Data Analysis

Planning and performing data analysis is an important part of the scientific process. It is a tool that allows our reasoning to be more objective and less biased. Having a foundation in data analysis along with an understanding basic statistics and hypothesis testing will help you in multiple disciplines and allow you to critically evaluate science.

**Summary statistics**

Whenever you gather data, you need to calculate summary statistics to describe the individual dataset. These stats will help condense information from many pieces of data.

**The mean**

The mean ($\overbar{x})$ is the average of a set of observations. It is a measure of central tendency telling you where most of the values occur. Calculate the mean by summing ($Σ)$ the individual observations ($x\_{i}$) and dividing by the total number of observations ($n$)

$$\overbar{x}= \frac{Σx\_{i}}{n}$$

**The sample variance**

The sample variance ($S^{2}$) is a measure of dispersion telling you how much your data vary around the measure of central tendency. Estimating dispersion is very important because variability is an inherent part of collecting data. Some variability arises from real differences among individual values, while some arises from measurement error. The sample variance has a minimum of zero (if all your observations are the same) and increases with variation among observations. Calculate the sample variance using the following formula:

$$S^{2}= \frac{Σ(x\_{i}-\overbar{x})^{2}}{n-1}$$

**The standard deviation**

The standard deviation ($S$) is another measure of dispersion. If you can calculate the sample variance, you can easily calculate the standard deviation; the standard deviation is simply the square root of the variance. Like the sample variance, the standard deviation has a minimum of zero (if all your observations are the same) and increases with the variation among observations.

$$S=\sqrt{S^{2}}$$

**The standard error of the mean**

A quantity that is related to the sample variance and standard deviation is the standard error ($SE$) of the mean. One useful property of the standard error is that, all things being equal, it decreases as sample size increases. This property makes sense. An estimate of a mean will become more precise as you collect more data, simply because you have more information. This precision is reflected in your standard error, which will also become smaller with sample size. Put another way, your confidence in your estimate of the mean will increase as sample size increases. Because the standard error reflects the precision of, and therefore your confidence in, the estimate of the mean, the mean and the standard error should always be presented together. In contrast to the standard error, neither the variance nor the standard deviation decrease as sample size (n) increases. As such, the variance and standard deviation are estimates of how variable your data are, while the standard error is an estimate of how precise your estimate of the mean is. It can be calculated by dividing the standard deviation of your sample by the square root of the sample size:

$$SE=\frac{S}{\sqrt{n}}$$

**Statistical tests**

Sometimes you have more than one dataset or more than one variable. This may require you to make comparison or describe a specific relationship to answer your research question. In these cases, you will have to use statistical hypothesis testing.

Scientists often want to compare groups of observations to see if they differ. Groups can be defined based on a categorical variable or factor created in nature (e.g., males and females), or a variable manipulated by the scientist (e.g., the amount of fertilizer given to a plant). The goal of such comparisons is to determine whether the two groups differ in a continuous variable (for example, height), which in turn can be used as support for a claim of causation, e.g., that an increase in fertilizer causes increases in plant growth. If you find a difference in mean (average) values for the continuous variable between the members of each group you sampled, you might be tempted to conclude that your grouped observations or experimental treatments have revealed a meaningful effect on the variable you measured. Unfortunately, conclusions drawn from mean values can be misleading. Measurement error and uncontrolled environmental variation can cause two means to differ somewhat, but it would be wrong to interpret differences caused by these factors alone to represent differences caused by the factor that you are investigating. To determine whether a difference between two means is scientifically meaningful, we need to partition out the variation in our data that is caused by a given variable or experimental manipulation from that caused by measurement error and environmental variation. In other words, we need to analyze our data using statistical hypothesis tests.

A statistical hypothesis test allows you to discriminate between an alternative hypothesis, which is your estimation of the effect of a given variable or experimental manipulation on the data you have collected, and a null hypothesis, which is the idea that the variable or manipulation you're studying will have no effect on your data. Statistical hypothesis testing requires you to calculate new statistics, called test statistics. Generally, as the absolute value of a test statistic increases, so does your confidence that you can reject the null hypothesis. How large the test statistic must be for you to reject the null hypothesis declines with sample size. You will never be certain that the alternative hypothesis is true; all you can have is some defined level of confidence that the null hypothesis is false. By convention, when the absolute value of a test statistic is so large that there's less than a 5% chance that the null hypothesis is true, (a so-called alpha or P-value of 0.05 out of 1), scientists reject the null hypothesis and tentatively accept the alternative.

Here are several of the most common statistical tests:

1. The t-test. A t-test is used when you want to compare the means of two groups. A t-test addresses the alternative hypothesis that an observed or manipulated variable (the independent variable) that falls into two categories affects a second variable (the dependent or response variable) that is measured on a continuous scale.

2. The analysis of variance (often abbreviated as ANOVA). An ANOVA is used when you want to compare more than two groups (e.g., an independent variable that falls into more than two categories).

3. Regression analysis and correlation analysis. Regression and correlation analyses are related techniques that are used to look at the relationship between two variables. These analyses differ from the t-test and ANOVA because they are used when your data do not fall into discrete groups, but are instead continuously distributed. Correlation analysis allows you to examine the association between two variables without making any assumption about the causal relationship between them. In other words, you can use correlation analysis when you do not know if variable 1 is influencing variable 2, or vice versa. In contrast, regression analysis is used when you believe that one variable (the independent or predictor variable) influences the other (the dependent or response variable). One result of a regression analysis is a line or curve showing the best prediction of the true relationship between variables, given the data at hand; for this reason, it is sometimes referred to as "curve-fitting."

4. 2 tests. Chi-squared (pronounced "k-eye-square") analysis and related techniques are used when both the independent and dependent variables are categorical (i.e., fall into discrete groups), so what needs to be analyzed is the pattern of frequencies of different outcomes.

**Tutorial**

The free Google Sheets (<https://www.google.com/sheets/about/>) add-in XLMiner Analysis ToolPak (<https://chrome.google.com/webstore/detail/xlminer-analysis-toolpak/aolkkmkpomhckcckjbmnibeakmhdmflj>)

or the built-in functions within Excel work great for summary stats and simple stat tests. Make sure you specify the correct input and output range. Your TA will give a short tutorial showing you how to analyze and graph the sample data. Please follow along on your laptop.

**Making figures and tables**

Figures and tables frequently will help the reader to understand complicated data more easily than a written description. Note, however, that if the data can be easily summarized in the text (e.g., they consist of 2-3 numbers), a figure or table is not necessary. The text should tell the reader the important points to be noted on the graphs or tables, or call out specific examples from the figure or table to illustrate a point. Obviously the same data should not be presented in two different forms (e.g., a table should not contain raw data that is summarized in a graph), so decide which form helps you tell your reader what you want him/her to know. Graphs of any kind, as well as other pictorial materials, are referred to as “Figures” in the text. Tables are called “Tables” in the text, and are numbered separately from figures. Both are numbered in the order in which they are referred to in the text (so you shouldn't refer to Figure 3 before Figure 1 is mentioned). In a table, make sure you label all columns and give it a title and an appropriate description/footnotes. All figures should have a title and description. The figure should be a standalone product that does not need additional information outside the description to understand. Thus, it is necessary to include a written legend explaining each color or line type, a mention of the statistical test if necessary, and other descriptors such as sample size and location/date of samples. Remember that bar graphs are used when the independent variable is categorical whereas line graphs are used to show continuous relationships. Always include units on both axes. The independent variable (manipulative variable) is on the x-axis and the dependent or response variable is on the y-axis. Gridlines and boarders should be used carefully or not at all in order to not obscure the graph or table. Also, you may want to use symbols to denote significant results especially in tables.

Table 1: Guide to using and graphing statistical analyses based on goals.

|  |  |  |
| --- | --- | --- |
| **Goal** | **Statistical analysis** | **Suggested figure** |
| Showing the distribution of data |  | Frequency histogram |
| Presenting mean, range,… | Summary stats | Bar graph or box plot with standard error (SE) bars |
| Comparing the mean of 2 groups | t-test | Bar graph or box plot with standard error (SE) bars |
| Comparing the mean of more than 2 groups | ANOVA | Bar graph or box plot with standard error (SE) bars |
| Association of 2 variables | Correlation | XY scatter plot with fitted line |
| Causal relationship of 2 variables | Regression | XY scatter plot with fitted line |
| Comparison of frequency patterns | Chi-squared | Table or histogram |

Some of the content within this lab handout was taken verbatim from the Grinnell College Investigation Course website (<https://www.grinnell.edu/file/investigations-spring-2015-final.pdf>) and is used for educational purposes only.