STATISTICS: REGRESSION ANALYSIS, HYPOTHESIS TESTING

1. Introduction to Hypothesis Testing

In many situations, often only a small data sample can be obtained and a decision needs to be made about a population on the basis of this sample. This is the realm of statistics and hypothesis testing. For instance, a pharmaceutical company has to decide about the efficacy of a drug based on a clinical trial on a small sample of the population. Interventions, Treatments, and Therapies need to be validated for their efficacy based on a small sample. Randomized trials are frequently resorted to in such situations and statistical techniques are used to estimate the likelihood of finding a specific sample if it was drawn at random from the population under a given hypothesis. For instance, given a Polio treatment – how likely is that treatment preventing the occurrence of Polio? Is that vaccine effective? These are the types of questions that Hypothesis testing can help us uncover.

In order to perform hypothesis testing, we'll make extensive use of probability theory – concepts such as Central Limit Theorem are at the heart of these techniques.

The gold standard for hypothesis is a randomized, double blind trial. In such a trial, all other conditions are kept constant and only the main intervention (e.g. a drug or a procedure) is varied between the two groups. Neither the investigators nor the subjects know which population they have been assigned to and in such a setting stronger claims about causality (i.e. the drug improved the disease condition or the intervention caused a change in the behavior) can be made.

2. Null Hypothesis Significance Testing

A well-regarded, and frequently used approach to determine if interventions and claims that are being made are valid is Null Hypothesis Significance Testing. In this approach, we formulate a Null Hypothesis (you can think of this as maintaining the *status quo*) and then test it against an alternate or Research hypothesis. For example, let's say that a tire company claims that it has engineered a better tire – a tire where the treads wear out slowly. How can you verify if this claim is true? The Null Hypothesis could be that the tire treads wear out at the same rate on these new tires compared with the old technology. An alternate hypothesis in this case can be that the new technology tires have longer tread wear out time. In order to test the hypothesis, we conduct a trail where we acquire samples of tires under both these technologies (old and new) and then outfit them on cars. We then drive these cars under different conditions and then examine the amount of tread wear. Once we know the average rate of tread wear in both these sub-samples, we can then compute the likelihood of the estimated average wear of the new set of tires and either reject the Null Hypothesis that the tread wear is the same or fail to reject it. For instance, if the normal tire lasts for 60,000 miles and the new sample indicates that the mean tire life is 65,000 miles – we might be failing to reject the Null Hypothesis. On the other hand if the new sample indicates a mean tire life of 80,000 miles – we might have sufficient confidence to reject the Null Hypothesis.

When we perform Null Hypothesis Significance Testing, we are frequently comparing means of the two subsets – the one subject to the invention and the one that has not been exposed to the intervention. The Null Hypothesis is that the means of the two subsets are the same and the Alternate Hypothesis would be that the intervention did indeed affect (improve) the mean.

Two things govern whether the Null Hypothesis gets rejected or not – one is the effect size and the second one is the sample size. The larger the size of the effect, the more confidence we have in rejecting the hypothesis. In addition, the larger the size of the sample, the more confident we are on the accuracy of our estimates. The second is because of the weak law of large numbers. Since we have a larger sample, there is higher likelihood that the sample mean is close to the Expected value of the population.

Ideally, we want a large enough sample and a large enough effect to conclusively reject the Null Hypothesis and thereby provide evidence that the alternate hypothesis is plausible.

We note that Null Hypothesis Significance Testing is a Frequentist approach. We assume either the hypothesis is True or False and we calculate $P(D|H)$ when the hypothesis is true or false. That is, we calculate the likelihood of seeing a particular data sample D under the condition that the hypothesis is true. We reject the null hypothesis if this probability is lower than a significance threshold (typically 0.01) and we accept the null hypothesis otherwise.

3. Types of Errors

There are two types of errors that can occur when we perform null hypothesis testing. First, we can incorrectly reject a true null hypothesis. This is also referred to as a False positive. Examples include a Fire alarm going off when there is no fire or a medical treatment indicating that the intervention works when in fact it does not work.

The second type of error we can make is failing to correctly reject a false null hypothesis. This is a False negative or a Miss. Examples include a fire being present and the alarm fails to go off or the treatment does really work but the clinical trial fails to show that conclusively.

When we compare population means, a Type I error would be concluding that the two subset means are different when they are in fact not different. A Type II error would be concluding that the means are not different when they are in fact different.

The Table 1 below summarizes the relationship between the truth and falsehood of the null hypothesis and the outcome of the statistical test:

4. Regression Analysis

Together with Hypothesis testing, Regression Analysis forms a standard tool in the statisticians tool chest. Simple Linear Regression can be used to study the relationship between two variables: A dependent variable and an independent variable. When we do

	H_0 True	H_0 False
Reject H_0	Type I Error (False Positive)	Correct (True Positive)
Accept H_0	Correct (True Negative)	Type II Error (False Negative)

Table 1. Summary of the types of Errors

hypothesis testing, the dependent variable is usually the outcome that is of interest. For instance, in the case of a drug intervention, the outcome is the disease state or extension of lifespan. The independent variable could be the dosage of the drug, for instance. Independent variable is something that the experiment can manipulate and the dependent variable is the outcome that is being tested. Given the simple case of two variables: one dependent (y) and one independent (x), we can model the relationship between them using a Linear function (hence the name Linear Regression) thus: $y = b_0 + b_1x + e$. Here, b_1 represents the size of the effect and e represents the measurement errors and other irreducible/unmeasurable effects. Given a sample set with many (y_i, x_i) pairs, we can use Ordinary Least Squares to estimate the best b_0 and b_1 that reduce the error to the minimum. After fitting, we can evaluate the values b_0 and b_1 and see if they are significant. For example, we can evaluate the null hypothesis that b_1 is 0 and accept it or reject it based on the evidence. If b_1 is significant, then we can potentially conclude that the intervention is effective. For example, we may want to study if exercise reduces total cholesterol. The dependent variable would be total cholesterol in the body as determined by a fasting blood test and the independent variable would be the amount of exercise done per week by people. For instance, in a randomized double blind study (the gold standard of such tests), the subjects will be randomly assigned to one of the many groups (no exercise, 5 minutes a day, 10 minutes a day, 20 minutes a day, for example). In addition, the researchers will not know which participants have been assigned to which group. At the end of the measurement period, there will be many pairs of (y_i, x_i) where x_i will be average amount of exercise per week and y_i will be total cholesterol after the conclusion of the evaluation period. In such a sample, we can perform Linear Regression to fit b_0 and b_1 and determine if b_1 is significant. That is, the Null Hypothesis is that b_1 is zero or that exercise has no effect on cholesterol. If $b_1 = 0$ is rejected, we can conclude that it is not true that exercise has no effect on cholesterol. When there are multiple variables involved, we can still perform Linear Regression and measure the significance of each of the variables. For instance, the mpg of the car is a function of the engine horsepower, the air drag coefficient of the car, the weight of the car etc. We can perform a linear regression analysis to fit a model and measure the significance of each of the variables on the dependent variable (in this case, the mpg).

Once you perform a fit, you can also compute a confidence interval (typically a 95% confidence interval) that states that if you perform 100 such trials, 95 of them will have fits in a given range. For each of the coefficients in the linear fit, you can compute a 95% confidence interval. Clearly it is desirable to have the confidence interval as close as possible to the estimated coefficients in a given experiment. The size of the confidence interval is determined by the number of samples you have. The larger the number of samples, the tighter the confidence interval. This comes from the weak law of large numbers. The more samples you have, the closer you are to the expected value of the underlying quantity.