Analyzing your Science Fair Data

Step 1: Determine your variables

Experiments are generally trying to determine if there is a relationship between two different variables. For example:

- Does playing music to your plant (jazz, classical or none) impact how much taller it gets in one week (measured in centimeters)?
- Does the type of praise given to a student after successfully completing a math problem ("you're smart" or "you're persistent") impact whether the student successfully completes the next problem (yes or no)?
- Does the amount of time a participant slept the previous night (measured in hours) impact their driving ability (accuracy score from a Driver's Ed simulator)?

What are the variables that you studied? You may have measured several different variables when conducting your experiment. For now, think about the two most important or most interesting variables. Then talk with a teacher or mentor about the ways you can incorporate your other variables into your analysis.

Step 2: Classify your variables

Each variable can be classified as either categorical or numeric. A categorical variable puts items (experimental units) into groups, while a numeric variable can take on any value in a reasonable range. For example, in the plant experiment, you would have three different groups of plants, one group that was exposed to jazz, one group that was exposed to classical music, and one group that was not exposed to music. This is a categorical variable. The amount of growth could be any value between 0 cm to 10 cm (maybe more?) so this is a numeric variable.

Don't be fooled by categorical variables that look like numbers. For example, if instead of playing music to your plants you decided to experiment with different amounts of water given to a plant each day. Some plants get 1 ounce of water each day, some plants get 5 ounces of water each day, and some plants get 10 ounces of water each day. The amount of water a plant receives cannot take on any value between 1 and 10, it can only take on the values 1, 5 or 10. This is a categorical variable.

Once you've classified each of your two variables as either categorical or numeric, your experiment will be one of the three following combinations:

- Two categorical variables (example: type of praise and success on next math problem)
- One categorical and one numeric variable (example: type of music and amount of growth)
- Two numeric variables (example: amount of sleep and score on driving test)

Each combination gets analyzed differently. Focus on the scenario that fits your data.

Step 3: Summarizing your data numerically and graphically

- Two categorical variables (example: type of praise and success on next math problem)

Display information in a two-way table or relative frequency two-way table and a side-by-side bar chart.

Two-way table

		Type of Praise		
		"You're smart"	"You're persistent"	Total
Success on next problem	Success	41	28	69
	Did not succeed	9	2	11
	Total	50	30	80

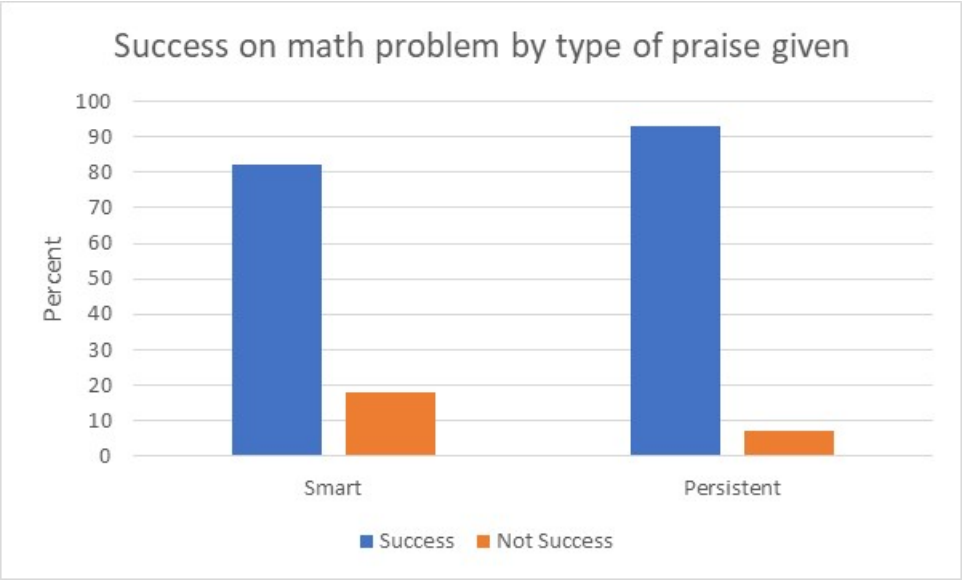
Relative Frequency Two-way table

		Type of Praise		
		"You're smart"	"You're persistent"	Total
Success on next problem	Success	41/50 = 82%	28/30 = 93%	69/80 = 86%
	Did not succeed	9/50 = 18%	2/30 = 7%	11/80 = 14%
	Total	100%	100%	100%

If the treatment groups ("you're smart" and "you're persistent) have the same number of individuals, you may consider using the two-way table with counts of individuals. In this example, there are different numbers of individuals receiving each type of praise, so a relative frequency two-way table is appropriate and allows for meaningful comparisons. Example "Overall, 86% of the students were able to successfully complete the next problem. However, while 93% of the students who were told that they're persistent were successful, only 82% of the students who were told that they're smart were successful."

You may decide whether or not to include the calculations (in grey above) in the table. If the table looks too busy, remove the calculations but be sure to include the number of participants in each treatment group somewhere prominent on your poster.

Side-by-side bar chart



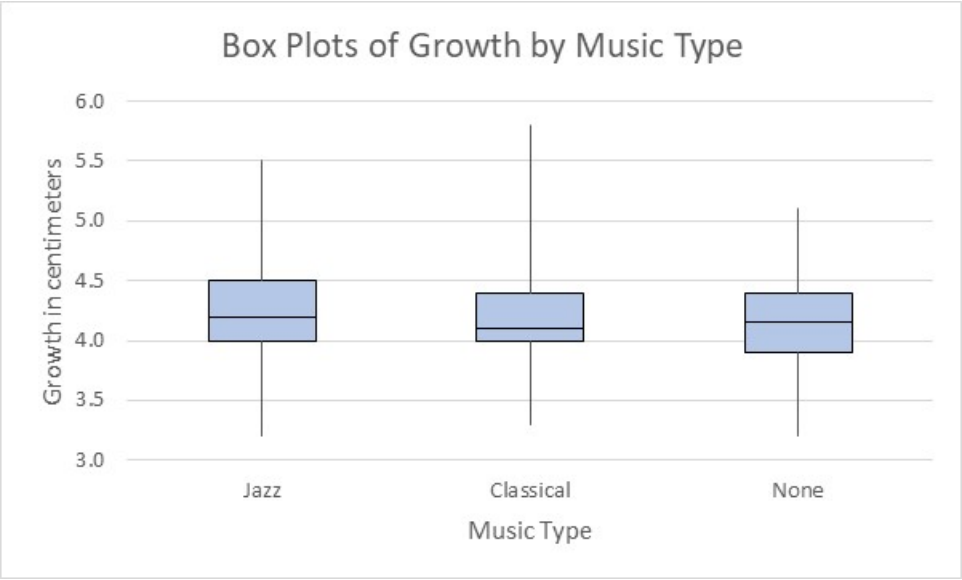
- One categorical and one numeric variable (example: type of music and amount of growth)

Summarize your data for each group

	Type of Music		
	Jazz	Classical	None
Number of Plants	28	30	25
Mean	4.23 cm	4.31 cm	4.18 cm
Standard deviation	0.56 cm	0.55 cm	0.53 cm
Minimum	3.2 cm	3.3 cm	3.2 cm
First Quartile	4.0 cm	4.0 cm	3.9 cm
Median	4.2 cm	4.1 cm	4.15 cm
Third Quartile	4.5 cm	4.4 cm	4.5 cm
Maximum	5.5 cm	5.8 cm	5.1 cm

In this experiment, plant heights were measured to the nearest tenth of a centimeter. In reporting means and standard deviations you may report one additional decimal place of accuracy than your measurement accuracy.

You may want to make a histogram for the data for each treatment group (jazz, classical, none). However, side-by-side box plots are an easier way to compare treatment groups at a glance on the same axes.



The graph above was created in Excel. Unfortunately, Excel does not have the functionality to create box plots, so the graph above was created by creating a stacked bar chart and then formatting each of the sections. Search 'box plot Excel' and you will find step-by-step instructions.

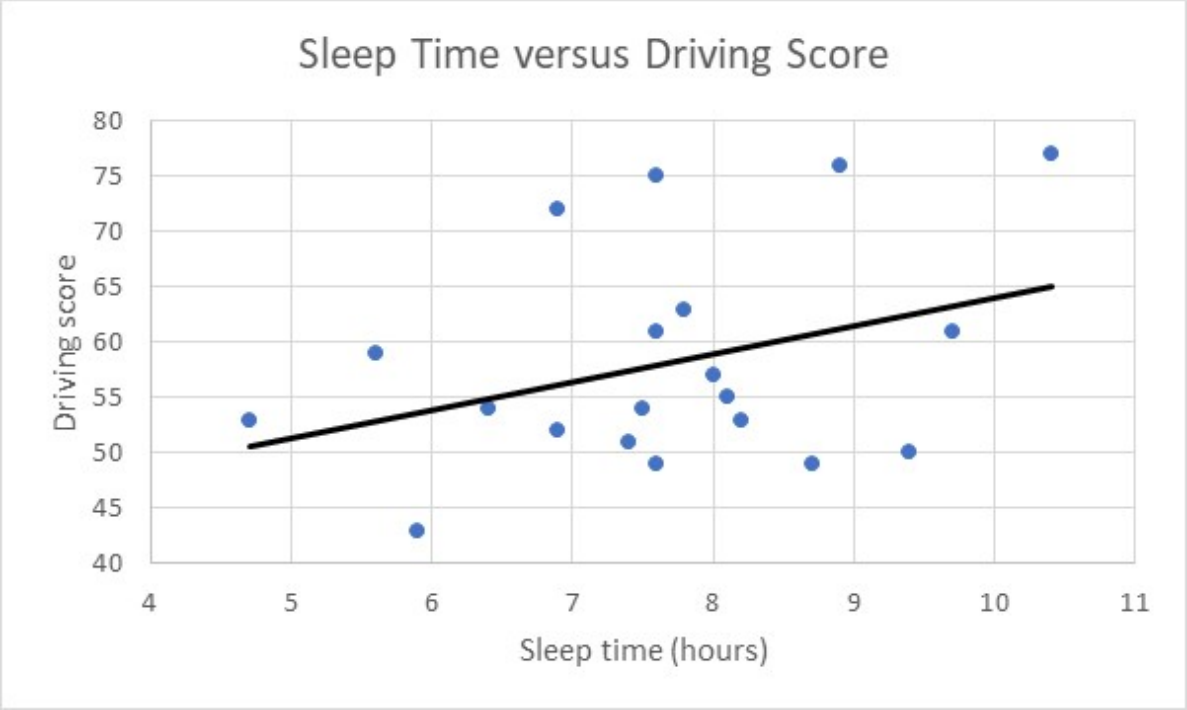
- Two numeric variables (example: amount of sleep and score on driving test)

You may want to summarize each of the variables separately. This does not address the relationship between the variables, but will give the judges some familiarity with the data.

Example: Twenty high school students were given a driving test in a simulator. The driving scores can range from 0 to 100. For our participants, the low was 43 and the high was 77 with a mean of 58.2 and a standard deviation of 9.8. After the driving test, students were asked what time they went to bed last night and what time they woke up this morning. Sleep time for each student was calculated. The low was 4.7 hours and the high was 10.4 hours with a mean of 7.67 hours and a standard deviation of 1.39 hours.

Note that the driving scores are given as whole numbers and don't have a unit of measurement. The mean and standard deviation for the driving scores have one additional decimal place than the data. Sleep time was recorded to the nearest tenth of an hour. Mean and standard deviation are reported with one additional decimal place. Sleep time is measured in hours, so all reported values about sleep are followed by the label of hours.

The appropriate graph for this type of data is a scatter plot.



Here the regression line is included. Summary statistics reported for a regression line are R^2 and the equation of the line. Example: There appears to be a weak linear relationship between sleep time and driving score with an R^2 value of 12.9%. The equation of the regression line is: predicted driving score = 2.5*(sleep time) + 38.8. This means that for each additional hour of sleep, the driving score is 2.5 points better, on average.

Be careful not to imply that additional sleep time causes the increased driving score. Perhaps students who are more risk adverse choose to give themselves more time to sleep and tend to be more careful drivers. So it's not the additional sleep that improves driving, but rather a student's risk adverse nature that causes both.

Step 4: Write your conclusion

In each scenario, you will discuss the relationship shown by the data.

- Students who are told "you're persistent" seem to be more likely to succeed on the next problem than those who are told "you're smart".
- Growth in plants seems similar regardless of whether they were exposed to jazz, classical music, or no music.
- There seems to be a relationship between sleep time and driving score. Longer sleep times were weakly correlated with higher driving scores.

If you randomly assigned treatments to your experimental units, you may make a conclusion about cause-and-effect. For example, if for each student you flipped a coin and If it landed heads you told them they were smart, and if it landed tails you told them they were persistent, then you can assess cause-and-effect. Both groups of students should have the same probability of successfully completing the next problem, the only difference was the praise they were just given.

If you did not control either variable, then the strongest statement you can make is that there is a relationship, or an association between the variables. See the sleep versus driving score example on the previous page.

Put on your conservative, legal-ese hat when writing your conclusion. You don't want to make too strong of a statement.

- First, most likely your project was not funded by a million-dollar National Science Foundation grant, but rather by the bank of mom and dad. So, your sample size is probably small-ish and your ability to control extraneous factors was limited. It's a good idea to admit the limitations of the study if asked by the judges.
- Second, even if you see very compelling results, there is a chance that those results are due to random chance. Perhaps if you were to select another 20 students and assessed their sleep time and driving score you'd see no relationship, or even a negative relationship.

Possible next step: Increasing the statistical sophistication of your data analysis.

Statistically, you can assess the probability that the relationship you saw in your data was due to chance. This process is called "Test of Hypothesis". Each different variable set up has a different type of test that is appropriate.

- Two categorical variables (example: type of praise and success on next math problem)
  - If each variable has only two options, you can perform a two-proportion Z-test
  - Given any number of options for either variable, you can perform a chi-squared test of independence
- One categorical and one numeric variable (example: type of music and amount of growth)
  - If your categorical variable has only 2 categories, you can perform a two-sample t-test
  - Regardless of the number of categorical options, you can perform an ANOVA test
- Two numeric variables (example: amount of sleep and score on driving test)
  - You can perform a test of hypothesis on the slope of the regression line (if the slope is 0, then there is no relationship between the two variables)

For more information contact a statistics teacher at your school.