**Reflection after Class 6**

As I read through Chapter 3 of *Practical Statistics for Data Scientists*, I learned a great deal about statistical experiments, specifically about how to create them and use them effectively in a given scenario. Sections 1 to 3 laid out, in great detail, numerous methods of testing to use to verify or invalidate certain hypotheses.

**What is A/B Testing?**

An A/B test is an experiment with two groups to determine the superior choice of two factors, whether they be treatments, products, or procedures. These factors (to which the groups are exposed) are called **treatments**. Often, one of the two treatments during A/B testing is either the currently existing standard treatment or no treatment at all. If this standard (or nonexisting) treatment is used, it is called the **control**. In these A/B tests, the hypothesis is usually that the experimental treatment is better than the control.

**What do you need in order to conduct an A/B test?**

In A/B tests, the group exposed to the experimental treatment(s) is called the **treatment group**, while the group exposed to the control (either a standard or nonexistent treatment) is called the **control group**. Each group contains subjects, which are the items (such as people or objects) that are exposed to the treatments. Ideally, subjects are **randomized** (randomly placed in either group), to offset bias, so that the results of the experiment are either caused by the effectiveness of the treatments or just plain random luck. To measure the effectiveness of the treatments in comparison to each other, a **test statistic**, or a metric common to both experiments, is used.

**Why might you need more than just A/B?**

While A/B may be useful when you want to note the difference between two treatments, data scientists aren’t often looking at just two treatments. Instead, they are often thinking something along the lines of “Out of all of these treatments, which one is the most effective/is the most useful/has the greatest impact?” Thus, while A/B testing can be quite useful in fields such as marketing or web design, where specific elements are often compared, it is not as useful in other fields where many different elements are compared.

**What are hypothesis tests? (Feel free to use outside sources to answer this). Why might they be useful in helping us understand observed effects?**

Hypothesis tests, also known as significance tests, are tests whose purpose is to find out whether random chance or luck has a role to play in the observed effect that occurs in the result. Using this testing method, one must test their hypothesis (the **alternative hypothesis**, or what you hope to prove/why chance is not to blame) by comparing it with the **null hypothesis** (the counterpart to the alternative hypothesis, or the hypothesis that chance is to blame for the observations recorded in the results of an experiment). In hypothesis tests, the null hypothesis is only rejected if its probability falls below a predetermined significance level, in which case the hypothesis being tested is said to have that level of significance. Hypothesis tests are useful in helping us understand observed effects as they provide evidence concerning the plausibility of the hypothesis based on data from the observed effects seen in the results.

Hypothesis tests can be **one-way tests** (hypothesis tests that only count chance results in one direction) or **two-way tests** (hypothesis tests that count chance results in two directions), and this is dependent on whether or not you care about chance favoring both hypotheses or only one. For example, if you are running a hypothesis test on the observed data from an A/B test and are using a control treatment as one of the two treatments, you likely would care less if chance favored the control, as the goal is to definitively prove that the experimental treatment is better than the control. On the other hand, if you ran an A/B test where both treatments were experimental treatments and your goal was to find the “better” treatment, you would run a two-way hypothesis test to confirm that chance did not favor either side.

**What are t-tests used for?**

A very common type of significance test that is frequently used is the t-test. All significance tests require a **test statistic**, or a metric for the difference or effect of interest, and the t-test is no different. This is used to measure the effect you are interested in and help you determine whether that observed effect lies within the range of normal variation in luck/chance.

Before the common use of computers, resampling tests were not practical (reshuffling large amounts of numeric data to resample was not optimal), and statisticians used standard reference distributions. A test statistic could then be standardized and compared to the reference distribution (**t-distribution**). One such widely used standardized statistic is the **t-statistic**, or a standardized version of common test statistics, such as means. Nowadays, t-tests are used to simply compare the means of two groups.

**Why is resampling useful in data science?**

**Resampling** is the method of drawing additional examples from the observed data set either **with replacement** (by returning each removed item to the sample before the next draw) or **without replacement** (by not returning each removed item to the sample before the next draw) for the purpose of assessing the random variability of a statistic. There are two procedures involving resampling: bootstrapping and permutation testing.

Bootstrapping (the procedure of drawing random elements from the population with replacement, repeating this step until there are as many elements drawn as there are in the population, recording a statistical value of the dataset created from this repeated drawing, and repeating the previous steps an arbitrarily large number of times) is used to determine the reliability of an estimate.

Permutation testing (the procedure of combining the samples of the different groups, shuffling the combined data, resampling from the new dataset without replacement to create a new sample of the same size as the first group, repeating the last step for every group, recalculating the statistic measured in each original sample for each resample, and repeating the previous steps an arbitrarily large number of times to yield a permutation distribution of the test statistic) is used to test hypotheses. Without resampling, neither bootstrapping nor permutation testing would be possible.