Statistics for Science Fair Cheat Sheet

Statistics is a scientist's powerful ally. Used properly, statistics allows your students to interpret the results of their experiments and report conclusions with measured confidence. Statistics shouldn't be scary—in fact, the basic ideas are quite simple. It's the details that get messy. This handout goes a bit beyond the basics and seeks to pack a lot of information into a tiny space. **Applying statistical tools will dramatically increase the quality of your students' science fair projects.** Here is a map to becoming a science fair statistical sleuth.

1. **It starts with experimental design**. Statistics can't help your students if their data isn't any good. That's why you should begin with the end in mind: think statistics from the start. Keep in mind these principles of experimental design:

    a. **Control.** To draw conclusions, we need to control, to the best of our ability, all of the things that we can. Our goal is to make it so that the only difference between our experimental units is the independent variable.
    b. **Replication.** Broadly speaking, if you have less than 15 replicates, you probably aren't ready for statistical analysis; if you have at least 15 replicates, you might be in the clear. It is best to have *at least* 30 replicates.
    c. **Randomization.** This is the principle we are often least familiar with, but it may be the most important. Letting unbiased chance, such as a penny, a dice, or a table of random numbers, do the picking for us is essential if we are going to let the power of statistics work for us. If your project allows it, consider a matched pairs design.
    d. **Flowcharts and diagrams** are helpful tools for conveying your experimental design.

2. **We snoop around a bit by doing exploratory data analysis (EDA)**. Neglecting EDA and skipping straight to inference is a quick way to make a fool of yourself. Start with graphs and then use numerical measures to characterize the S.O.C.S. of your data: spread, outliers, center, and shape.

    a. For **categorical variables**, choose from bar graphs, pie charts, and two-way tables.
    b. For **quantitative variables**, choose from stemplots, histograms, relative cumulative frequency plots, and timeplots. Scatterplots are most useful for comparing relationships among quantitative variables. Remember that bar graphs and histograms are two different things: bar graphs are for categorical variables, and histograms are for quantitative variables.
    c. Calculate and compare the mean ("average value") and the median ("typical value"). Determine the standard deviation. **Always report a measure of center with a measure of spread.**
    d. Ask yourself: Is the distribution skewed or symmetric? Unimodal, bimodal, or multimodal? Are there outliers?
    e. Start with graphs, proceed to numbers, and make a preliminary interpretation of the data.

3. **We start to close in on the story by using statistical inference**. Valid inference depends on appropriate data production, skilful EDA, and the use of probability. When you use statistical inference, you are acting as if the data come from a *randomized* experiment, which is one of the reasons why randomization is such an important part of experimental design. One of the big ideas of inference is the *p*-value. The ***p*-value** is the probability that the observed result is due to chance. It is the probability that, from a randomized, controlled experiment, the null hypothesis is correct. Whenever doing an inference procedure, always remember to specify your null hypothesis, H_0, and your alternative hypothesis, H_a. **We can consider three basic models for analyzing science fair projects.**

    a. **The relationship between two quantitative variables.** If a student is looking at the relationship between two variables and used multiple levels of those variables, **a scatterplot and regression analysis** is probably a good analysis framework. Here are some tips:
        i. Plot the independent variable on the x-axis. Look for the overall pattern of the scatterplot and for deviations from that pattern. Discuss the direction, form, and strength of the pattern.
        ii. If the pattern appears to be linear, calculate a correlation coefficient, **r**, which measures the strength and direction of the relationship between two quantitative variables.
        iii. Use least squares regression to determine a mathematical model of the relationship between the two variables. Be sure to look at a residual plot; there should be no systematic pattern to the plot.

        iv. Consider performing inference about the regression slope. This is a *t*-test with the null hypothesis that there is no linear relationship. See section B (below) for more on *t*-tests.
        v. Least squares regression and the correlation coefficient can only be used for data with a linear relationship. If the relationship is not linear, more advanced methods can be used to transform the relationship so least squares regression can be used.

    b. **Comparing data from two different groups.** If your student is examining differences between two groups of data, then an examination of **boxplots and a *t*-test** is probably the statistical path of choice.
        i. Plot the data using a side-by-side boxplot. Based on the plots, ask, "Does there appear to be a difference between these groups?" "Does the experimental group seem to be larger or smaller than the control group?"
        ii. Based on the context of the project, determine the alternative hypothesis, H_a. Do you think that the experimental group has a smaller mean than the control group? That the experimental group has a larger mean than the control group? Or simply that the groups are different from one another?
        iii. Define a level of significance (e.g. $P< 0.05$). Carry out the two-sample *t*-test, examine the P-value, and determine if the difference between the groups is statistically significant.
        iv. Evaluate the practical significance of the difference between the groups.
        v. Report the results of the inferential analysis in the context of your experiment.
        vi. If you are trying to determine if a set of data is different from a specific value, then the *t*-procedures are probably still the statistical tool of choice. However, instead of using a two-sample *t*-test, use a one-sample *t*-test. Also, if your student used a matched pairs design, the *t*-procedures are slightly different; consult one of the references below for details.
        vii. When comparing differences between more than two groups, ANOVA may be a more appropriate choice. See one of the references below for details.

    c. **Inference for categorical variables.** If a student is doing a genetics project, they are probably interested in knowing whether the results of their experiment match the expected distribution of phenotypes or genotypes. Other projects that have categorical variables that are expected to take a specific distribution can also be analyzed with this technique.
        i. Start with a two-way table. Calculate marginal distributions and determine the differences between the marginal distributions of the experimental and control group(s).
        ii. If possible, use a bar graph or pie chart to show the distribution differences among the various groups. Remember that bar graphs are often easier to make and understand.
        iii. Use a chi-square test for goodness of fit. Determine the test statistic, X^2, and the *p*-value.
        iv. If the chi-square test finds a significant result, examine the distribution to find the largest components of the chi-square statistic.

4. **We combine the results of exploratory data analysis and inferential analysis with our analytical intuition to figure out what the data are telling us**.

5. **We clearly present our results using the language of statistics**.
    a. State the statistical hypothesis along with your scientific hypothesis in the hypothesis section of your presentation/paper. Be sure to express the reasoning behind the hypothesis.
    b. Use a flowchart to show the experimental design. In the methods/procedures section, point out how replication, control, and randomization were utilized.
    c. Show both exploratory data analysis and inferential analysis in the data analysis section. Discuss the meaning of graphs and numerical measures. State the level of significance used in your tests (e.g. $P <0.05$). State the conclusion of the significance test.
    d. State the statistical and practical significance of the results in the conclusions section. Be sure to state whether the null hypothesis was accepted or rejected.

Suggested References:

The Cartoon Guide to Statistics by Larry Gonick and Woollcott Smith, ISBN 9780062731029
Cliff Notes Quick Review: Statistics by David Voelker, Peter Orton, and Scott Adams, ISBN 9780764563881
The Practice of Statistics, 3rd ed. by Daniel Yates, David Moore, and Daren Starnes, ISBN 9780716773092, www.whfreeman.com/tps3e.