Statistics and Data Analysis in MATLAB
Kendrick Kay, kendrick.kay@wustl.edu
February 3, 2014

Lecture 2: Hypothesis testing and correlation

1. Exploring a more complex dataset: one variable, two conditions

- Suppose we measure a quantity not just for one condition (which was the subject of Lecture 1), but for two conditions. For example, suppose we measure the heights of male adults and the heights of female adults. What can we do with the data?
- One thing we might want to do is to determine whether the two conditions differ with respect to the mean. In Lecture 1 we saw that the mean of a sample drawn from a population is subject to sampling error. Here, we face an analogous problem: the difference in the means of two samples is also subject to sampling error. Thus, even if we find that the means of the two samples are different, the samples could actually have been drawn from the same underlying probability distribution.
- To determine whether two conditions differ with respect to the mean, we use a statistical approach known as *hypothesis testing*. In hypothesis testing, we pose a *null hypothesis* and ask: if the null hypothesis is true, how likely is the observed pattern of results? This likelihood is known as the *p*-value, and indicates the statistical significance of the observed pattern of results. If the *p*-value is less than some threshold that we decide upon (e.g. $p < 0.05$), we reject the null hypothesis.
- The *t*-test is the classic method for testing whether two conditions have different means (ANOVA is the generalization of the *t*-test to multiple conditions). In essence, the *t*-test poses the null hypothesis that the means of the two conditions are equal, makes some assumptions on the distribution of the data (specifically, that the two conditions are Gaussian-distributed with equal standard deviation), and then calculates a *p*-value analytically.
- An advantage of parametric tests like the *t*-test is that they tend to be powerful, i.e., if a difference exists, parametric tests are likely to detect that this is the case. However, parametric tests rely on assumptions, and these assumptions may be invalid for a given dataset. Moreover, we may be interested in comparing something other than the mean. To address these scenarios, we can use nonparametric techniques for hypothesis testing. This is especially viable given that nowadays we have plenty of computing power.

2. Nonparametric alternatives to the *t*-test

- *Randomization (or permutation) tests*. Let's pose the null hypothesis that the two sets of data come from the same probability distribution (not necessarily Gaussian). Under the null hypothesis, the two sets of data are interchangeable, so if we aggregate the data points and randomly divide the data points into two sets, then the results should be comparable to the results obtained with the original data. So, the strategy is to generate random datasets, compute some statistic from these datasets (such as difference in means or difference in medians), and then compare the resulting values to the statistic computed from the original data. We count the number of randomly obtained values that are more extreme than the actual observed value and divide this by the total number of simulations that were run. The result is the *p*-value. Notice that we have used raw computational power to calculate the *p*-value directly instead of relying on analytical formulas (which are valid only if certain assumptions are met).

- *Bootstrap methods*. In the previous example, we posed the null hypothesis that the two sets of data come from the same distribution and then used randomization to generate new datasets. A slightly different method is to use bootstrapping to generate new datasets instead of randomization. The difference is that in the case of randomization, we enforce the constraint that none of the data points are repeated—akin to drawing data points without replacement—whereas in the case of bootstrapping, we generate new datasets by drawing data points with replacement. The advantage of randomization is that it may be easier to understand. However, bootstrapping has a more solid probabilistic foundation, and samples drawn from the bootstrap distribution may better approximate the underlying data distribution compared to samples generated through randomization. Practically speaking, the two methods likely give similar results.
- *Other applications of hypothesis testing*. We have introduced hypothesis testing in the context of testing differences in means (or medians) of two groups, but hypothesis testing can be applied in other circumstances. One case of note is testing whether the mean (or median) of a single group is different from zero. To address this case, let's pose the null hypothesis that the observed data come from a probability distribution that has zero mean. A reasonable choice for this null distribution is the data itself but recentered at zero. Under the null hypothesis, if we draw samples from the null distribution and compute the mean (or median) of these samples, then the results should be comparable to the actual mean (or median) of the original data.

3. Exploring a more complex dataset: two variables, one condition
- Suppose we have one condition and measure not just one quantity (which was the subject of Lecture 1) but measure two distinct quantities. For example, suppose we measure both the heights and weights of male adults. What can we do with the data?
- The *scatter plot* is an extremely useful visualization that illustrates the relationship between two variables. Before getting into technical details, you can learn a lot by simply looking at your data.
- One possibility is that there is no relationship between the two variables, that is, the variables are *independent*. Technically, this can be expressed as follows: the *joint probability distribution* of the two variables is equal to the product of the *marginal probability distributions* for the two variables. Intuitively, independence means that knowing the value of one variable provides no information about the value of the other variable.
- The other possibility is that there is some relationship between the two variables. A simple case is that there is a *linear relationship* between the variables. Such a relationship can be positive (increase in one variable is accompanied by increase in the other variable) or negative (increase in one variable is accompanied by decrease in the other variable). All other relationships can be lumped under the category of *nonlinear relationships*. An example of a nonlinear relationship is a U-shaped curve.

4. Correlation
- Linear relationships can be quantified using the metric of *correlation (r)*. Correlation values lie in the range –1 to 1, where –1 indicates a perfect negative linear relationship, 0 indicates no relationship, and 1 indicates a perfect positive linear relationship. (Note that we are discussing Pearson's product-moment correlation, but there are other variants of correlation.)
- Correlation can be calculated as follows:

$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\text{std}(x)} \right) \left(\frac{y_i - \bar{y}}{\text{std}(y)} \right)}{n}$$

In words, we z-score each variable (subtract off the mean, divide by the standard deviation) and then compute the average product of the variables. (Technical note: in the above formula, std should be computed using a version of standard deviation where we normalize by n instead of $n - 1$.)

- Correlation can be given a nice geometric interpretation: Treat each set of data as a vector in n-dimensional space. Subtract off the mean of each vector and normalize each vector to be unit-length. Correlation is now simply the dot product of the two vectors.
- It is common to compute r^2, which indicates the fraction (or percentage) of variance in one variable that can be explained by the other variable. The range of r^2 is 0% to 100%. (We have not yet formally addressed the concept of variance; we will do that in a later lecture.)
- Note that correlation captures only linear dependencies. This means that the concepts of correlation and dependence are related but not identical. If two variables are correlated, they are necessarily dependent. However, two variables can be dependent but still have zero correlation.
- To capture all types of dependencies between two variables (not just linear ones), we can use the metric of *mutual information*, but that is outside the scope of this class.

5. Error bars and p-values on correlation

- How do we obtain error bars on the correlation observed in a set of data? There are parametric methods that make assumptions about the distribution of the data, but let's address this nonparametrically. Once again, the bootstrap comes to the rescue: treat the observed data as an empirical probability distribution, draw bootstrap samples from this distribution, calculate correlation values for these bootstrap samples, and construct a confidence interval from the obtained correlation values.
- How do we test whether the correlation observed in a set of data is significantly different from zero? Let's address this nonparametrically using a randomization test. The null hypothesis is that there is no relationship between the two variables (in other words, the variables are independent). If the null hypothesis is true, then we should be able to shuffle the order of each variable and obtain datasets that are equivalent to the original dataset. So, to obtain a p-value, shuffle each variable, calculate a correlation value for the shuffled data, repeat this a large number of times, and count the number of times that randomly obtained correlation values are more extreme than the actual observed correlation value.
- While we are on the topic of p-values on correlation values, it is convenient to introduce here *Monte Carlo methods*. Monte Carlo methods are a very general class of methods that make use of randomly generated data to test various hypotheses. To illustrate, let's look at a simple example. Imagine that someone tells you that they found a correlation of $r = 0.4$ for a sample size of 10. Assuming we do not have access to any other information, what can we do to evaluate the claim that an actual positive correlation exists? Let's run some simple Monte Carlo simulations in which we draw samples of size 10 from two independent Gaussian distributions and compute the correlation value observed in each sample. Such simulations reveal that it is actually quite likely to obtain a correlation of $r = 0.4$ for a sample size of 10, even when the underlying distributions are independent. Thus, we should not believe the claim.