Inferencial statistics

A/B testing and confidence interval

Nawal Yala

Contents

Chapter

I

Introduction

In the field of statistics, there are two main branches: descriptive statistics and inferential statistics. When analysing data, we will often use both together, so it will be useful to first describe each branch and how they differ from one another.

1 Population & Sample

Before any analysis, should we determine what the data you are dealing with. It represents a population or a sample. A population is a collection of all elements of interest. It can be all product in a factory, all individuals in a certain country/city, all patients of a certain illness...etc. The number of all elements in a population is denoted with an uppercase **N**. A sample is a subset of the population and is denoted with a lowercase **n**. Populations are hard to define and hard to observe in real-life (all patients of a certain illness). A sample, however, is much easier to contact. It is less time consuming and less costly. Time and resources are the main reasons we prefer drawing samples compared to analyzing an entire population.

2 Descriptive statistics

Descriptive statistics is about describing/summarizing our data (data can represent a population or a sample) using the following measures: measures of center, measures of spread-/variability, outliers and shape of our distribution (use plots of our data to gain a better understanding). Measures of center can be the *mean*, *median* or/and the *mode*. Whereas the measures of spread can be the variance, standard deviation, and the range. The mean is the common measure of center used, while the standard deviation (the root of variance) is common one used to measure spread. Standard deviation can be difficult to interpret as a single number on its own. Basically, a small standard deviation means that the values in a data are close to the mean of the data set, on average, and a large standard deviation means that the values in the data set are farther away from the mean, on average.

> **Remember.** The standard deviation measures how concentrated the data are around the mean; the more concentrated, the smaller the standard deviation.

3 Inferential Statistics

Inferential Statistics is about using our collected data (sample data) to draw conclusions to a larger population. Performing inferential statistics well requires that we take a sample that accurately represents our population of interest. A common way to collect data is via a survey. However, surveys may be extremely biased depending on the types of questions that are asked, and the way the questions are asked. This is a topic we should think about when collecting our sample data.

4 Measurement formulas in Samples and Population

In the field of statistics, we call:

Notation.
Parameter: is a numerical value that can be computed from a population of data. It can be any measure can be calculated by descriptive statistics.
Statistic: is a numerical value that can be computed from a sample of data. It can be any measure can be calculated by descriptive statistics.

We also use different symbols to denote a parameter and a statistics:

Table 1. Denotation of parameter and statistic.

Measure	parameter	statistic
mean	μ	$\bar{x}$
standard deviation	σ	s
proportion	π	p

In statistics, we generally use different formulas when working out the population data and sample data. When we work out the population, each data point is known. So we are 100% sure of the measures we are calculating. There is no need to any conclusion about population from a sample. However, when we extract a sample from a population, any calculated statistic is interpreted as an approximation of the population parameter. Moreover if you extract 10 different samples from the same population you will get 10 different measures. Statisticians have solved the problem by adjusting the algebraic formulas for many statistics to reflect the issue.

The sample mean is the average of the sample data points. While the population mean is the average of the population data points. Since We use N and n, there are two different formulas but they are computed in the same way as below:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.1}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.2}$$

But what about the variance? is it calculated in the same way? Standard deviation measures the dispersion of a set of data points around their mean value. Population variance denoted by σ^2 is equal to the sum of square differences between the observed values and the population mean divided by the total number of observations. While sample svariance is equal to the sum of square differences between the observed values and the sample mean divided by the total number of observations minus one. The two formulas are below:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{4.3}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4.4}$$

The number of values in a sample is called *degree of freedom.* Why the dgree of freedom is decreased by 1 in the statistic (second) formula? Let's know about this with a practical example.

> **Example** (Population).
> We have a population of 7 observations. So N = 7.

> Population = [1,1,1,2,3,4,5,5,5]
> The mean μ is 3.
> The population variance σ^2 is: 2,888889.
> The standarad deviation of the population σ is: 1,699673.

We have the population thus we don't need any estimation of the population parameter. Imagine now, we have only a sample of this population. We don't know the data in the example obove and we need to estimate population parameter from the sample we have.

> **Example** (Sample)**.**
> We have a sample of 5 observations. So n = 5.
> Population = [1,2,3,4,5]
> The mean $\bar{x}$ is 3.
> The sample variance s^2 is: 2.500000.
> The standarad deviation of the sample s is: 1.581139.
>
> > **Remark.** The variance above is calculated using **n-1** in the denominator. Let's now calculate it using **n** in the dominator instead.
>
> The sample variance (n in denominator) s^2 is: 2.
> The standarad deviation of the sample ((n in denominator) s is: 1.414214 .

When we decrease the degree of freedom by one (i.e using n-1 in the variance formula), we obtain a value close to that obtained by the population. A sample is used here to estimate population parameter, so the closet values to the population parameter are the most accurate ones. To conclude, the degree of freedom is decreased to correct the variance upwards.

> **Remember.** The number of freedom in the sample variance formula is decreased by one to correct the sample variance upwards. s^2 is the unbiased estimate of population variance calculated from a sample.

The mean and the standard deviation of any random variable are calculated with same formulas given above. For instance, X is the random variable that represents the proportion of individuals having a certain characteristic in a data. Let's suppose there are m individuals having a characteristic among N individuals (so there are m 1s and N-m 0s). Proportion of individuals who support this chacarcteristic π is equal to $\frac{m}{N}$. The proportion is thus the mean of individuals that have a characteristic. This is true in a population as well in a sample, the difference is just in the notation (in a sample, $\mathbf{p} = \frac{m}{n}$).

The standard deviation of the proportion is calculated as the standard deviation of any variable. For the population:

$$\sigma_\pi^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \pi)^2 \tag{4.5}$$

x_i are either 1 or 0. and $\pi = \frac{m}{N}$ So,

$$\sigma_{proportion}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \pi)^2 = \frac{m(1-\frac{m}{N})^2 + (N-m)(0-\frac{m}{N})^2}{N} \tag{4.6}$$

After some simplification, we get:

$$\sigma_{proportion}^2 = \frac{1}{N}m(1-\frac{m}{N}) \tag{4.7}$$

And finally:

$$\sigma^2_{proportion} = \pi(1 - \pi) \tag{4.8}$$

For a sample, the variance is obtained by decreasing the degree of freedom by one.

Chapter

II

inferencial statistics

5 Introduction

In the previous part, we discussed descriptive statistics, and we said that we can use it to summarize our data. In fact, is practically impossible to know all items in a given population. So descriptive statistics is no longer way to describe and summarize the population data. Often we have only a sample of a population and we have to use it to estimate population parameters. Statisticians, in this case, use the sample data and calculate its *statistics* (using descriptive statistics), and from that, draw an inference that the parameter of the entire population falls within a specified interval of values. Inferencial statistics refers to methods and procedures that used to calculate such interval. There are two main approches of inferential statistics. The first, is the estimation of the parameters (such as mean, median, and standard deviation) of a population based on those calculated for a sample of that population. The estimation of parameters can be done by constructing a range of values in which the true population parameter is likely to fall. The second approch of inferential statistics is hypothesis testing. It is to determine the effectiveness of an experimental treatment. This is done by determining if the treatment yields results that are significantly different from those obtained from a sample given no treatment at all.

6 Distribution

In this section, we learn about distribution because it is the core of the inferential statistics approches. If you feel confortable with this topic you can pass to the next section. In statistics, which says distribution says distribution of probability. A distribution is a function that shows the possible values for arandom variable X and how often they occur. Statisticians use the following notation to describe probabilities: $\mathrm{P}(\mathrm{X} = \mathrm{x})$ = the probability (likelihood) that random variable takes a specific value of x. The sum of all probabilities for all possible values must equal 1.

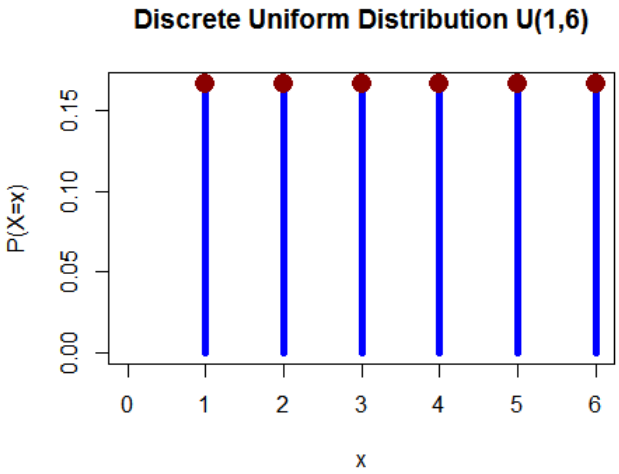
Probability distributions describe the dispersion of the values of a random variable. The type of probability distribution is determined by the type of variable. In statistics, distributions of a random variable are divide into the following two types:

1. Discrete probability distributions for discrete variables

2. Probability density functions for continuous variables

6.1 Discrete probability distribution

Discrete distribution can assume a discrete number of values. Die toss and the number of books sild each year are examples of this first type of distribution. These are discrete distributions because there are no in-between values. We can have only 1, 2, 3, 4, 5 and 6 as results in a die toss. Similarly, we can count 400 or 402 books sold, but nothing in between. For discrete probability distribution, each possible value has a non-zero likelihood. Furthermore, the probabilities for all possible values must sum to one.



The probability of getting X between 2 and 4 is equal to the sum of probablity of getting each value: P(2<= X <= 4) = P(X=2)+ P(X=3)+ P(X=4) = 3(1/6) = 0.5

> **Remark.** The probability values create the shape and the type of a distribution ant not the opposite.

> **There are other discrete distributions, the following are examples:.**
> - Binomial distribution to calculate the probability of getting binary values (outcomes), such as coin tosses.
>
> - Poisson distribution to model count data, such as the count of books sold per year.
> - Uniform distribution to claculate the probability of multiple events with the same probability, such as rolling a die.

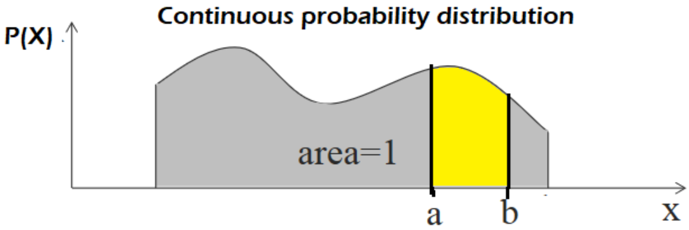
6.2 Continuous probability distribution

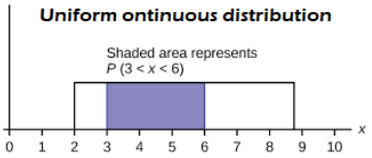
Where the random variable can take any real value (an infinite possible number), we have know that we have a continuous or density distribution. Age, Time, Height, weight, and temperature are examples of a continuous distribution. Unlike discrete distributions where each particular value has a non-zero probability, specific values in continuous distributions have a zero probability because a single value doesn't exist. For instance, the probability of measuring a height that is exactly 166.67cm is zero because no matter how correct measuring devices are, accuracy still are not enough with respect to the nature of a real number. More clearly, We cannot reach mathematically to measure the exact value of the hieght; 166.67cm can be 166.67777776777777! So between two measures extremely close there are thousand of real values. Thus, P(X = 166.67) = 0. it is meangful to talk about P(X<= 166.67).

So, probabilities for continuous variable are measured over interval of values rather than single value. A probability indicates the likelihood that a value will fall within an interval.

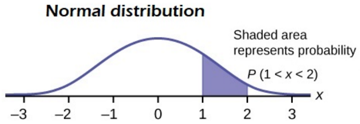
In discrete distribution, the probabilities for all possible values must sum to one. On a continuous probability plot, the entire area under the distribution curve equals 1. The proportion of the area under the curve that falls within an interval of values along the X-axis represents the probability that a value will fall within that intarval. Finally, you can't have an area under the curve with only a single value, which explains why the probability equals zero for an individual value.

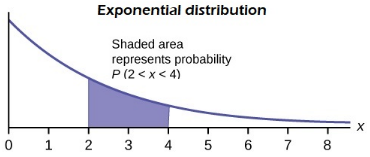
In the plot below of continuous distribution, the probability that X fall within an interval [a, b] is equal to curve air under that interval (yellow area).





As in discrete distribuion, there are many continuous probability distributions, including:.
- Uniform - Normal - Students
- Exponential - log_normal - Weibull ... etc





Each probability distribution has parameters that define its shape. Most distributions have between 1-3 parameters. Specifying these parameters establishes the shape of the distribution and all of its probabilities entirely. These parameters represent essential properties of the distribution, such as the central tendency and the variability.

Important.
Normal distribution and Students distribution are the most distributions in statistics. This is due to the following reasons:

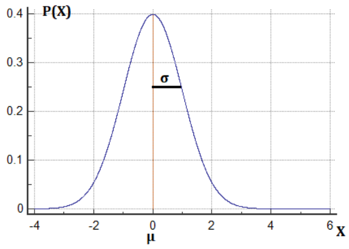
(a) With enough sample sizes (degree of freedom), distribution of some sample statistics are approximated to normal distribution.

(b) All computable statistics are elegant.

(c) Decisions based on normal distribution insites have a good track record.

We will see all this in detail in next and later sections.

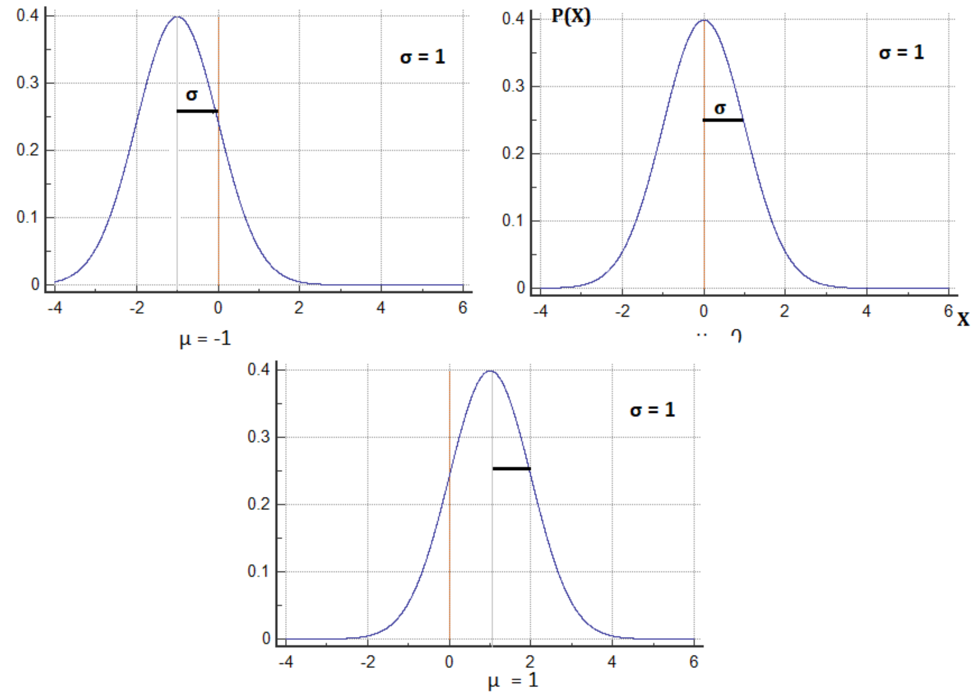
6.3 Normal distribution



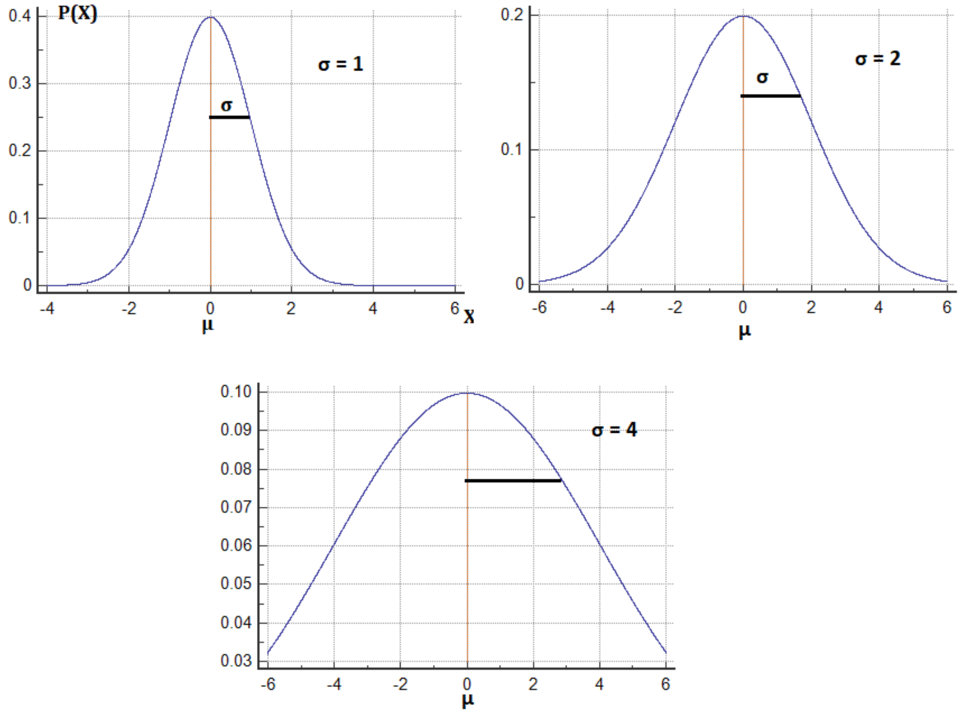
As it is shown below, normal distribution has a shape like a bell, that's why it is called bell curve. It is symmetrical and its mean median and mode are equal.It is perfectly centered around its mean It is denoted in this way: $N \sim (\mu, \sigma^2)$. N stands for normal. The tilde assigned ($\sim$) denotes it is a distribution and in brackets we have the mean and the variance of the distribution. You can notice that the highest point is located at the mean because it coincides with the mode. The spread of the graph is determined by the standard deviation. P(X) is the function of probability of X. It is a gaussian function:

$$P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right\} \tag{6.1}$$

With same standard deviation and different means, we obtain a graph that moves from left to right.



With same mean and different standard deviations, we obtain a graph that moves from narrowest to widest.



As any continuous distribution, probability of X fails within an interval is calculated by the curve air under that interval. For instance, However, computation probablities in normal distribution are direct and elegant. Let's summarize all normal distribution properties :

Normal Distributions Properties

1. The normal curve is symmetrical about the mean μ;
2. The mean is at the middle and divides the area into halves;
3. The total area under the curve is equal to 1; it is completely determined by its mean and standard deviation σ:
   - Approximately 68% of the area under the normal curve is between $(\mu-\sigma\ ,\ \mu+\sigma)$. In other word, P$(\mu-\sigma <X< \mu+\sigma) = 0.68$.
     That also means that 68% of X values fits within one standard deviation of the mean.
   - Approximately 95% of the area under the normal curve is between $(\mu-1.96*\sigma\ ,\ \mu+1.96*\sigma)$. In other words, P$(\mu-\mathbf{1.96}*\sigma <X< \mu+\mathbf{1.96}*\sigma) = 0.95$. That also means that 95% of X values fits within one standard deviation of the mean.
   - Approximately 99% of the area under the normal curve is between $(\mu-\mathbf{2.58}*\sigma\ ,\ \mu+\mathbf{2.58}*\sigma$. In other words, P$(\mu-2.58*\sigma <X< \mu+2.58*\sigma) = 0.99$. That also means that 99% of X values fits within one standard deviation of the mean.

One can ask why these intervals are calculated from of $\mathbf{k}*\sigma$ of the mean? For any distribution, it is useful to know how dispersible the X values are. We would probably be interested in knowing how our data is far from the mean. Standard deviation can be greatly affected if the mean gives a poor measure of center. We will see the importance of that in later sections. Another question can be raised: what if the probability distribution **isn't a Normal**, how the probability of X falls within an interval of $[\mu-\mathbf{k}*\sigma\ ,\ \mu+\mathbf{k}*\sigma]$ is calculated? To answer this question, Chebychev has a theorem extracted by experience. the theorem is known as ***Chebyshev's inequality*** and says:

Theorem 6.1 (Chebyshev's inequality)**.**
If X a random variable follow any probability distribution: X $\sim D(\mu,\sigma)$ with μ finite and σ non-zero, then $1-\frac{1}{k^2}$ of the distribution's values (X values) will lie within k standard deviations of the mean.
We transform the text into a formula, P$(\mu-\mathbf{k}*\sigma <$X$< \mu+\mathbf{k}*\sigma) >$ 1 - $\frac{1}{\mathbf{k}^2}$

The sign (>) reveals that the theorem provides a worst-case look at data dispersion within any data distribution including the Normal distribution.

Let's claculate probabilities for different values of k:
k=2: 1 - (1 / 22) = 1 - 0.25 = 0.75 (75%)
k=3: 1 - (1 / 32) = 1 - 0.11 = 0.89 (89%)
In these cases, Chebyshev's inequality states that at least 75% of the data will fall within 2 standard deviation of the mean, and 89% of the data is expected to fall within 3 standard deviations of the mean. With the concept of (at least), the theorem is true even for Normal distribution, however with less precision, since 95% and 99% of X vlaues in Normal distribution fall within 1.96 and 2.58 standard deviation, respectively.

6.4 Standard Normal distribution

The standard normal distribution is a particular case of the normal distribution. If a random variable X follows a normal distribution with parameters ($\mu = 0, \sigma^2 = 1$), i.e. $X \sim N(0, 1)$ then we say X is a standard normal random variable. We use Z to denote a standard normal distribution instead of N. So we say $X \sim Z(0, 1)$. The density of standard normal therefore is,

$$P(X = x) = \frac{1}{\sqrt{2\pi}} \exp\left\{\left(-\frac{x^2}{2}\right)\right\} \tag{6.2}$$

The standard normal distribution is easy to understand and has several important applications. Every normal distribution can be standardized. To transform a normally distributed variable X to one with a standard normal distribution, we create a variable called z-score. It is equal to the originale variable X minus its mean divided by its standard deviation:

$$z - score = \frac{X - \mu}{\sigma} \tag{6.3}$$

If we take a data and subtract its mean from each data point and then calculate the mean again, we'll gat zero. Once again you will get zero. Let's take an example:

Example (Standardizing a normal distribution).
data1 = [1, 2, 2, 3, 3, 3, 4, 4, 5]
The mean μ is 3.
The standarad deviation of the population σ is: 1.2.
Now let's substract the mean from all data points.
We get a new data =[-2, -1, -1, 0, 0, 0, 1, 1, 2].
Let's calculate the mean of the new data. It is 0.

Remark. The propability or the number of occurence of each point in new data remains the same as in the old data.

So far we have a new distribution which is still normal but with a mean of 0. The standard deviation is still 1.2. The next step of standarization is to set the the standard deviation to 1. To do this, let's first remeber that the standard deviation is equal to the average of the distance of each data point to the mean. If we want to set its value to 1, we just divide all points by its old value.

data_standarized = [-1.63, -0.82, -0.82, 0, 0, 0, 0, 0.82, 0.82, 1.63].
If we calculate the standard deviation of this data, we will get 1 and the mean is still 0.

Using standard normal distribution makes inference much easier. We will see that in the next sections.

7 Sampling distribution

Before given any definition, let's say we have a population under study (the population of interest), which can be objects, people, measurement..etc. Since it is impossible to collecte all items for a population, we collect a sample from it. Statiticians when collect data from a population say that they sampling it. Let's say that we have a chance to collect more than one sample from this population.

Sampling the population.
Let's say that we collected $\boldsymbol{L}$ samples of size $\boldsymbol{n}$ each.
sample_1 $=[x_1^1, x_2^1, ..., x_n^1]$
sample_2 $= [x_1^2, x_2^2, ..., x_n^2]$
.
.
.
sample_L $= [x_1^L, x_2^L, ..., x_n^L]$

Let's now calculate a **satatistic** from each sample. This could be any statistic that can be calculated by *descriptive statistics*. Let's calculate, for example, the **mean** of each sample.

Calculate statistic from each sample.
Statistic in this example is the mean. It can be any statistic.
$\bar{x_1}$ = mean(sample_1)
$\bar{x_2}$ = mean(sample_2)
.
.
.
$\bar{x_L}$ = mean(sample_L)

Now, if we plot all the values of $\bar{x_i}$, i=1,...,L, x-axis carries the values $\bar{x_i}$ and y-axis carries the probabilities (frequenties of occurence), and we'll get what is known as **sampling distribution** of the sample mean. If the statistic is the proprtion, we'll get the sampling distribution of the sample proportion and so forth on any statistic. Our statistic $\bar{x_i}$ changes because the sample collected changes each time. The sampling distribution allows to know hos the statistic changes grom one sample to another.

Remember. The sampling distribution is the distribution of a statistic, this could be any statistic.

7.1 Central Limit Theorem and Law of Large Number

There are two very important mathematical theorems that are related to sampling distributions: The Law of Large Numbers and The Central Limit Theorem.

Theorem 7.1 (Central Limit Theorem (CLT)).
The Central Limit Theorem states that no matter the distribution of the entire population (entire data), it can be binomial, uniform or another distribution. If we repeatedly take random samples of size **n** from the population, then when **n** is enough large, the distribution of the means of samples (the sampling ditribution of sample mean) will follow a normal distribution. Moreover their distribution will have the same mean as the population mean μ and a standard deviation $\frac{\sigma}{\sqrt{n}}$. So the **sample mean** $\sim N(\mu, \frac{\sigma}{\sqrt{n}}$.
When we work out thhe sampling distribution, it would be more accurate to talk about the **standard error** 7.2 rather than the standard deviation.

Remark. It turns out that the Central Limit Theorem is true for more than just the sample mean. It actually applies for these well known statistics:

1. Sample means ($\bar{x}$)

2. Sample proportions (p)
3. Difference in sample means ($\bar{x_1} - \bar{x_2}$).
4. Difference in sample proportions ($p_1 - p_2$).

And it applies for additional statistics, but it doesn't apply for all statistics! For example doen't apply for variance/standard deviation.

7.2 Standard Error

So far we saw the **standard deviation** of any random variable. It is calculated as the square root of the variance:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{7.1}$$

Standard deviation is a measure of how far each data point is from the mean. It is the best measure of spread of an approximately normal distribution.

What is so the standard error?
It is the standard deviation of the sampling distribution formed by the sample means. The variable here is the mean not the data points themself. Standard error is a measure of how far each sample mean from the mean of the sampling distribution.

Since the mean of the sampling distribution (the mean of the sample means) is the true population mean (by the central limit theorem), the standard error tells us how the mean of any sample is far from the true population mean μ. The standard error can be calculated from the srandard deviation of the population using the following formula:

$$StandardError(SE) = \frac{\sigma}{\sqrt{n}} \tag{7.2}$$

n: sample size

> **Important.**
> The standard error decreases as the sample size increases. As the sample size gets closer to true size of the population, the standard error becomes close to zero and the mean of any sample becomes the mean of the population. As the sample size decreases, the sample means are more spread out, and it becomes more likely that any given sample mean is an inaccurate representation of the true population mean.

To assimilate what that said on sampling distribution, a practical excercise is available under the name: 'sampling distribution.ipynb'.

8 Confidence Interval

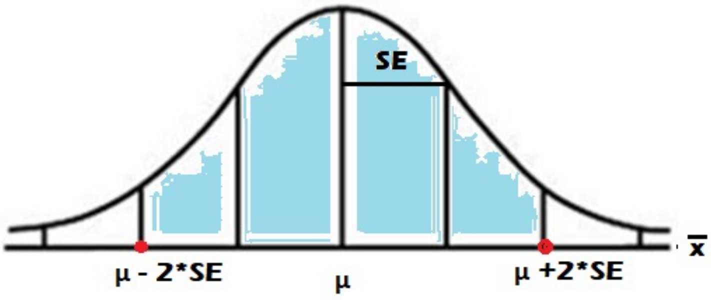
As mentionned in the introduction, one of the two approches of inferecial statistics is to estimate the population parameter from a sample statistic. A fast way to estimate a parameter is to approximate it to the sample statistic. For instance, we say: the population mean μ is approximately equal to sample mean $\bar{x}$. As sample size increases, $\bar{x}$ becomes very close to the true mean μ. Such estimate is called a ***point estimate***. The limitation of this point is that doesn't allow knowing how it is may be far from the population parameter.

However, an interval of values is more reliable and provides more information than a single value. **Confidence Interval** is a method used by statisticians to calculte a range of values to estimate a population parameter, with some level of confidence.

There are two methods to compute the confidence interval of a parameter, tradtional and modern methods. The first method is based on assumptions and formulas, while the second is based on bootstraping. We'll see the two methods.

8.1 Confidence Interval: traditional method

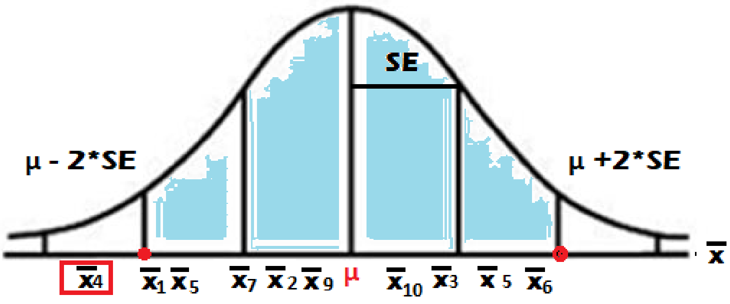
Before given any formula here, let's first suppose we want to estimate the population mean from the sample that we have. We suppose that our sample size **n** is large enough, so the Central Limit Theorem hollds and we can say that: the sampling distribution of the sample mean follows a normal distribution of mean equal to the population mean and strandard error equal to the standard deviation of the population divided by square of **n**. So, sample mean $\sim N(\mu, \frac{\sigma}{\sqrt{(n)}})$. The sampling distribution thus has all the properties that has any normal distribution 6.3.



Sampling distribution of sample mean

We emphasize that the points from which the sampling distribution is constructed are the means of samples not the data points themself. The air shadded on the plot represents the propbability of the sample mean is within two standard error from the mean of the means. This propbability is equal to 0.95. So, P(μ -1.96*SE < $\bar{x}$ <μ+1.96*SE) = 0.95. In other words, 95% of means (x-axis) are located within an interval of [μ -1.96*SE , μ+1.96*SE]. It is essential to understand this concept.

Let's suppose now that the sampling distribution in the plot is constructed by 10 sample means, i.e we took 10 samples and we calculated thier means and then we ploted the sampling distribution of sample mean.



As we can see 95% of $\bar{x}_i$ (9 means) are within an interval of [μ -1.96*SE , μ+1.96*SE] and 5% of $\bar{x}_i$ (1 mean) are ouside the interval. This also means that the probability of any sample mean will be within 1.96*SE from the distribution mean is 0.95 (P(μ -1.96*SE < $\bar{x}_i$ <μ+1.96*SE) = 0.95). Remember, we have often the possibility to collect one sample. Let's $\bar{x}_1$ is the mean of this sample, so if $\bar{x}_1$ is within [μ -1.96*SE , μ+1.96*SE], μ is within [M -1.96*SE , M+1.96*SE]. Morever, each time we collect a sample from a population, μ is within [$\bar{x}_i$ -1.96*SE , $\bar{x}_i$+1.96*SE] 95% of the time.

Finally, the formula of the confidence interval of the mean: We have a sample from a population with a size large enough. The mean of the sample is $\bar{x}$, so, the confidence interval of the population mean μ is:

$$CI_{95\%} = [\bar{x} - 1.96 * \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 * \frac{\sigma}{\sqrt{n}}] \tag{8.1}$$

We also can compute an interval with 99% confidence. By coming back to the Normal distribution properties, this interval will be large than the one with 95% confidence and it is computed as follows:

$$CI_{99\%} = [\bar{x} - 2.58 * \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58 * \frac{\sigma}{\sqrt{n}}] \tag{8.2}$$

For a number Z of standard deviation, there is a 1-α% of confidence. α determines the level of confidence and it is generally fixed to 0.01 or 0.05. The general formula of confidence interval is thus:

$$CI_{1-\alpha\%} = [\bar{x} - Z * \frac{\sigma}{\sqrt{n}}, \bar{x} + Z * \frac{\sigma}{\sqrt{n}}] \tag{8.3}$$

You are definitely noticing that the formula requires the standard deviation of the population σ! Generally we know nothing about the population. So if the propulation variance is known we can apply directly the formula. If the variance is unknown, we will estimate it.

8.2 Compute confidence Interval: known σ

If the population standard deviation σ is known, we should use a large sample size to assume the normality of the samplig distribution of the sample mean. Let's take an example:

> **Example.**
> We have a sample with size n = 100.
> Sample mean $\bar{x}$ = 5.
> Population standard deviation σ =2.5.
> Confidence level is 95% ==> Z = 2
> Standard error, SE = $\frac{\sigma}{\sqrt{n}}$ = 0.25.
> So, the $CI_{95\%} = [\bar{x} - Z * SE, \bar{x} + Z * SE]$ = [4.5,5.5]
> **Interpretation:** Population mean μ will be 4.5 and 5.5 95% of the time.

8.3 Compute confidence Interval: unknown σ

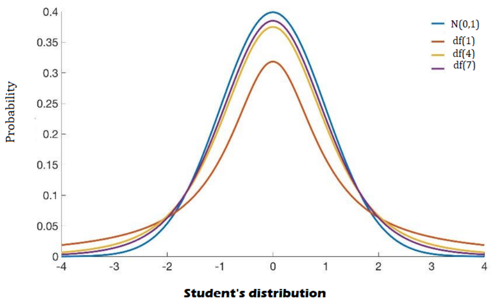
If σ is unknown, we have to estimate it using the sample that we have. Recall, the population σ is used to compute the standard error SE of the sampling distribyution. σ is unknown, so it is estimated by the standard deviation of the sample $\boldsymbol{s}$ that can be calculated by the formula already seen:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{8.4}$$

We use here the corrected standard deviation of the sample by decreasing the degree of freedom by one. Then the standard error is computed:

$$SE = \frac{s}{\sqrt{n}} \tag{8.5}$$

The news is that we cannot use the same z number in the previous confidence interval formula. If the sample size is small we cannot assume the normality of the sampling distribution and thus the number of standard deviation z correspond to a certain level of confidence will be different. So, instead of normal distribution used when σ is known, we use a Student's distribution (t-distribution) when σ is unkown. The t-distribution allows inference through small samples with an unknown population variance.



The Student's t distribution looks much like a normal distribution but generally has fatter tails fatter tails. Faster tails means higher dispersion and therefore there is more uncertainty in the inference. The larger sample size the more the t-distribution looks like the normal distribution. In the table bellow, the number of standard deviation t in t-distribution confidence. *df* is the degree of freedom, it is equal to n-1.

Table 2. number of standard deviation at a level of confidence _ t-distribution

df = n-1	$t_{95\%}$	$t_{99\%}$
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

As you can see, for sample sizes larger than 20, the distribution is almost exactly like the normal distribution ($t_{95\%}$ at df >=20 is equal to $z_{95\%}$, and $t_{99\%}$ at df >=20 is equat to $z_{99\%}$).

So, the formula of the confidence interval becomes:

$$CI_{1-\alpha\%} = [\bar{x} - t * \frac{\sigma}{\sqrt{n}}, \bar{x} + t * \frac{\sigma}{\sqrt{n}}] \tag{8.6}$$

Let's take an example:

> **Example.**
> We have a sample: [5, 6, 2, 9, 3], n=5
> sample mean, $\bar{x} = 5$
> standard deviation of sample, s = 2.73
> standard error of the sampling distribution, SE = $\frac{s}{\sqrt{n}} = 1.225$
> Now, we look at the table of t-distribution to extract t: the number of standard deviation. We have df =5-1 = 4, so $t_{95\%}$ =2.776.
> $CI_{95\%} = [5 - 2.776 * 1.225, 5 + 2.776 * 1.225] = [1.5994,\ 8.4006]$
> Interpretation: population mean μ is between 1.5994 and 8.4006 95% of the time.

8.4 confidence Interval: modern method

We saw in the previous section how we compute the confidence interval based on formula. Formula or tradional method have underlying assumptions that may or may not be true. Furthermore, the Central Limit Theorem doesn't apply to all statistcs 7.1. One educated, but potentially biased opinion on the traditional methods is that these methods are no longer necessary with what is possible with statistics in modern computing, and these methods will become even less important with the future of computing. Instead of relying to theorems, we could simulate the sampling distribution of any statistic. This introduces a technique known as ***bootstraping***. Bootstraping in statistics means sampling with replacement. If we want to bootstrap 5 numbers from this set [1,2,3,4,5,6,7,8,9,10], we could randomly sample these 5 numbers: **set1**:[5,4,3,8,1], and if we want to sample another 5 numbers, each number in the **set1** might end up being sampled again. This is what we mean by sampling with replacement. An example of bootstrap sampling is when we flip a coin or a die several times. The fact that once we flip a head or roll a 6, we can continue to flip or roll that value again means that it is replaced even after being selected. To build an confidence interval for a popullation parameter using bootstrap technique, we follow the step below:

> **Confidence interval with Bootstrap technique.**
> All we have is a **sample** from a population with n size.
>
> **Step 1:** boostrap sampling, i.e sampling the **sample** with replacement
> Without computer, it is impossible to bootstrap manually. In this step we select randomely L bootstrap samples of size <= n.
>
> bootstrap_1 = $[x_1^1, x_2^1, ..., x_n^1]$
> bootstrap_2 = $[x_1^2, x_2^2, ..., x_n^2]$
> .
> .
> .
> bootstrap_L = $[x_1^L, x_2^L, ..., x_n^L]$
>
> **Step 2:** Calculate statistic value from each bootstrap sample
> Statistic here is the mean. But is can be any statistic.

$\bar{x_1}$ = mean(bootstrap_1)
$\bar{x_2}$ = mean(bootstrap_2)
.
.
.
$\bar{x_L}$ = mean(bootstrap_L)

Step 3: $\bar{x_i}$, i=1...L, are used to construct the sampling distribution of sample statistic. In our case sampling distribution of sample mean. This distribution follows a normal distribution.

Step 4: all we have to do now is to build the interval with 95% (or 99%) confidence in the middle of the distribution, using the percentile function. So we cut off the bottom 2.5% and the top 2.5%.

If we truly belive that our collected data are representative to our population of interest, the bootstraping method should provide a better representation of where the parameter is likely to be. However with large sample sizes, confidence interval formula should be provide very similar results to those obtained by bootstraping method. With smaller sample sizes, using traditional methods likely has assumptions that are not true of your interval. Small sample sizes are not ideal for bootstrapping methods though either, as they can lead to misleading results simply due to not accurately representing your entire population well.

The python code to copute the confidence interval is:

Listing 1. Confidence interval

```
import numpy as np
ci_level = 95
nb_boot_samples = 10000
sample_size = len(sample)
means = []
for _ in range(nb_boot_samples):
    bootstrap_sample = np.random.choice(sample,sample_size, replace=True)
    bootstrap_sample_mean =bootstrap_sample.mean()
    means.append(bootstrap_sample_mean)
lower = np.percentile(means, 2.5)
upper = np.percentile(means, 97.5)
ci = [lower, upper]
```

8.5 Confidence interval for the difference in two parameters

We saw so far how to bulid a confidence interval for a single parameter. Sometimes we interest to estimate the difference in two parameters. For instance, we have a question: what is the difference in the birth weight owhen the mother baby is smoker and when she is not smoker. In order to build a confidence interval for the difference in the mean in the birth weights for these groups we can do something similar to what we have done. The python code of the confidence interval in this cas is as follows:

Listing 2. Confidence interval for the difference in two parameters

```
import numpy as np
ci_level = 95
nb_boot_samples = 10000
sample_size = len(sample1)
```

```
diff_means = []
for _ in range(nb_boot_samples):
    bootstrap_sample1 = np.random.choice(sample1,sample_size, replace=True)
    bootstrap_sample_mean =bootstrap_sample.mean()
    bootstrap_sample2 = np.random.choice(sample2,sample_size, replace=True)
    bootstrap_sample1_mean = bootstrap_sample1.mean()
    bootstrap_sample2_mean = bootstrap_sample2.mean()
    diff_means.append(bootstrap_sample1_mean - bootstrap_sample2_mean)
lower = np.percentile(diff_means, 2.5)
upper = np.percentile(diff_means, 97.5)
ci = [lower, upper]
```

diff_means represents the sampling distribution of the difference in the sample means. For instance, if the confidence interval doesn't contain zero, therefore this would suggest that there is a difference in the population means. So that's how we look at the confidence interval for the difference in two parameters.

While building a confidence interval of the difference in two parametrs using formulas continues to require to set up assumptions, the bootstrap technique recommand one thing: the sample data is representative of the entire population.

9 AB testing

As we saw, confidence interval is the procedure of using a sample data to calculate a range of values in which the population parameter is likely falls. The AB testing, on other side, is the procedure of making an assumption (hypothesis) on the population and doing a statistical test to know if the assumption is supported by the sample data or not.

The first thing to do when performing an AB testing is to transform our assumption to two mutually exclusive hypothesis, i.e two opposed hypothese:

1. The null hypothesis denoted as H_0: this is what we believe is true about the population.

2. The alternative hypothesis denoted as H_1: is the contrary to the null hypothesis.

AB testing compares between the two hypotheses and tells us which hypothsis is best supported by the sample data. The null hypothesis would be the mean of the span of writing hand is equal to the span of non writing hand, where the span is distance from tip of thumb to tip of little finger of spread hand. while the alternative is the span of writing hand is not equal to the span of non writing hand.

Here are some rules for setting up null and alternative hypotheses:.
The H_0 is true before you collect any data.

The H_0 usually states there is no effect or that two groups are equal.

The H_0 and H_1 are competing, non-overlapping hypotheses.

H_1 is what we would like to prove to be true.

H_0 contains an equal sign of some kind either $=$, $\leq$, or $\geq$.

H_1 contains the opposition of the null either $\neq$, $>$, or $<$.

In the example of span of writing and non writiong hand, the statements of two hypotheses are;

H_0: $\mu_{writinghand} = \mu_{nonwritinghand}$
H_1: $\mu_{writinghand} \neq \mu_{nonwritinghand}$

As you can see, we denoted the mean by the symbol μ . This is because jypothesis testing is about population not a sample.

9.1 Type of errors

When performing AB testing, we could make mistakes, i.e, choosing the alternative when the null is true or choosing the null when the alternative is true. The first error is called **Type I Error**, while the second is called **Type II Error**. Type I Error is the error that we should avoid the most. Let's orgnize these errors in a table of truth by considering a juridical example: "Innocent until proven guilty". The hypothesis statements of this example is:
H_0: innocent
H_1: guilty

		TRUTH	
		GUILTY	INNOCENT
DECISION	GUILTY	Type II Error	correct choice
	INNOCENT	correct choice	Type I Error

Table 3. Type of Errors.

9.2 Choice between hypotheses

After setting up the hypothesis and fening the type of errors, how do we Choose between hypotheses? We need now to use our sample data to figure out which hypothsis we actually think is more likely o be true. There are two approaches used to choose one of the hypotheses. One is the approach based on confidence intervals, where we simulate the sampling distribution of our statistic, then we calculate the confidence interval in which the population parameter is likely falls. Then we see if our hypothesis is consistent with the confidence interval computed. The second approach is simulating what we belive to be true possible nder the null hypothsis and we see if our data is consistent with that. Let's understand all of that with examples.

9.2.1 How to AB test based on confidence interval

Let's consider this example: The mean of data scientist salary is $113k, according to Glassdoor. We have a sample of data scientist salaries and we want to know if the mean of all data scientist salaries is equal to 113k is supported bu our sample.
sample = [117 313, 104 002 , 113 038 , 101 936 , 84 560 , 113 136 , 80 740 , 100 536 , 105 052 , 87 201 , 91 986 , 94 868 , 90 745 , 102 848 , 85 927 , 112 276 , 108 637 , 96 818 , 92 307 , 114 564 , 109 714 , 108 833 , 115 295 , 89 279 , 81 720 , 89 344 , 114 426 , 90 410 , 95 118 , 113 382]
It is a one sample ab testing and the procedure of the test based on confidence intervals is as follows:
Step 1: State null and alternative hypotheses:

- $H0$: $\mu = \$113\ 000$
- $H1$: $\mu \neq \$113\ 000$

Step 2: Set the confience level. ci = 95%.
Step 3: the statistic is the mean since the population parameter is the mean.
Step 4: simulate the sampling distribution of the sample mean using bootstrap technique.
Step 4: compute the confidence interval using bootstrap technique.
Step5: interpret what we obtained.
below the python code of the procedure of computon the confidence interval.

Listing 3. AB testing using confidence interval

```
import numpy as np
nb_boot_samples = 10000
sample_size = len(sample)
means = []
for _ in range(nb_boot_samples):
      bootstrap_sample = np.random.choice(sample,sample_size,  replace=True)
      bootstrap_sample_mean =bootstrap_sample.mean()
      means.append(bootstrap_sample_mean)
lower = np.percentile(means, 2.5)
upper = np.percentile(means, 97.5)
ci = [lower, upper]
```

The confidence interval obtained is [\$96109.084, \$104120.57]. So the true mean of all data scientist salaries is within this interval with 95% confidence. Now, is the null hypothsis is consistent with our sample data? The null says that the population mean equal to 113000, we lok if the interval falls in the null hypothesis space or in the alternative hypothesis space. As we can see, the interval is entirly below 113000 and therefore we would suggest to **reject the null hypothesis**.

9.2.2 How to AB test based on simulating the null hypothesis

Consider the same sample, the procedure of AB testing by simulating the null hypothesis is as follows:

Step 1: State null and alternative hypotheses:

- $H0$: $\mu = \$113\ 000$
- $H1$: $\mu \neq \$113\ 000$

Step 2: the statistic is the mean since the population parameter is the mean.
Step 3: Before defining this step, it is essentiel to recall some concept. By the central limit theorem, the sampling distribution of the sample mean would follow a normal distribution with the mean is equal to the true mean of the population and the standard error is equal to the population standard deviation divded by the root of sample size. In the AB testing based on confidence interval, we have computed the sampling distribution. The mean of all means (np.mean(means)) is equal to \$100189. One can ask: hey you just said that the sampling distribution follows a normal distribution with a mean is equal to the population mean, but as we can see our sampling distribution has a mean of 100189.93 and that is different to 113000. Right! in fact this is due to the sampling error, or in other word to the effect of the sample size. The more the sample size is closer to the size of entire population, the more the mean of the sampling distribution is closer to true mean of the population.

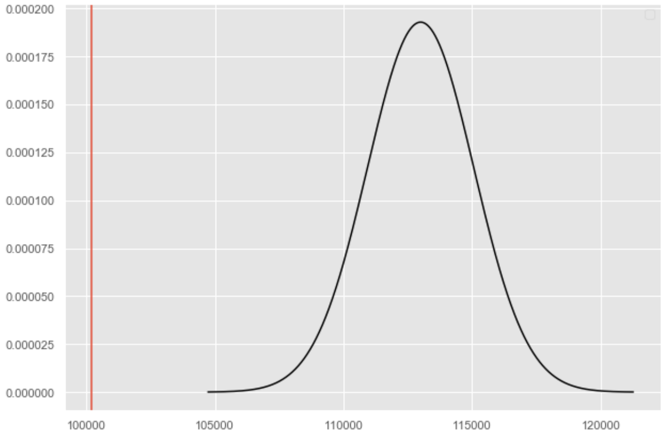
Now let's define the step 3:
We assume that the null hypothesis is true and we know what the sampling distribution would look like if we were to simulate from the closet value to the alternative that is still in the null space. That is, this value of 113000. What we mean by the null is true? we mean that the mean of the means (the mean of the sampling distribution) is equal to the exact value of the population mean.

So, to simulate the null hypothesis we need to know the mean and the standard error. The mean is the closet value to the alternative i.e 113000 .To calculate the standard error we need the standard deviation of the population. As we haven't this value, we can estimate it as explained in the earlier sections. The standard error = the standard deviation of the sample divided by the root of the sample size (SE = $\frac{s}{\sqrt{n}}$). We can directly calculate the standard error from the sampling distribution of the sample mean by computing its standard deviation. Each value in the null distribution represents a possible mean under the null hypothesis.
Step4: calculate the sample mean (called the observed value) and ask the question: where the sample mean falls in the null distribution?
Step5: interpret what we obtained.

Listing 4. AB testing by simulating the null

```
import numpy as np
nb_boot_samples = 10000
sample_size = len(sample)
means = []
for _ in range(nb_boot_samples):
       bootstrap_sample = np.random.choice(sample,sample_size,  replace=True)
       bootstrap_sample_mean =bootstrap_sample.mean()
       means.append(bootstrap_sample_mean)

# compute standard error of the sampling distribution
SE = np.std(means)
# We can also calculate the SE by this method
#SE1 = np.std(sample, ddof =1)/np.sqrt(sample_size)


#Simulate the null hypothesis
mu = 113000
null_val = np.random.normal(mu, SE  , nb_boot_samples)
# the observed value
sample_mean = np.mean(sample)
```

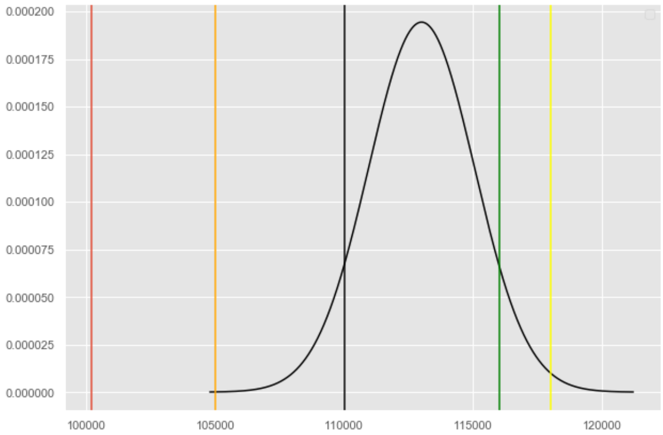
Outputs:
SE = 2049.83
SE1 = 2095.67
As we can see the values of SE and SE1 are very similar.
sample_mean = 100200.37



We can see that the sample mean falls far from the null distribution. Our sample mean is far enough from the tail and we don't think probably it came from the null values. If the sample mean were to fall closer to the center value 113000, it would be a value that we would except from the nul hypothesis and therefore we think the null is more likely to be true.

9.2.3 Probability of value (p-value)

In the previous example, we said that each value in the null distribution represents a probable sample mean under the null hypothesis. We also saw that the sample mean falls far from the null distribution so we don't think it came from the null. That was clear by the plot of the distribution and the position of the sample mean in the plot. Now imagine we had one of the sample mean that falls in the null distribution as shown in the figure below:

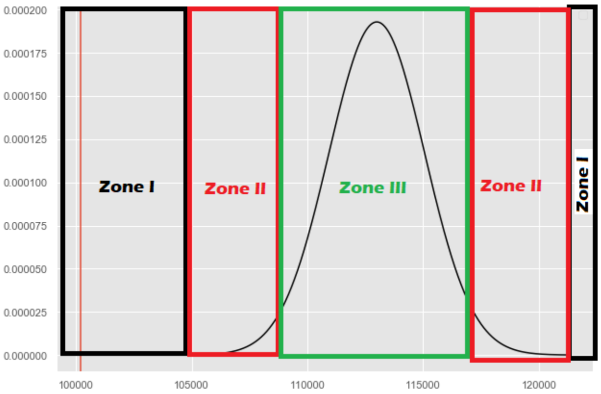


How we can in these cases choose the hypothesis that likely to be true? We recall that our null hypothesis is stated as :

- $H0$: $\mu =$ \$113 000
- $H1$: $\mu \neq$ \$113 000

So the more the sample mean is closer to the center of 11300, the more brobably it came from the null. But how we quantify that? When we can say that a certain sample mean is closer or far enough to the center of the null? To resolve or answer to this question, statiticians proceed as follows:

- Determine the critical zone or insecure zone. It is determined according to the alternative hypothesis sign. In the case of our example, the insecure zone is when we move away from the center. In the figure below, there are 3 zones.



When a sample mean falls in the **Zone I** it is obvious that it didn't come from the null. If the sample mean is on either the ends of the tail -the begining of the **zone II-**, despite it is within the null distribution the sample mean is in insecure zone. We consider the insecure zone the zone of contact with the outside of our null hypothesis. When this zone ends? Statisticians generally set the area of this zone to 0.01 or 0.05. This value is called significance Level α and it represents the type I error, i.e, the alternative is chosen when the null hypothesis is true. In the alternantive hypothesis $\neq$ sign, the two zone I are equidistant from the null center and each area is set to $\frac{\alpha}{2}$ = 0.005 or 0.025.

Now, when the sample mean falls in the zone III we can say that we accept the null hypothesis or **we fail to reject** it, since the null has been assumed to be true. If it falls in the zone I and II **we reject** the null hypothsis.

- As we saw, we can directly choose the hypothesis that is likely to be true based on the plot only. But when we haven't the plot we need to do some calculation to know if the sample mean is within the α (critical) zone or not. We thus calcul a measure called the probability of value (**p-value**) and it defined as follws:

> **Definition 9.1.** p-value is the probability of observing our statistic (or one more extreme in favor of the alternative) if the null hypothesis is true.

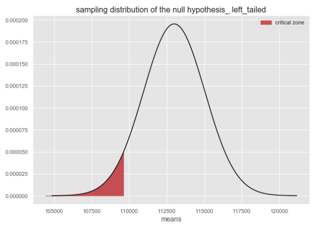
To simplify the definition let's continue with our example. The statisic in the defintion is the mean in our example. We know what is the sample mean, but what about a vlaue that is more extreme in favor of the alternative? In our example the sign of the alternative hypothesis is $\neq$. A value in favor of the alternative is thus a value that is different from the center of the null. Now, observing a sample mean or a value more extreme in favor og the alternative is a value that is far from the null center by a value equal to (null center - sample mean) and this for the two sides. We denoted it as more_ extrem. Finally, p-value is the probability of obtaining more_ extrem value. More precisely, p-value = probability of null values < (null center - more_ extrem) + probability of null values > (null center + more_ extrem). To make e decision , if the p-value is $< \alpha$, that means we are in the critical zone, so we reject the null hypothesis. If the p-value is $> \alpha$, we fail to reject the null.

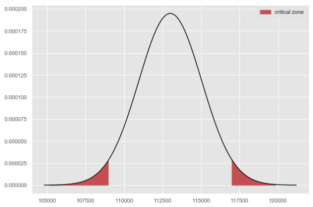
Listing 5. AB testing by simulating the null_ compute the p-value

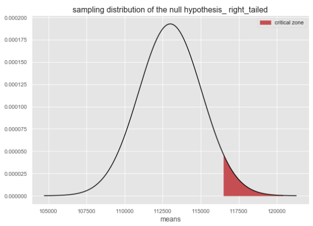
```
import numpy as np
nb_boot_samples = 10000
sample_size = len(sample)
means = []
for _ in range(nb_boot_samples):
        bootstrap_sample = np.random.choice(sample,sample_size, replace=True)
        bootstrap_sample_mean =bootstrap_sample.mean()
        means.append(bootstrap_sample_mean)

# compute standard error of the sampling distribution
SE = np.std(means)
# We can also calculate the SE by this method
#SE1 = np.std(sample, ddof =1)/np.sqrt(sample_size)


#Simulate the null hypothesis
mu = 113000
null_val = np.random.normal(mu, SE , nb_boot_samples)
# the observed value
sample_mean = np.mean(sample)

more_extreme = np.abs(mu - sample_mean)
p-value = np.mean(null_val < mu - more_extreme) +\
                np.mean(null_val > mu + more_extreme)
```

9.2.4 AB Testing types

There are 3 types of AB testing. One of them is that of the example above, when the hypothesis statement contains the = and ≠ signs. This type is known as **two tailed test**. The second type is when the sign of the null hypothesis is ≤ and the alaternative contains the sign >. This test is known as **right tailed test**. The third test is when the sign of the null hypothesis is≥ and the laternative contains the sign <. This test is known as **left tailed test**. As we said, the critical or alpha zone is determined according to the sign in the alternative statement. Let's first take the second test, the right tailed test. The null hypothesis is when the mean (or any other parameter) is less than or equal to a certain value, which is also the center of the null. So, the insecure or critical zone is situated at the end of the right tail. The p-value in this case is a one side probability or area and it is defined as the probability of the null values is greater than the sample mean (the observed value). Similary, the critical zone of the left tailed test is situated at the end of the left tail. The p-value is the probability of null values is less than the sample mean.







As in confidence interval we can AB test the difference in two parameters. Instead of taking the sample statistic we take the difference in the two statistics.