effect on the populations you created. In other words, it is time to determine whether the worms **responded** to the selection imposed upon them. Over the course of the experiment, some students in the “selection” group may have noticed that the worms appeared to move “faster” as the experiment progressed. Other “selection” group students may have noticed that many more *C. elegans* crossed the finish line in the later generations than in the parental generation. Although these informal observations are important, we need a more conclusive determination and characterization of the response of our *C. elegans* populations to the selection pressure imposed - in both the “selection” and the “control” groups. In order to do this, each student must analyze the data. Using the results of these statistical analyses, we can determine between which generations significant differences (if present) occur in the “selection” and “control” populations. Each student must also produce figures that illustrate these findings, and that are presented and discussed in the lab paper.

With all of this in mind, the goals of this exercise are:

1. To analyze **your** *C. elegans* data and obtain the appropriate descriptive and comparative statistics for each generation,
2. To analyze the **pooled class** *C. elegans* data, for both the *selection* and *control* groups, and obtain the appropriate descriptive and comparative statistics for each generation,
3. To determine whether mobility speed of these *C. elegans* populations differ statistically between generations, and,
4. To produce figures which present these results in a clear, easy-to-interpret manner.

Before performing the data analysis described below, review the handout **Laboratory 3: Introduction to Statistics and Data Analysis** – available at the *Evolution* lab website at: <http://academic.bowdoin.edu/courses/s02/bio216/labs/statistics/index.shtml>.

Procedures

We will be using **Statview 5.0.1™**, available on the Macintosh lab computers in Druckenmiller 124, to perform all statistical analysis for this experiment. Before beginning your analyses, you must clearly formulate the **hypotheses** that this **experimental study** examines – recall the pre-lab discussion from the first week of the semester. Specifically, what did you *expect* to happen to the mobility speed of *C. elegans* populations in the **selection** group over the course of selection? What did you *expect* to happen in the **control** group?

Recall from the lab presentation on data analysis and statistics that each statistical test performed examines a **null hypothesis (H_0)** specific to that statistical examination. With each of the statistical tests you perform in this exercise, you should formulate a H_0 , and from the results of the comparative statistic determine whether you **reject** or **accept** that H_0 .

Using Statview 5.0.1 to analyze your Data

On the desktop of each lab computer is a folder entitled “**Evolution Student Work 02**”. Inside this folder is another folder (“**Selection Lab Data**”) that contains the data collected by each student over the course of the artificial selection experiment. There are three data files in the “Selection Lab Data” folder:

1. **SelectionData02(15min-student)**: This Statview file contains the finishing times for all *C. elegans* to cross the plate in the first 15 minutes of each generation, sorted alphabetically by student name,
2. **SelectionData02(top10-student)**: This Statview file contains the finishing times for the **10 fastest** *C. elegans* from each student’s experiment for each generation, sorted alphabetically by student name,
3. **SelectionData02(top100-total)**: This Statview file contains the finishing times of the **fastest** 100 control and **fastest** 100 selection worms (pooled over the entire class), sorted by generation.

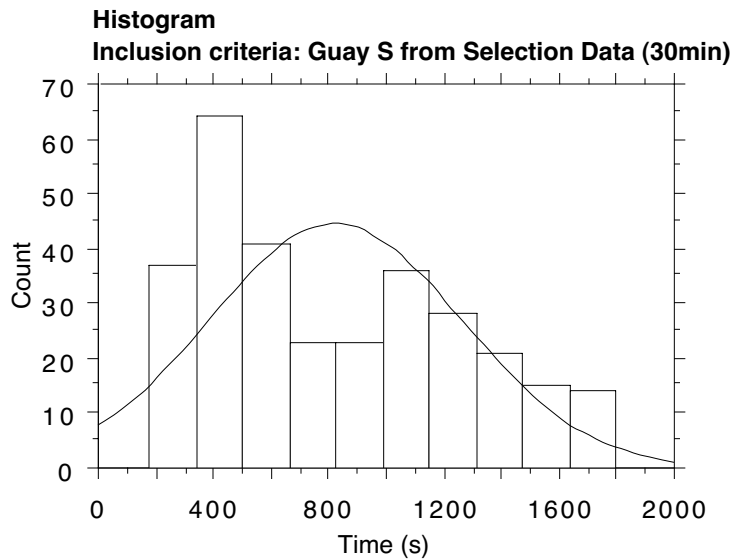
A. Statistical Analysis of Finish Times for *C. elegans* to Cross the Racetrack in 15 minutes:

Each student will begin by examining the populations of *C. elegans* crossing the racetracks in the first 15 minutes for each generation assessed. (Note: although the experimental protocol called for each student to measure race times for 45 minutes, due to experimental circumstances several students were only able to measure worm speed for 30 minutes. Thus, for the sake of consistency, we only examine worms crossing the racetrack in the first 15 minutes of the experiment for each student.) Open the file “**SelectionData02(15min-student)**” by double-clicking on the icon. Notice that there are columns indicating the student, generation, finishing time (s), and experimental condition (control vs. selection) for each worm.

In order to perform the statistical analyses, you must set the **Criteria** for each analysis. The “Criteria box” located at the top of Statview window is used to assign a subset of data that the computer will analyze. Start by analyzing *your own* data first. To do this, click on the “Criteria box” and scroll down to your last name. Each student’s name is followed by a C (indicating you were a member of the *control* group) or S (indicating you were a member of the *selection* group).

Once you have set the criteria to your data, perform the analysis outlined in the steps below:

1. In order to determine whether our data fit a normal distribution (and thus whether it is more appropriate to perform a *parametric* or a *nonparametric* statistic), under **Analyze**, go to **Graph ► Histogram**. A window will open asking you to “Assign Variables for ‘Histogram’”. Statview 5.0.1 is a drag-and-drop application - to assign a variable, one must drag the appropriate variable with the mouse and drop in the appropriate variables box. In this case, the variable you want to plot is **Race Time(s)**. After clicking OK, a graph similar to the following will appear:



In this graph, the line plot illustrates an *ideal normal distribution* for your data. The histogram shows the *actual distribution* of your data. From this graph, determine whether it is more appropriate to perform a *parametric* or a *nonparametric* statistic; if the *actual distribution* of your data mirrors the *ideal normal distribution*, it is appropriate to perform *parametric* statistics. If not, it is more appropriate to perform a *non-parametric* statistic. **Do not** close the Statview Viewing window!

- Analysis of Variance (ANOVA) – Parametric:** If you determine that your data fits a normal distribution, the ANOVA is the appropriate statistical analysis to perform – recall from the **Introduction to Statistics and Data Analysis** handout that the ANOVA is a comparative statistic used to compare between 3 or more groups (in our case, between the *parental*, *F2*, and *F4* generations). *Please note: All students are recommended to perform the ANOVA first regardless of whether your data fit a normal distribution – it is easier to see the general trends in your data using the ANOVA than the nonparametric test.*

To perform the ANOVA, under **Analyze**, go to **ANOVA and t-test ► ANOVA or ANCOVA**. You must then assign the variables for your analysis. In this case, **Race Time(s)** is the *dependent variable* and **Generation** is the *factor* (or, the *independent variable*). There is no covariate in this data set.

Before clicking OK, you should clearly formulate your H_0 for this statistical examination. In this case, the H_0 is that “there is no difference in the finishing times of *C. elegans* crossing the racetrack in 15 minutes between the parental, F2, and F4 generations”. Remember that *this* is the hypothetical model that the ANOVA is testing.

Once clicking OK, the following tables will appear in the viewing window:

ANOVA Table for Time (s)

Inclusion criteria: Guay S from Selection Data (30min)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Generation	2	1938995.604	969497.802	4.936	.0078	9.872	.814
Residual	313	61474686.355	196404.749				

The **p value** in the table is less than our $\alpha=0.05$, so therefore we *reject* the H_0 and conclude that there are some significant differences between the mean finishing times of parental, F2, and F4 worms.

Means Table for Time (s)

Effect: Generation

Inclusion criteria: Guay S from Selection Data (30min)

	Count	Mean	Std. Dev.	Std. Err.
Parental	62	745.790	337.867	42.909
F2	147	764.884	491.800	40.563
F4	107	594.374	425.581	41.142

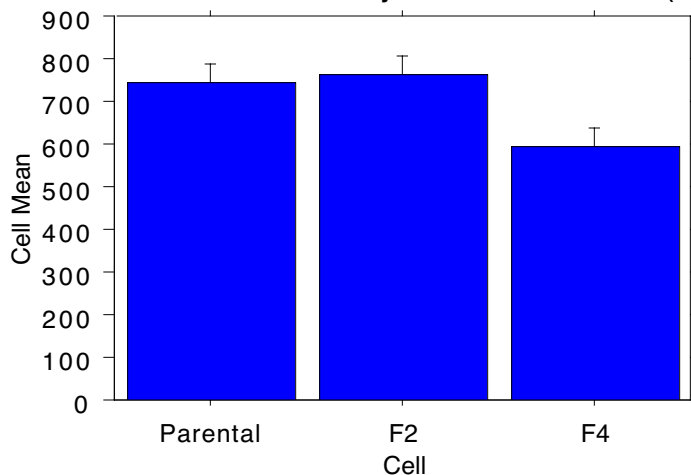
The table above gives the descriptive statistics (mean, standard deviation, and standard error) for each of the generations of selection analyzed. You should *write these numbers down* (or copy & paste them into a Microsoft Word document) – *you will need them to produce your figures.*

Interaction Bar Plot for Time (s)

Effect: Generation

Error Bars: ± 1 Standard Error(s)

Inclusion criteria: Guay S from Selection Data (30min)



Fisher's PLSD for Time (s)
Effect: Generation
Significance Level: 5 %
Inclusion criteria: Guay S from Selection Data (30min)

	Mean Diff.	Crit. Diff.	P-Value	
Parental, F2	-19.094	132.046	.7762	
Parental, F4	151.416	139.175	.0331	S
F2, F4	170.511	110.808	.0027	S

The *interaction bar plot for time*, and the *Fisher's PLSD* examination should be considered together. The Fisher's PLSD test is a *post hoc* examination, meaning that it is **only** applicable if the results of the ANOVA allow you to reject your H_0 . The Fisher PLSD test compares each possible pair of groups analyzed in the ANOVA individually in order to determine which groups *in particular* differ significantly from one another. In the case above, there is no significant difference in worm finishing time between the parental and F2 *C. elegans* populations, whereas there is a significant *decrease* in worm finishing time (i.e., the worms got faster) between the F2 and F4 generations, and between the parental and F4 generations. Note that significant differences are calculated at the 5% level, and are indicated by an "S" at the far left of each row in the Fisher PLSD table.

- Kruskal-Wallis Test – Nonparametric:** If you determine that your data do not fit a normal distribution, the Kruskal-Wallis test is the appropriate statistic to perform. Please review the **Introduction to Statistics and Data Analysis** handout for a brief description of how the Kruskal-Wallis test differs from the ANOVA.

To perform the Kruskal-Wallis test, under **Analyze**, go to **Nonparametrics** ► **Kruskal Wallis**. Assign the variables for your analysis. In this case, **Race Time(s)** is the *measured variable* and **Generation** is the ≥ 3 *group variable*. Again, before clicking OK, you should clearly formulate the H_0 that the Kruskal Wallis test will be examining.

Once clicking OK, the following tables will appear in the viewing window:

Kruskal-Wallis Test for Time (s)
Grouping Variable: Generation
Inclusion criteria: Guay S from Selection Data (30min)

DF	2
# Groups	3
# Ties	68
H	11.247
P-Value	.0036
H corrected for ties	11.247
Tied P-Value	.0036

Kruskal-Wallis Rank Info for Time (s)
Grouping Variable: Generation
Inclusion criteria: Guay S from Selection Data (30min)

	Count	Sum Ranks	Mean Rank
Parental	62	11096.500	178.976
F2	147	24517.000	166.782
F4	107	14472.500	135.257

The **Tied p value** (which corrects for ties in ranking procedure performed by this statistic) in the table above is also less than our $\alpha=0.05$, so we *reject* the H_0 and conclude that there are some significant differences between the finishing times of parental, F2, and F4 worms.

For the purposes of your paper, to determine the generations that differ significantly from one another, you will use the Fisher PLSD examination resulting from the ANOVA you performed earlier, regardless of whether your data set is parametric or non-parametric.

You may choose to print out the results of these statistical examinations. We also recommend that you save the statistical algorithms in the Statview viewing window – please save it in the “**Evolution Student Work 02**” folder under your name. Note that this will *not* save the actual results of the statistical examinations, but rather will save the algorithm used for these statistics. In order to view your results, one of the Statview data spreadsheets must be open, and the **criteria** must be set appropriately.

If you are going to continue immediately, *keep the Statview viewing window open.*

Once you have performed the statistical analyses for *your* worms, perform similar analyses for the *pooled class selection and control* data sets. In order to do this, you must change the **Criteria** set earlier. This can be done in three ways: (1) if conducting this analysis immediately after the steps above, under **Manage**, go to “**Edit/Apply Criteria . . .**”. Choose the new criteria you wish to apply (in this case, scroll down to **SELECTION Class Data**) and click on the “**Apply**” button. Return to the Statview viewing window, and you should notice that the statistical examinations now show the results for the pooled class *selection* group data. Again, you may print out or write out the important statistical results. (2) If you are starting fresh *and* you saved your Statview viewing window, open the appropriate Statview data sheet, and *your* saved Statview viewing window by double-clicking on both icons. Go to the “Criteria box” and scroll down to “**SELECTION Class Data**”, and the statistical examinations should show the results for the pooled class *selection* group data. (3) If you are starting fresh *and* you *did not* save your Statview viewing window, open the appropriate Statview data sheet, and in the “Criteria box” scroll down to “**SELECTION Class Data**”. Repeat the analysis steps 1-3 outlined above.

Once you have completed the statistical analysis of the pooled *selection* data set, change the **Criteria** to “**CONTROL Class Data**” and conduct the appropriate statistical analyses as described in the paragraph above.

B. Statistical Analysis of Finish Times of the 10 fastest *C. elegans* to Cross the Racetrack:

Each student will now examine their 10 fastest *C. elegans* crossing the racetracks for each generation assessed. Open the file “**SelectionData02(top10-student)**” by double-clicking on the icon and set the **Criteria** for the analysis. Again, start by analyzing *your own* data first. To do this, click on the “Criteria box” and scroll down to your last name.

Once you have set the criteria to *your* data, perform the analysis similar to the one performed earlier for worms crossing in 15 minutes (outlined in steps 1-3 in the pages above).

Similarly, once you have analyzed *your* worms, perform analyses for the *pooled class selection and control* data sets, following procedures similar to those outlined above. In this analysis, the pooled class data set includes the 10 fastest worms for each student pooled into either the *selection* or *control* groups; in other words, you are not examining the 10 fastest worms from the entire class, but rather a data set of 80 fast *selection* worms and 60 fast *control* worms.

C. Statistical Analysis of Finish Times of the 100 fastest *C. elegans* pooled over the entire class for each generation:

The final statistical examination you will perform is to look at the 100 fastest *from the entire class* to cross the finish line for both the *selection* and *control* groups. (Note that for this data set, you *will not* examine your individual data set, only the pooled data for the entire class). Open the file “**SelectionData02(top100-total)**”, and set the **criteria** to “**SELECTION Class Data**”. Perform statistical analyses similar to those performed above for the “**SelectionData02(15min-student)**” data set. Once you have completed that analysis, perform the same statistical examination on the *control* group – setting the criteria to “**CONTROL Class Data**”.

Using CricketGraph III to graphically illustrate your Results

Once you have performed the statistical analysis of your data, you must present these results in a way that is clear and easy to interpret. Often, the most effective way to present results is graphically in figures - a figure can often communicate information that would be difficult to describe in prose. Producing these figures can be quite a challenge, but can also be satisfying and fun! Illustrated below are suggested figures for presenting your artificial selection data using graphical program *CricketGraph III*. You may also choose to use *Microsoft Excel* or some other graphing program to produce figures.

Remember also that each figure you produce must have a **figure legend** that accurately describes for the reader what is presented in each figure. See **The Format of a Scientific Paper** for details about figure legends.

A. Entering Data into the CricketGraph III Spreadsheet.

In order to produce figures for your Artificial Selection Lab, a template **CricketGraph III** spreadsheet has been placed in the “**Evolution Student Work 02**” folder on the lab Macintosh computers: **Selection Graph Template**. Using the **means** and **standard errors** recorded from the analyses above, **CricketGraph III** can be used to illustrate this information. In order for **CricketGraph III** to plot these means and standard errors, both values must be entered into separate data columns in the spreadsheet.

The template spreadsheet is designed to contain ALL of the data needed to make ALL of your figures – there are 17 data columns in total. **Generation** is listed in the first column, and successive columns are organized into pairs: the first in the pair is for the **means** of a particular analysis, and the second column is for the associated **standard errors**. To use the **CricketGraph III** template spreadsheet, enter the data from your analysis into the appropriate columns. Once you have entered your data, it is very important that you **save** the data file. Please save your **CricketGraph III** spreadsheet under **your last name** in the “**Evolution Student Work 02**” folder.

B. Producing Figures with CricketGraph III.

For this report, you will be producing 4 figures:

1. One figure illustrating the *pooled class selection and control results* for worms crossing the plate each generation in 15 minutes on the same plot.
2. One figure illustrating the *pooled class selection and control results* for top 10 worms from each generation on the same plot.
3. One figure illustrating the *pooled class selection and control results* for top 100 worms from each generation on the same plot, and
4. One figure illustrating *your top 10 worms and the worms to finish in 15 minutes* from each generation – on the same plot.

To produce each of your figures, from your **CricketGraph III** data spreadsheet go to the **Graph** function and select **New Graph**. Choose the **Line** graph. Once you have done this, a window opens asking you to set up the axes of your graph. Use the following criteria:

1. For the **Horizontal (X) Axis** choose “**Generation**”,
2. For the **Vertical (Y) Axis**, while holding down the Apple Key, **⌘**, select:
 - For the figure illustrating *your top 10 worms and the worms to finish in 15 minutes* from each generation select “**MEAN Student 15min**”, and “**MEAN Student top10**”.
 - For the figure illustrating the *pooled class selection and control results* for worms crossing the plate each generation in 15 minutes select “**MEAN pooled Selection 15min**”, and “**MEAN pooled Control 15min**”.
 - For the figure illustrating the *pooled class selection and control results* for top 10 worms from each generation select “**MEAN pooled Selection top10**”, and “**MEAN pooled Control top10**”.
 - For the figure illustrating the *pooled class selection and control results* for top 100 worms from each generation select “**MEAN pooled Selection top100**”, and “**MEAN pooled Control top100**”.

3. Select the **Graph Format**. Click on this box and choose the “Worm Selection Figures” format.
4. Now click **OK**, and a figure will appear.

To improve the image of this figure, the labels on either axis can be altered by clicking on the **ABC** tool, and then clicking once on the text you wish to change. For example, when the graph first appears you will see “**MEAN Student 15min**” labeling the Y-axis. This label may not be appropriate - you may wish to change this to “**Mean Race Time(s)**” or simply “**Mean Time(s)**”

You may want to alter the increments used to label the Y-axis. To do this, double click on the axis itself and changing the **increment** to what you feel would be best. Similarly, you may also want to change the **minimum** or **maximum** value shown on the axis.

As for the rest of the figure, the design and look is up to you. You may prefer certain fonts, sizes, or styles that you can play with.

C. Putting error bars on your figures

Once you have your figure produced, you will need to put **error bars** to show the dispersion of each **mean** value on the figure. In order to do this, follow the procedure below:

1. In order to produce a figure with the appropriate error bars, go to the **Options** function and choose “**Y Error Bars . . .**”. A window will appear. At the top left-hand corner of the window is a box labeled “**Selected Series:**”. In order to place error bars on each mean value, click once on the value (i.e., “**MEAN Student 15min**”) in this box.
2. Make sure that **Both Errors** is shown in the window at the top right hand side. Go to the window labeled “**Source of Error:**.” Currently, the source of your error should be “**none**” - you want the standard errors that you entered into the **CricketGraph** spreadsheet. Click on “**none**” and select *your CricketGraph* data sheet as the source of error (this should be at the bottom of the window).
3. A list of columns will appear. Select the standard error (SE) data that corresponds to the **mean** value in the “**Selected Series**” window. Under the **Separation** option, set that number to 0. Do not click **OK** yet!
4. Now, click on the next **mean** in the “**Selected Series**” window. Repeat steps 2 and 3 above until all of the **means** have associated **SE** values.

D. Figure Legends.

The actual figure is only part of the figure as a whole. Any figure or table that you present must be clearly labeled (i.e. Figure 1.) and must be described by a **figure legend**, located below the figure. The figure legend should be a concise description of the results illustrated in the figure. It is important that a figure and its legend be able to “stand alone” and be easily interpretable to the reader. This means that the reader should be able to understand what is shown in the graph *without* having to read the results section.

Figure legends typically include a brief description of the results being presented and what the error bars represent (i.e., Error bars represent one standard error . . .). You may also want to **briefly** talk about any statistical differences that are found between any of your means.