Review of Introduction to Statistics

Different introduction to modeling

Conceptually there are two approaches to represent reality: one that assume perfect knowledge and one that assume that humans will always have some missing information. The debate seems trivial, but the arguments are compelling from both perspectives. However, from practical perspective the lack of information approach is preferred. The main tool used to implement this is statistics, which assume that we have access to all the information, but not to all the facets that describe the process of interest.

For example we can measure with high accuracy and precision the height and age of a tree but we cannot predict with the same level of accuracy and precision the height from age. So, we observe the object, meaning we have access to all relevant information, but we miss some of the related information, which generically can be described as covariate, or a variable that is related with the main variable of interest. To ensure the complete description of a process, which is one of the logical requirements of an argument, the missing part of the information is represented stochastically. Therefore, we associate a particular chance of an event to occur, in our example the chance to observe (measure) a particular height at a given age. Do not worry for the time being of the condition "given age", the essence of the argument is the same. The theoretical limitations are not the only one that hinder the complete representation of reality. There is also the practical aspect, as we will rarely, if not ever, measure all the individuals of a population. We are operating thru representatives, or samples. The inference to the population from the sample was and still is one of the central topics of understanding the surrounding reality, or basically modeling.

Central Limit Theorem

Now, that we know the philosophical foundation of the class, let review the main result of the stochastic, sampling based approach, to reality. I hope that you remember it: it is the Central Limit Theorem.

The formal statement of the CLT is:

Let assume a set of random variables $X_i, i = \overline{1,n}$, that are independent and identically distributed. The distribution of each X_i is the same and has a finite mean, μ, and variance, σ^2. Then, for large n, the mean of the random variables is normally distributed:

$$\bar{X}_n = \frac{X_1 + X_2 + .. + X_n}{n} \sim Normal$$

This is the main results. However, the theorem is even more powerful because CLT is also providing us the mean and variance of the distribution of $\bar{X}_n$, which is μ, and σ^2/n, . Therefore, CLT is formally written as $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

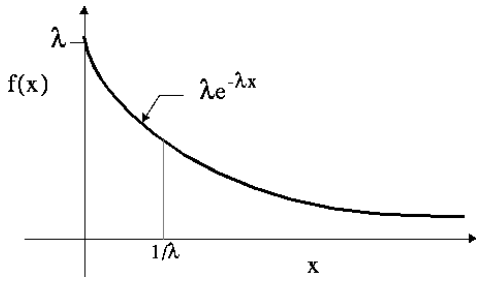
Side note: One of the requirements of the CLT is that the variables have the same distribution; however, there is a form of CLT that has this requirement relaxed. This form was developed by Lyapunov and in essence has the same for as the classical CLT, if some additional conditions are fulfilled.

Important **observations**:

1. Irrespective the distribution of the random variable *Xi*, their mean of will be normally distributed. Now, let see how this major results, for many people the most important results of statistics, help us in understanding reality. Remember, we are unable to measure all the individuals, so we measure few individuals, n, which make a sample. Assuming that the individual measured are

from the same population of interest (let say Spruce) and that the variable measured is the same (let say dbh), then the distribution of variable is the same. This means that the dbh of spruce comes from the same distribution. Then, the mean dbh is normally distributed. This is the main results, because it relates theory to reality.

2. The distribution of the sample mean is fully known. This is the most important tool to represent numerically the world around us. Why, because it allows us to make valid inferences from limited information. A normal distribution is fully described by two parameters: mean and variance. The CLT provides us with the main tool on estimation of the two parameters: 1) mean of the sample is the same with the mean of the distribution of the original variables, and 2) the variance of the sample mean is the variance of the original variables divided by the sample size.
3. As the sample sizes increases, the variance of the sample mean decreases. Therefore, for large samples, the variance of the sample mean converges to zero. This trial results, is one of the "Achilles' heel", as it allows us to prove anything empirically.



Let see how CLT operates with an example with Excel from an exponential distribution. I have chosen the exponential distribution because is a distribution very different from normal, as it is "open at one end", meaning has the largest skewness possible (i.e., asymmetry). We will do this by creating a set of random samples with 5 and 50 values (i.e., n=5 and n=50). To ensure representatively we will choose 20 samples of each size. The exponential distribution is defined by ne parameter, λ, and has the probability density function $pdf(x) = \lambda e^{-\lambda x}$. The mean of a RV exponentially distributed is λ^{-1} and the variance is λ^{-2}. For our example,

let choose λ= -2, which renders a mean of 0.5 and a variance of 0.25. To generate a random variable with an exponential distribution in Excel we will use the function: -1/λ×ln(1-y), where y is a random number, uniformly distributed from 0 to 1. The syntax is: =-1/2*LN(1-RAND()).

One small issue with CLT: to compute the mean and variance of the sample mean we need the mean and variance of the original distributions. In real word, not only that we cannot measure all the individuals, but also we do not know the mean and variance. Because, the linear property of the mean and CLT, we can replace the population mean with the sample mean. But we cannot do the same for variance. So what we can do? Lucky for us Gosset, while working for the Guinness Brewing Company, developed a distribution that bear his name, the t-distribution, which can be obtained as the ratio between the mean and standard deviation of a set of *n* random variables normally distributed:

$$t = \frac{\bar{X}}{s_X/\sqrt{n}}$$

Looking at the formula, you should realize that there is a connection between CLT and t, but it is not straight forward. So how t-distribution is related to CLT? The simplest way of explain this relationship is by starting with the main theorem:

As IN CLT, let assume a set of random variables $X_i, i = \overline{1,n}$, that are independent and **normal** distributed with mean μ and variance σ^2. Then the ratio of the difference between the variable mean and sample mean and the standard error of the sample has a standard normal distribution, *N(0,1)*.

Analytically, the theorem is written as:

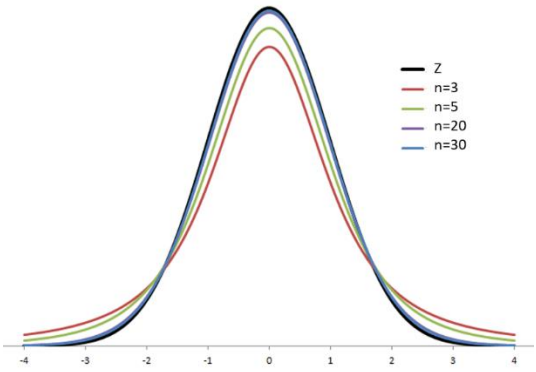
For $X_i \sim N(\mu, \sigma^2), i = \overline{1,n}$, the statistics

$$\bar{X}_n = \frac{X_1 + X_2 + .. + X_n}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

determine a random variable $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ which follow a t-distribution with n-1 degrees of freedom.

The main finding of the theorem is not that the sample mean is the same with the population mean, which we already know from CLT, but the fact that the standard deviation of a sample can be used as an estimate of the standard deviation of the population. Remember, CLT states that for a set of ANY iid random variables $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$, whereas the Gosset results states that $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)$.



Are those results important? The mathematician George Polya, who coined the term central limit theorem in 1920, stated that CLT plays a pivotal role in probability theory.

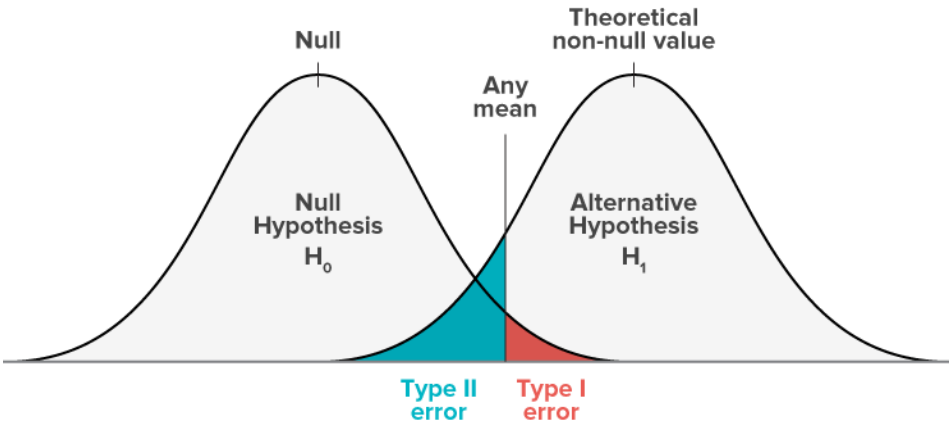
Application of CLT

Now that we have a more clear understanding of the main results of statistics, let see how we can use it. But first let revisit the foundation of science, again. Remember, we want to infer population parameters from sample statistics. Because of this "expansion" from ample to populations errors would be present. Therefore, it is important to know what errors are important. There are two types of possible errors, called false positive and false negative. A false positive error is an error that

supports a statement when in fact the statement is not true. In statistics false positive errors are called type I errors. A false negative error, is an error that does not support a statement when in fact the statement is true. In statistics false negative errors are called type II errors. It is not simple to see that false positive and false negative are not mutually exclusive, in the sense that if one happens the other doesn't. In fact there are many results that shows a nonlinear relationship between these two errors. Now that we know what errors we can encounter, which one is more important false positive or false negative? This is an ethical question, and the society decided few thousands of years ago, that false positive errors are not preferred. Therefore, this should be the focus of the test. Why? An example will clarify the reason why: let assume that you are judged for a crime that can end with capital punishment. The society decided that it is not preferred to be found guilty when innocent, false positive, even that this means that a true criminal will go unpunished (false negative).

To assess the presence or absence of an error we are using various tests. Any test is a mathematical solution to a statement. For simplicity, the statement is called hypothesis, and because it is easier to work with equalities than with inequalities, a hypothesis is stated as an equation. The test evaluate the truthfulness of the hypothesis. Because we are in the fundamental assumption that information is missing, the assessment itself is prone to lack of information, therefore we will use a stochastic approach. This means that instead of stating with certitude that the hypothesis is true or false, the test will tell with a certain degree of certitude if the hypothesis is true, and consequently with a certain degree of certitude that the alternative, or contrary, hypothesis is true. Because the hypothesis to be tested is an equality, it is customary referred to as The Null Hypothesis.

One could ask rightly, what is the connection between CLT and hypothesis testing? The answer is rather simple. Remember, that from the CLT the mean of the sample is normal or t-distributed. A hypothesis simply states whether or not the mean of the sample has a particular value, given or computed. The test of Type I error is a measure of how far or close the sample mean is to the theoretical mean when the null hypothesis is true, whereas the test of Type II error is a measure of how close the sample mean is to the theoretical mean when the alternative hypothesis is true (as we see in the figure). The two shaded areas have a particular names: α for Type I, the red area, and β for Type II, the green area. α is the probabilities of Type I error, which is compared with the p-value, and β is the probability of Type II error. You should notice that small α leads to larger β.



All this discussion serves to test, for example, if a particular sample from a population is different than a given value. Let assume for example that you would like to thin a particular stand. You know from you silviculture class that thinning can occur if the mean dbh is larger than 10 cm. Therefore, you went in the field and measured 10 plots. The following values were obtained: 10 8 11 8 9 7 11 9 6 13. The question is: should you thin or not? Let answer this question using hypothesis testing.

First, let convert the question of interest, thin or not thin, in a statement that can be tested. The null hypothesis in this case is H0: the dbh ≥ 10 cm. Always when you setup the null hypothesis you have to state the alternative hypothesis, which in this case is Ha: The dbh<10. Now let ask yourself, why dbh <10 and not dbh ≠10. The answer is practical, remember the intent is to thin, and if the dbh <10 then you will not thin. Now that we know what we have to test, let appeal to our friend CLT. According to CLT, the mean of the sample is normally distributed with mean the mean of the population. Therefore, $\bar{X} = \frac{92}{10} = 9.2$ cm.

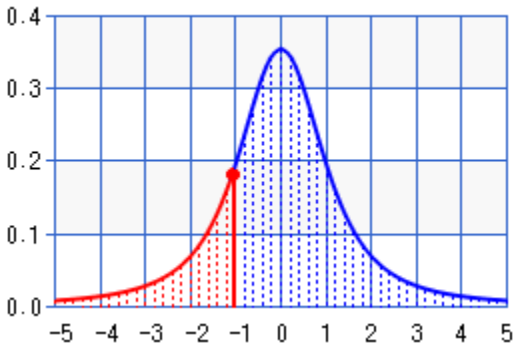
Next step is to see if you have all the information required by CLT: do we know the variance of the population? The answer is no. Therefore, we have to use Gosset's t-distribution. First let compute the variance of the sample:

$s^2=1/9\times[(10-9.2)^2+(8-9.2)^2+(11-9.2)^2+..+(13-9.2)^2]=4.4$

The statistics that follow a t-distribution is by convention labeled t empiric or t computed or t data, and is computed as

$$t_{empiric} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{9.2 - 10}{\sqrt{4.4/10}} = -1.2$$

We have estimated all the values in the formula except the μ, which according to H0 is 10. The Type I error is now focused on the left side of the t curve, as the Ha is focused on the values lower than 10. Using Excel, we can compute the probability that is left on the t empiric:



t.dist(-1.2, 10-1, TRUE)=0.13. Depending on the preset acceptance it can be stated that for H0 is not rejected for an α=0.05 or rejected for and α=0.2.

Expectation

In probability and statistics a central role is played by the Expectation. Simply stated, the expected value of a random variable is the arithmetic mean of that variable. Expectation is written as E(.), where . stands for the parameters or statistic of interest. For example the mean of X is simply E(X), which is μ. Depending on the type of variables there are two possibilities: discrete and continuous.

The expected value of a discrete random variable, X, is found by multiplying each X-value by its probability and then summing over all values of the random variable. $E(X) = \sum_{All\ values\ of\ X} p(x) \times x = \mu$

For a continuous variable X ranging over all the real numbers, the expectation is defined by

$$E(X) = \int_{-\infty}^{\infty} f(x) \times x dx = \mu$$

The variance of a random variable X is defined as the expected squared deviation of the random variable values about its mean. Therefore,

$$var(X) = E[(X - \mu)^2] = E(X^2) - \mu^2 = \sigma^2$$

For a discrete RV

$$var(X) = \sum_{all\ values\ of\ X} p(x)(x - \mu)^2$$

The covariance of two random variables is

Cov(X,Y)=E[(X-E(X))(Y-E(Y)]=E(XY)-E(X)E(Y)

Properties of expectation

- *E(aX)=aE(X)*
- *E(X+Y)=E(X)+E(Y)*
- *Var(a+bX)=b^{2}var(X)*
- *Var(X±Y)=var(X)+var(Y)±cov(X,Y)*