Data Analysis Guide

Kenneth Fortino

September 16, 2009

Introduction When we gather information about anything we are collecting data. Data are simply all of the pieces of information that we have collected about our subject. If we follow certain guidelines during data collection we can use statistical tools to test questions about relationships within our dataset. A comprehensive discussion of these guidelines and the reasons for them are beyond the scope of this course. However in general to use statistical tools to analyze data we typically need measurements of the following things[1]:

1. A measure of the **location** of a group of measurements – usually this is the mean ($\bar{x}$).
2. A measure of the variation around the mean **within** each group.
3. A measure of the variation **between** the means of each group.
4. A measure of how the variation in one group of measurements relates to (or **correlates** with) the variation in another group of measurements.

Measures of Location Measures of location tell us where most of the values of the data lie. There are 3 common measures of location: mode, median, and mean. The **mode** is the most frequent value in a set of measurements. The **median** is the value in a set of measurements that has equal numbers of greater and lesser than observations. The **mean**($\bar{x}$) is the sum of all the values (Σx) divided by the number of values (n). Mathematically the mean is defined: $\bar{x} = \frac{\Sigma x}{n}$.

Within Variation The mean is an estimate of the location that a group of numbers would converge on if we took an infinite number of observations. Since we rarely have an infinite sample, we need to have some measure of how much each of our individual measurements differs from the mean. Three common measures of **within** variation are:

1. **Variance** (s^2) which is defined as the sum of the squared difference between each value and the mean value divided by the 1 minus the number of values. Mathematically: $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$.

[1] Definitions and concepts adapted from: Gotelli, N. J. and A. M. Ellison, 2004. *A Primer of Ecological Statistics.* Sinauer Associates, Inc. Sunderland, MA

2. **Standard Deviation** (s) which is defined as the square root of the variance.

3. **Standard Error of the Mean** ($s_{\bar{x}}$ or SE) which is defined as the standard deviation divided by the square root of the number of observations. Mathematically: $SE = \frac{s}{\sqrt{n}}$.

4. **95% Confidence Interval** (CI_{95}) which is defined as SE multiplied by 1.96. The CI_{95} says that the actual measure of location for the group you are sampling (i.e., the *true mean*) will be contained within the range described by ± the CI_{95} 95% of the time you calculate the mean and CI_{95}.

The specific differences and justifications for each of these measures of within variation is way beyond the scope of this lab but note that the larger any of the above measures are, the more variation within the group of measures and the less effectively the mean represents the location of the group.

Between Variation Often our scientific questions will lead us to a comparison of measurements from two different groups (e.g., the recovery speed of groups on two different drugs). If we calculate and compare the mean value for each group, the difference is called the **between** variation. However, remember that there is **within** variation around the mean of the groups so lots of the values from one group may overlap the values of the other group even if the means are different. If we have lots of overlap in the two groups then they may not really be different. We can use comparisons of the **within** and **between** variation to provide evidence for whether the two means are really different.

One of the easiest ways of doing this is to determine whether the CI_{95} of the two groups overlap. Recall that the CI_{95} of a group says that the *true mean* will be inside the CI_{95}, 95% of the time it is calculated so if there is no overlap between the CI_{95} of the two groups then 95% of the time that we make the comparison we would conclude that the two means are different. It is a statistical convention to refer to differences that occur 95% of the time as "significant" (i.e., real) differences. Note that we cannot say the with absolute certainty that the two means are different, we can only say that it is unlikely that they are not different.

Correlation Another comparison we may wish to make is to see if paired measurements from individual samples (e.g., the paired height and weight of a group of people) have any relationships in their variation. For example if one of the paired measurements (e.g., weight) increases when the other measurement (e.g., height) increases then they are positively correlated. A negative correlation indicates that one measurement decreases when the other increases. Correlations can vary from a perfect negative correlation (indicated by -1) to a perfect positive correlation (indicated by 1). If there is no relationship between the two measures then the correlation is 0. If the correlations are perfect (i.e., -1 or 1) then all of the variation in one measurement is captured by the variation in the other measurement.