From the Expert: Data Analysis

Hypothesis testing is considered as a statistical inference where data from the sample is used to draw conclusion about the population parameters. The process begins with making the assumption about the population parameters called null hypothesis (H_0). Then, the alternative hypothesis (H_a), which is the opposition of the null hypothesis, is formed. The hypothesis testing attempts to determine whether the null hypothesis should be rejected.

However, the error from using sample data for the hypothesis test may be possible. In general, the error can happen in two different ways:
1) H_0 is rejected when H_0 is actually true (Type I error or false positive)
2) H_a is accepted when H_a is actually false (Type II error or false negative)

To control the probability for type I error to happen, the level of significance (α) is specified. In other words, the level of significance can be described as the maximum probability that is allowed for type I error. The most common level of significance are α=0.05 and α = 0.01. For example α=0.05, it means that there is only 5% of making mistake for Type I error. Note that, most experiments would stop here. To make the experiments more complete, type II error could be further investigated. Type II error is represented as β. The power of the test is denoted with 1-β, which is the chance that the null hypothesis is rejected when it is not true. In order to avoid the false negative, a certain number of sample size is required. The formula and automatic calculation of the sample size can be found in many sources. In general, the larger sample size, the greater the power to detect the difference.

To make a conclusion (accept or reject the null hypothesis) for the test, the p-value and α is compared. p-value is the probability of the observed results from the experiments or surveys given that the null hypothesis is true. If the p-value is less than α, the H_0 will be rejected. Otherwise, we cannot reject the null hypothesis. The term “statistical significance” is often used when p-value is less than α. The meaning of significance is simply means the “difference” in the null hypothesis.

The hypothesis test can be employed on population parameters such as mean, variance, standard deviation, and proportion. In addition, the hypothesis test can be performed on regression, correlation coefficients, and probability distribution.

Experimental design or design of experiments (DOE) was proposed by Ronald A. Fisher. DOE is involved with the analysis of data that is generated from an experiment. The focus of the experimental design is on the effect of the process or intervention (called treatment) on the objects (called experimental units), such as: patients, plants, animals, etc. The study is widely used in medicine, biology, agriculture, industrial production and marketing research.

These terms are often mentioned in DOE:

- **Treatments** are the different methods that we administer to experimental units so that we can compare effects such as different design layout webpages, different membership fees, or different temperatures.
- **Experimental units** are things on which we apply the treatments such as customers who see different webpage designs, or plants growing in different temperatures.
- **Responses** are outcomes that we measures from experimental units after applying the treatments such as customer likes/dislikes of the design, change in profit, or quality of products.
- **Factors** are the general type or category of treatments. For example, 2 groups of students with 3 different learning methods. The learning methods are treatments and three "types of learning" methods are three levels of factor.
- **Control** is the treatment that is the baseline for comparing with other treatments. The control may be with no treatment (null treatment) or common use treatment such as standard web page without variation or common pain killing medicine.
- **Placebo** is often used in human subjects. It is a null treatment that simulates the treatment to deceive the subjects. For example, patients with headache are given inactive treatment such as sugar pill as a placebo for pain killing medicine.
- **Confounding** is the term to describe the effect of one factor or treatment cannot be separated from another factor or treatment.

Different types of variables

- Independent variable is something that you want to manipulate
- Dependent variable is something that you want to measure
- Control variable is something that you hold constant
- Random variable is something that you allow to vary randomly
- Confounding variable correlates with independent variable

There are different types of experimental design, which can be classified by treatments, experimental units used, how treatments are assigned, and how response are measured:

Completely randomized design is a common practice where objects are randomly assigned to an experimental group. The treatment groups are ensured to be as similar as possible.

Randomized complete block design is concerned on the differences among objects within the experimental group. Therefore, the objects are divided into homogeneous group called "block" before they are randomly assigned to a treatment group. For example, patients may be divided into different age group levels (block). Then, patients would be randomly assigned to the treatment within the particular age level.

Factorial design emphasizes two or more factors at the same time. The effects of several factors and their interactions can be identified.

A/B testing is a simple randomized design that has been applied in a new context. This technique has been widely embraced by major web companies such as Google, Facebook, and Amazon, etc. for web page optimizations. Here, the statistical hypothesis testing is employed on two variants A and B. The process begins with creating another version of the original webpage. The page visitors are randomly assigned between the original page (control) and the variation version (treatment). Finally, the data regarding to the webpage performance is collected. The webpage that is not well performed will be cancelled.

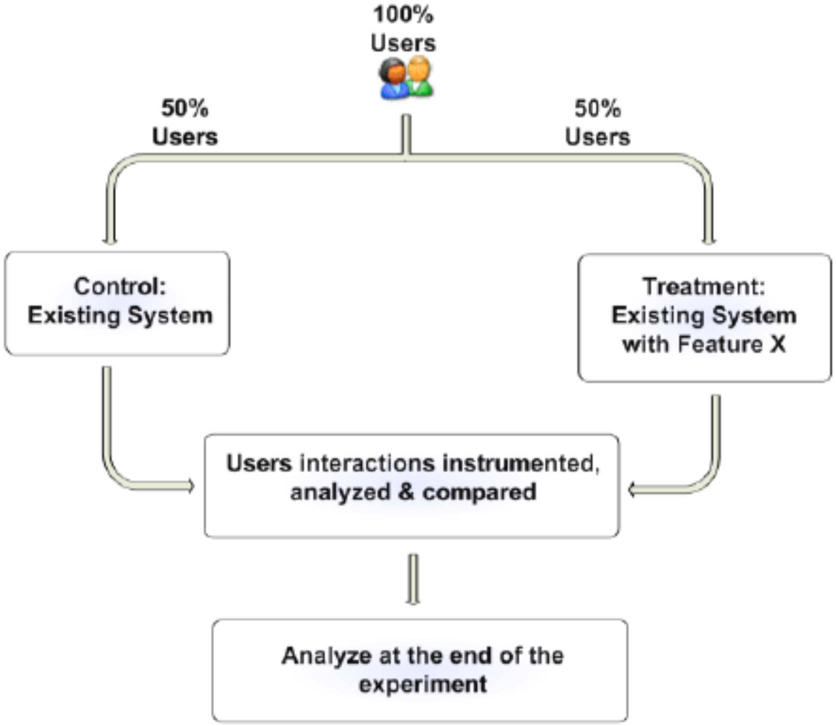
100% Users
50% Users
50% Users
Control: Existing System
Treatment: Existing System with Feature X
Users interactions instrumented, analyzed & compared
Analyze at the end of the experiment

Figure shows A/B testing from Kohavi, et al. (2007) who worked for Microsoft. The detail about A/B testing can also be found in the paper referenced.

Here are some useful suggestions for A/B testing:

- Always do A/A testing first. They are supposed to have similar performance. However, you may detect some variation that you cannot ignore before you run A/B testing
- When doing A/B testing, always test them at the same time (i.e. randomly assign control page and the variation page to the customers). Different days or different weeks of testing different versions might impact the results
- Do not stop the experiments too early. This might affect the statistical confidence if there are not enough samples.
- Avoid making a surprise to regular visitors. It is recommend including only new visitors in the test.
- Do not let your feeling rule out the test results. The results may not be what you expected.

Regression analysis is related to identifying the relationship between one dependent variable and one or more independent variables. The variable that is not changed by other variables is called the independent variable (or predictor variable). Whereas, the variable that could be changed or depends on other variables is called the dependent variable (also called response or outcome variable). For example, time spent on studying can cause changes in the test score. Therefore, time spent on studying is the independent variable and the test score is the dependent variable. Regression analysis is widely used for prediction. The equation is derived from the line that best fit the data points with minimal deviations from the line. Given the sample (x, y) pairs, the parameters $(b_0, b_1, \ldots, b_k)$ are estimated and utilized for creating the regression equation. Once the regression equation is obtained, you can predict the value of the response variable (i.e. score) from the predictor (i.e. study time).

Simple linear regression assumes that there is a linear relationship between one dependent variable and one independent variable. The response variable should be continuous. Usually, the relationship is expressed in the form of a mathematical equation that fits the relation the best. The terminology is only suggested that X predicts Y but **not** X causes Y.

$$y = b_0 + b_1 x$$

where b_0 is the regression coefficient or intercept
b_i is the regression coefficient for x_1 or slope
y is dependent variable
x is independent variable

Logistic regression is similar to simple linear regression except the dependent variable must be a categorical variable with 2 possible outcomes (dichotomous) such as (click, not click) or (yes, no) etc. With basic transformation, logistic regression can be changed into linear regression.

$$L = logit(y) = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

where p is the proportion of success
b_0 is the regression coefficient or intercept
b_1 is the regression coefficient for x_1 or slope

Multivariate (multiple) linear regression is an extension of simple linear regression. The model is the relationship between one dependent variable (continuous variable) and two or more independent variables (categorical or continuous). This model describes the impact of multiple predictors on a single response variable.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

where b_0 is the regression coefficient or intercept
b_i are the regression coefficients for independent variables x_1 through x_k

Normality is the underlying assumptions for many statistical testing. Two main methods for normality assessments are graphical methods and statistical tests.

For graphical methods, there are several ways to check the normality in the data:

1) Construct the histogram. Then, look at the data that is distributed. The histogram should resemble the normal distribution (bell shape curve).

2) Use quantile-quantile plot (QQ Plot). The actual scores, which are ranked and sorted, would plot against its corresponding expected values. If the data is normally distributed, the data points will lie along the diagonal line from lower left to the upper right.

3) Residuals scatter plot. Residuals are the difference between obtained and predicted values. The scatter plot between the predicted value and the residual that is normally distributed will show the majority at the center of plot with some distributed on both sides evenly.

References

Abrams, D. R. (n.d.). *Introduction to Regression.* Princeton University. Retrieved from http://dss.princeton.edu/online_help/analysis/regression_intro.htm.

Chopra, P. (June 24, 2010). *The ultimate guide to A/B Testing*. Smashing Magazine. Retrieved from http://www.smashingmagazine.com/2010/06/24/the-ultimate-guide-to-a-b-testing/.

Hansen, M. (n.d.). *Statistics 13 (A/B testing)*: UCLA lecture presentations, pg. 59. Retrieved from http://www.stat.ucla.edu/~cocteau/stat13/lectures/lecture6.pdf.

Kohavi, R., Henne, R. M., & Sommerfield, D. (2007). *Practical guide to controlled experiments on the web: Listen to your customers not to the HiPPO*. KDD 2007. Retrieved from: http://www.exp-platform.com/documents/guidecontrolledexperiments.pdf.

Oehlert, W. G. (2010). *A first course in design and analysis of experiments*. University of Minnesota: Retrieved from: http://users.stat.umn.edu/~gary/book/fcdae.pdf.

Siroker, D. (Mar 20, 2013). *Optimizely CEO - Best Practices from 100,000 A/B tests*:. [Video file] Retrieved from http://www.youtube.com/watch?v=lDfoVxHud7Y&feature=youtu.be.

Significant. A webcomic of romance, sarcasm, math, and language. Retrieved from http://xkcd.com/882/.