

# Letting the Numbers Talk

Data presentation is essential to attract interest and attention to your project. In this guide, we will be giving hints on how to do this in a number of ways. The key: visual graphics!

Sometimes, only one thing separates a science-fair project from excellence: the lack of simple statistical analysis. From the perspective of judges, good data analysis is important in science fair projects. Indeed, there is a 35 point section in the judging rubric that mentions "analysis of results using suitable graphs and statistics".

## Contents

1. Creating effective graphs
2. Different graphs for different purposes
3. Getting started with stats
4. More than just the "mean"

## Creating Effective Graphs

In science, graphs are one of the main methods of presenting information effectively and concisely. Graphs usually represent quantitative results from your experiments in a visual way. A few graphs on your board will not only reduce the amount of text cluttering your board, it can also be visually appealing to your audience. But like many other things, don't overdo it! Keep the number of graphs to a minimum.

Every graph should have:

- A **TITLE** that is easily located. The title should tell your audience what information the graph is trying to convey.
- **LABELS** on the axis, or in the form of a legend. A legend is a separate box near or on an empty part of your graph that clearly shows what each part of the graph represents, if there are multiple sections or sets of information.
- **NUMBERS**: remember, your graph is representing quantitative information. Your graph should have numbers labeled on the axis, or in the case of pie charts, near each section.

There are numerous types of graphs, and depending on what you use it for, there is often one type of graph that is the most effective and the most easily decoded by your audience for your purpose. Here, I will list the most commonly used types of graphs.

## **Pie Chart**

The purpose of a pie chart is to present information that is divided into different categories. It should be ordered from large to small sections, and should always have a legend to tell the audience what each of the sections represent. Its shape, a circle, offers a visual sense of a whole (100%) of something.

## **Bar Graphs**

When you have only a few data points (10 or less), and want to visually emphasize the size of your numerical data (by the height of a bar), the bar graph would be the right type of graph to use. Bar graphs are also handy when you want to compare two different sets of data, because bar graphs allow you to plot more than one data set on one graph. It is important to label your axis and their respective increments, as well as a legend if there are more than one set of data represented.

## **Line Graphs**

On a line graph, each value is represented by a data point. Line graphs are used when there are many data points, where each data point can be connected by a line, or a line of best fit can be drawn to show a trend. Like bar graphs, line graphs can have multiple sets of information, i.e. multiple lines on one graph.

## **Standard Deviation**

Even if you try very hard to make your experiments precise, there is always a possibility of your data being not as precise as you hoped. This is accounted for by standard deviation. Standard deviation is the measure of dispersion of a data set. Sometimes, you are expected to include the standard deviation on your graph. They are represented by a small bar protruding from each data point (refer to figure 2, the bar graph). The calculation of the value of the standard deviation can be complicated; but do not worry! The Excel program can do all that for you.

## **Different graphs for different purposes**

Once the data has been acquired and analyzed, it is time to present it, hopefully in the most attractive and perceptive way! Deciding on the format to use becomes easier when you know that there are specific presentation tools for each type of data:

1. Charts displaying frequencies,
2. Charts displaying distributions,

### 3. Charts displaying correlations.

I want to show:	I should use:	What type of data do I need?
Frequency of occurrence (how often), Comparison of magnitude (how much or how strongly)	<ul style="list-style-type: none"> <li>• Bar chart</li> <li>• Data table</li> <li>• Pareto chart</li> </ul>	Categorized data, data tallied in subdivision
Distribution of observations (how much variety, over what range)	<ul style="list-style-type: none"> <li>• Histogram</li> </ul>	Several measurements (repeats) of a single experimental quantity, i.e. all parameters are identical between each measurement (see Replication)
Correlation between parameters (is one variable related or affected by another)	<ul style="list-style-type: none"> <li>• Run chart (Trend over time)</li> <li>• Map (Correlation in space)</li> <li>• Scatter diagram</li> </ul>	Measurements of an experimental quantity, the dependent variable while varying some parameter of interest, the independent variable. Repeats, i.e. keeping the independent variable fixed and acquiring several observations, can be included. For a run chart, the independent variable is time; for a map, it is position. (See knowing the variables: Control!)

#### Representing frequency of occurrence

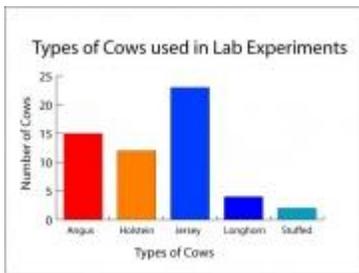
##### Data Table

Tables convey precise numerical information whereas graphs are more helpful to grasp the bigger picture. In most cases, graphs will make it easier to interpret your data.

However, in some occasions, tables may be preferred over graphs:

- Each value can be labeled or described easily, for instance to represent data from various sources,
- Values can be grouped in categories,
- Incomplete or disparate data sets are more accurately represented,
- The difference in order of magnitude can be better appreciated, e.g. to represent large, intermediate and small values at the same time.

Nevertheless, it is a good idea to include both tables and graphs of your raw data and results in the Data section of your report. Here is a good example of a table.



No pun intended, the Periodic Table is a very good example of arranging data by characteristic, and category. A graphical representation of all the elements would be very confusing and not very useful. Tables are beautiful in that we can see specific parameters of a given variable arranged neatly (Element is our variable, and the specifics are atomic mass, number, number of orbitals, etc.)

## Bar Chart

Bar charts present results that compare different groups. Bar charts work best when showing comparisons among categories. They are also the best way of showing "how much" or "how often" one category is affected in relation to another, depending on what you are trying to show. For example, if I wanted to show how many students get As after drinking Gatorade for the whole semester, compared to those that don't, a bar chart will clearly show the difference.

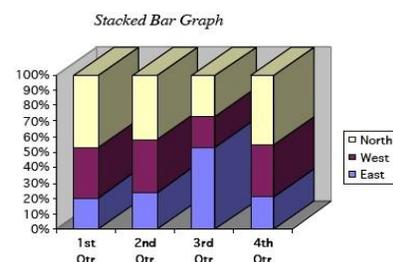
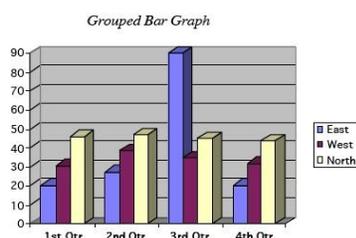
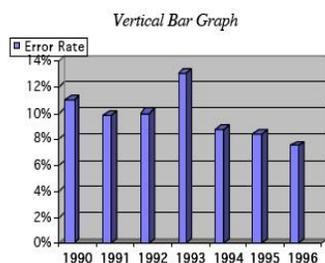
There are three different bar chart types that are used in different situations. 'Simple bar charts', 'Grouped bar charts' and Stacked Bar Charts'.

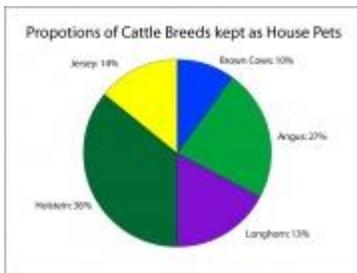
Simple bar charts sort data into simple categories. This is the type that you are most familiar with. Use vertical or horizontal bars; horizontal bars look better when long labels are being used.

Grouped bar charts divide data into groups within each category and show comparisons between individual groups as well as between categories. The groups within the categories have the bars "touching". For example, say I apply drug X to normal cells and cancer cells, and then apply "placebo drug" to normal cells and cancer cells. I would likely sort the bars of "Drug X treated" into one category (have them touch). I would also sort the bars of "placebo drug treated" into another category (have these bars touch on the sides). Another example is if you are comparing rates of smoking in different age groups in 2005 and 1995. You could group 2005 and 1995 bars touching, and then separate the groups based on different age groups.

Stacked bar charts show related groups one on top of each other. The components of each bar in a stacked bar chart should show each component as a fraction (or percentage) of the whole. It is recommended that you consider using stacked bar charts rather than pie charts if you need to compare more than three sets of data.

Here are a few examples courtesy of Concordia University.





**Whenever possible, use bar or pie charts to support data interpretation. Do not assume that results or points are so clear and obvious that a chart is not needed for clarity.**

### **Pie Chart**

The idea behind the pie chart is that many "pieces of pie" represent the relative proportion of various items in making up the whole quantity, i.e. how the "pie" is divided up. It is recommended that you do not to use pie charts for the following reasons:

- Pie charts take up a lot of space for the information they represent,
- The "pieces of pie" are not consistent in the way they are arranged spatially,
- In some occasions, it is difficult for the eye to gauge the difference between pieces of pie.

However, if you would like to use a pie chart, make sure you use it when a certain fraction is overwhelming small or large when compared to another. That is to say, pie charts are quick ways of representing two quantities if it is easy to determine that one is significantly different than the other.

Instead, the same information can be represented in the form of a bar chart or a table. If you decide nevertheless to use pie charts, be careful not to use too many notations. Keep it as simple as possible, only including enough information necessary to interpret the chart.

### **Representing distribution of measurements**

#### **Histogram**

Histograms, unlike bar charts, are used to display numerical/quantitative variables. For example, say you are measuring the pulse rate of 50 women to see how it differs between them.

Imagine using a bar chart to display this data - you would have 50 bars! This is ridiculous. We cannot do this because here we are looking at numerical data. Using a histogram, we slice up the entire span of values covered by the quantitative variable into equal-width piles called bins. Then we count the number of values that fall into each bin. The bin and the counts in each bin give the distribution of the quantitative variable. So for our pulse experiment, we would divide the X-axis into bins showing ranges of pulse rate. ie: Bin 1 would be 55 - 60, Bin 2 would be 60 - 65, Bin 3 would be 65 - 70, etc. Then we total the number of counts in each bin. This is the vertical height of each bin.

From this, we can see the distribution of the pulse rate among the 50 women. We might expect it to have a bell-shape.

## Representing correlation between variables

### Map

When thinking about a map, one might have in mind various city or countries with their geography represented in relation to others. Maps describe the position and shape of objects in space. You can take the concept to your advantage and use maps to convey spatial information about your results.

Numerical or qualitative information can be added on top of the spatial information on a map, for example:

- The amount of contaminant in soil samples can be mapped to their location of origin,
- Important tumour suppressor genes can be mapped on the 1D map of human chromosome 1.

This additional information can be carried by various visual effects; use the one which works best:

- Important sites and values can be indicated with labels and arrows,
- A variable can be represented at different sites in a bubble graph, the size of the bubble being related to the value of the variable at this site,
- The map can take different colors depending on the value of a variable; this is called a heat map. In this case, a color legend must be included.

Maps can easily represent information in 1D and 2D. Three-dimensional (3D) maps or renditions of objects can also be prepared, but note that, for the human eye, it is difficult to perceive perspective as represented on a flat surface such as the science fair board. If you have data with more than two dimensions (i.e. 3D, 4D, etc), you should seek an alternative way of representing it, for example:

- A panel with a series of 2D maps,
- A computer animation or an interactive 3D model,
- A small-scale mock-up model that you keep at your stand!

### Scatter Diagram

Scatter diagrams, also called Scatterplots, may be the most common and most effective display for data. By looking at them you can see patterns, trends, associations, relationships and outliers. Scatter plots are observing the relationship between two quantitative variables. For example, we could do a scatterplot of "year" and "expenses of the Enron Corporation". This is a special type of scatter plot called a time plot.

Scatter diagrams relate two quantitative variables and ask whether there is an association between them. Are grades higher now than they used to be? Are height and weight of a person related? Is the cost of traffic congestion per person related to the peak speed on the freeway? How does the number of cells vary over time with a particular treatment? Is the time spent at the dinner table related to how many calories you burn? Is the speed of the roller coaster related to the drop?

Note: Just because a relation "exists" between two quantities, it's really up to you to make sure that that relationship is logical, or meaningful. For example there might be, by coincidence, a remarkable relationship between ice cream prices and the weather in Timbuktu. But this doesn't tell us anything about either variable, because we know that in the real world, one did not cause the other.

## Network Diagrams

Network diagrams are instrumental in projects that wish to plot how one person might feature in an interconnected web of relationships. Examples where this representation might be useful include plotting the relationship one gene's expression might have over other genes in the same area of a certain chromosome; determining what words are most likely to occur in a reader's mind when subject to a certain kind of stimulus;

At first glance these diagrams can appear messy and convoluted, but they are the only types that will effectively capture in the interconnectedness of any given situation, provided that's what you want to highlight.

## Annotating Graphs

We will discuss three major components of graph annotation: labeling, making notes, and effective correlations.

### Labeling

Any type of graph has the sole aim of representing what is most valuable about your raw data, without actually writing each value down in a table, ad nauseam. That being said, what you label on your graph and how you choose to title it are both important components of the graph itself.

You can consider labeling the maximum and minimum values on your graph, if either is crucial to your experiment. You can also consider labeling specific intervals on the graph to indicate where a medium was replenish, where certain conditions were changed, etc. These sorts of labels should be concise and kept to a minimum of two to four words. Use a dialogue box type label, in which an arrow indicates the point of interest and the text box provides the information about that point.

Your graph's title should tell the reader in one phrase what this graph has tested, and what are the 'axes' involved. For example, in a project that constructed a fuel cell and tested its ability to generate hydrogen gas over time, one graph could possibly be titled "Measurement of H<sub>2</sub> Yield over Time" or "Fuel Cell Performance at 20, 30 and 40C." Both these titles immediately give the reader information as to what was measured, and in what conditions, or against what setting (Time, Temperature, Location, etc.).

Make sure that your labels and your title only add to the importance and informative nature of your graph.

## **Making Notes**

It is customary to include a few sentences as a caption to all your graphs, providing a brief analysis of the data above and explaining their significance. This can serve as a replacement or an addition to labeling, but usually both complement each other nicely. These few sentences shouldn't conclude anything about the result of the given experiment, but rather should explain what is going on, and what might have caused certain anomalies. You may be wondering how this is different from a conclusion, and to answer that concern, consider this example:

Imagine a project wishes to examine how long a microbial fuel cell can last without replenishing any nutrients for the bacteria present. A graph might show steady output of current and a few hours later, the curve will taper off towards zero. In the notes below the graph, one might write that nutrients were not replenished in this experiment, explaining the reasons for the curve dropping off. However, making a conclusion that nutrients need to be replenished every few hours to keep the current output steady is not to be done in these notes.

Notes can also include any specific information about the nature of the experiment, how many times it was repeated or specific concentrations used that were otherwise unspecified in the title or the labels.

## **Correlations**

In addition to what we have already told you about correlations, here are some more tips on representing correlations properly on your graph.

In regression curves, trend lines can be a very quick means of assessing the quality of a relationship between two parameters. The R-squared value approaches 1 as the relation becomes more and more evident. You can fit data to linear, exponential, logarithmic or power based regression best-fit trends, depending on your own inferences as to how the two parameters should be related.

On Excel or JMP an equation for the correlation might be calculated empirically by the program, and given along with the R-squared value. The equation might not be necessary unless you plan to extrapolate data that you didn't or couldn't measure. Usually the R-squared value is a label you should include in any correlation graph.

## **Getting Started with Stats**

Data presentation is essential to attract interest and attention to your project. The easier way to communicate an idea is through a picture, be it a bar-graph, flow-chart, histogram, etc. In this guide, we will be giving hints on how to do this effectively.

We will also be investigating Statistics. Whenever there is data and a need for understanding the

world, you need Statistics. Statistics help in handling variation in the data, reducing and summarizing data, interpreting data and drawing appropriate conclusions. Some questions that we can try to answer with statistics are: Is the global temperature increasing? Does vitamin C really prevent heart disease? Is a new engineering technique significantly better than the old one? We apply statistics to these questions to give us answers.

Sometimes, only one thing separates a science-fair project from excellence: the lack of simple statistical analysis. From the perspective of judges, good data analysis is important in science fair projects. Indeed, there is a 35 point section in the judging rubric that mentions "analysis of results using suitable graphs and statistics". Using stats can help you to determine the accuracy and clarity of your results. It can also provide "trustworthiness" to your conclusions.

On this page we won't be going through statistical methods, simply ideas that you should keep in mind when reading other sections.

One of the basic tenants of the scientific method is the idea of reproducibility: Will the same results be seen when the experiments are repeated? Scientists run the same experiments a number of times to see if the results are reproducible. However, there will always be some variation whenever experiments are repeated, due to experimental error. Statistical thinking helps determine whether differences observed between groups (e.g., control and treatment) are real or are just due to experimental error. Through the scientific method, a hypothesis proposes a model for your experiment. Then we look at the data. Is the data consistent with the model? Does it lend support to your hypothesis, or does it disprove the hypothesis? And what happens if the data is only \*slightly\* inconsistent with the model? What are the limits for deciding whether your hypothesis is right? Statistics allow us to establish a degree of "acceptance" or rejection". Mathematics is also useful in assessing the effects of measurement errors and other uncertainties in an experiment.

In the following guide, you will become well versed with natural variation of measurements, adequate sample sizes, graphing raw data properly for visual inspection, elimination of outliers, calculation of standard deviation and standard error of the data, and hypothesis testing. We will introduce you to Z-tests, T-tests, Chi Square, and when to use each. We will also go through normal distribution, linear regression, least-squares curve fit, and much more. The key is that statistics are used to get results and transition from the raw data to the conclusions.

Two types of statistics will be discussed:

1. Descriptive Statistics are ways of using statistics to describe data, through plots and calculations such as average, standard deviation, IQR, etc.
2. Inferential statistics are ways of inferring conclusions from the data. This includes the hypothesis test. This is the most powerful way that you can improve the quality of your science fair project.

## More than Just the Mean



Being able to describe a set of numbers may well be easy: you can just take the mean.

But there's more information than the mean that you want to talk about. Take these two sets of numbers:

{0, 1, -1} and {0, 1000, -1000}.

Both have the same mean, but they're very different sets of numbers. How do you distinguish between the two? Read on!

### Central tendency

Let's say we have a set of values all associated with a given quantity. A first goal of statistics is to be able to represent an "average" value for this quantity. An average summarizes a data set in a single value and provides information about the magnitude, the sign and the units of measured values.

Multiple samples require a tool that makes a comparison both easy and quick; central tendency does just that! Statistics provide you with various descriptors for the average: median, mode and mean among others. However, you must be careful as they represent your data in very different ways.

If you sort the set in increasing order, giving an index to each value, the median and mode can be found.

- The median is the value whose index lies in the middle of the set, namely half the values are below it and half above it. If the number of values  $n$  is odd, the median is the middle value. Counting in from the ends, we find this value in the  $\frac{n+1}{2}$  position. When  $n$  is even, there are two middle values. So, in this case, the median is the average of the two values in positions  $\frac{n}{2}$  and  $\frac{n}{2} + 1$ .
- The mode is simply the most frequent value(s). For example, the mode of {1,1,3,3,3,4,7} is 3, while the modes of {1,2,2,4,4,5} are 2 and 4.

Probably the most commonly used central tendency measure is the arithmetic mean:

- The mean is calculated by determining the sum of all values and dividing by the total numbers of values,  $n$ . Formally, for a set of numbers  $X = \{X_1, X_2, \dots, X_n\}$  the mean is  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ .

## IMPORTANT INFORMATION

A common mistake is to confuse the terms mean and average; while the term mean mainly refers to

the arithmetic mean (see above), average encompasses all measures of central tendency.

Other more specialized tools might be of interest for your project:

- The midrange is the mean of the maximum and the minimum values. It only summarizes the two extreme values in a set and can be biased if those are outliers.
- The weighted mean is a mean which is computed to give more importance to some data in contrast to other. This can be useful if you have a good reason to believe that some data are more reliable than other, for instance their uncertainty or dispersion is lower.

### Data dispersion

Two experiments are carried. For one experiment, values are all similar; for the other, values are very different. A question that statistics can answer is: how different are the many values with respect to each other?

The fact that values are not all the same, and the extent to which they differ, is called dispersion.

An important question, which can help you with your research, is: why is it that values are different, for the same experiment?

Common sources of dispersion are:

- Observations are drawn from a sample that is heterogeneous,
- Variation due to the random nature of the variable being observed,
- Errors and uncertainties in the measurements.

Consequently, data dispersion provides two types of information:

- Information about the variable(s) being observed, and
- Information about the quality of the methodology.

It is often difficult to separate the two, you have to use your judgment to decide on the probable cause of dispersion. Good projects try and find solid grounds to justify this kind of decisions.

Let's have a look at various methods that describe the dispersion of a data set: range, and standard deviation.

- The observed range is the difference between the maximum and the minimum observed (experimental) values. Since it does rely on only two extreme values, it conveys only scarce information. However, one can make arguments about the extent of dispersion and the precision of a method when comparing it to the potential range that is the difference between the maximal and minimal possibly observable values.
- The standard deviation takes into account how far each value is from the mean. Therefore it is very useful in thinking about how widely spread the values in a data set are. If many data points are close to the mean, then the standard deviation is small; if many data points are far

from the mean, then the standard deviation is large. If all the data values are equal, then the standard deviation is zero. How do we calculate standard deviation?

One way to think about spread is to examine how far each data value is from the mean. This difference is called a deviation. We first square the deviations and then average them. We square them to keep them from canceling out (ie: positive and negative deviations). Squaring also emphasizes larger differences. When we add up these squared deviations and find their average, we call the result the variance.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

The variance is not the ideal measure of spread because its data are in squared units. So we take the square root of  $s^2$ . The result,  $s$ , is the standard deviation. Formally, the standard deviation is the root mean square (RMS) deviation of values from their arithmetic mean. All together then, the standard deviation of the data is found by the following formula:

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

### Skewness

It can sometimes happen that the data is not distributed equally on both sides of the median. It is then said that the distribution of numbers is skewed, or asymmetric.

Skewness can be useful in analyzing data that is not distributed normally. Many statistical models just assume that data is distributed normally about the mean, but unless the skewness value is zero, this is not the case.

Skewness deals primarily with tail or snake values in a bar chart. Sometimes, values on either side of the bar chart will taper off in a manner that is different from a regular bell curve (one side tapers steeply compared to the other, etc.). It is for this reason that measuring skewness can lend more credit and meaning to your analysis.

There are two types of skewness negative skewness and positive skewness. Negative skewness refers to a tail that is longer towards the left. The majority of values are concentrated to the right in this case. A good example would be a bar chart with values {1, 2, 500, 700, 800, 1100, 1400}. It is very easy to see that this distribution has very few low values. Whether this type of distribution is intended or not, it is important to note that in this case the mean is lower than the median which in turn is lower than the mode. In a regular Bell Curve, all three quantities are equal.

Positive skewness is the exact opposite of negative skewness. Most values are low, and concentrated to the left of the distribution. The mean will be greater than the median, which will be greater than the mode.