Probability Distribution and Mean Testing

Elie Diwambuena,

June 20, 2023

1. Introduction

Every research begins with a question. We might be interested to know how much men and women make in term of salary. To find the answer, we can go out and travel the entire world to ask men and women their salaries but this is almost impossible. Does it mean we should give up? Absolutely no. Instead of travelling the world to ask women and men about their salaries, we can start from our homestreet. So, we take a seat in front of our house and ask every man and woman that passes by how much they make. Then at the end of the week, we sum the salaries of women and men. Let assume we get 49k in total for men and 51k for women. Are we conviced about this result? Most likely not. This is just the sample of salary taken from our hometown afterall.

The following week, we decide to travel to another country to visit a friend. Once there, we again take a seat in front of the house and ask every men and women that passes by for a week. Let assume we get 30k for women and 70k for men. Again, we are very skeptical about the result. This is just a sample from our friend hometown. Hence, we decide to travel to yet another country to see another friend and repeat the same process. This time we find 60k for male and 40k for female. Now we are even more confused than ever. Of course we do not have the means to travel again, so we return home and start wondering which of the 3 results that we obtained is true. This is a very difficult question. Put differently, we wonder what is the probability or the chance that any of the 3 results is true.

This is the challenge that most researchers face. In the following lines, we attempt to explain the steps that we might "supposedly" go through to answer this question. We say "supposedly" because we are going to take a very long path to reflect and discuss about every steps. Before we proceed, we would like to mention that we will put less emphasis on mathemitcal concepts. In practice, it is not common to solve statistical problems by hand. There are several statistical software that can handle the computations part. What really matters in practice is to understand the meaning of the statistical operations being performed. Hence, in this article we put a greater emphasis on the understanding of statistical concepts rather than on solving complex mathematical expressions. Hopefully, you will be able to connect the dots between the notions and the maths and walk out with a better understanding of the concepts.

2. Distribution and probability

2.1 Average or mean

Before we continue this discussion, let first clarify our research problem. We are not trying to compare the salary of men to the salary of women. Let this be clear because the approach is different for both questions. What we are attempting to do is to simple find what a typical man earn and what a typical woman earn. Now let return to our example. So, we have 3 different samples collected from 3 different places and we want to know which one is closer to the truth. First, we must look at each sample separately. In sample 1, the total salaries are 49k for men dand 51k female. But what information does this total convey? Nothing at all. Suppose we interviewed 40 men and 60 women. What does this imply? We have 60 women for 51k and 40 men for 49k. What should we do? We need to find a measure that accounts for the size of the sample. Thus, we divide 49k by 40 and 51k by 60 to get the salary per person for each group. We find that the salary per person for the men is 1.2 and for women is 0.85. This is known as the **average or mean**. The mathematical expression for the sample average is shown below.

$$\overline{x}(mean) = \frac{1}{n}\sum_{i=1}^{n} x_i = x_1 \ + x_2 \ ... + x_n$$

Where $x_1 \ + x_2 \ ... + x_n$ is the sum of all incomes and n is the number of people within each group sample (men or women).

2.2 Variance and standard deviation

It becomes obvious that the total can be misleading. **Even though our intention is not to compare**, but looking at only the total we might be tempted to believe that women make more than men. However, when we look at the average, the opposite can also be true. Then should we stop here? Not yet. First, how can we be sure that every man makes about 1.2k and that every woman makes about 0.85k? This may or may not be the case. Suppose that 35 out of 40 men that we interviewed make each 0.8k and the remaing 5 make each 4.2k. And suppose that 45/60 women that we encountered make each 1k per month and the remaining 15 makes 0.4k per month. The average values will still be the same. What does this implies?

It implies that the average can also be misleading. Thus, we cannot fully rely on the average neither. We need to find a way to verify whether the average is very close to the actual salary of a typical man or woman that we will meet on the street. For that, we can look at how close or far the salaries that we sampled are to their respective mean. Thus, we take the salary of every men we interviewed and substract the men average salary. We do the same for women, so we take the salary of every women we interviewed and substract the women's average salary.

$$man_1 sal - man_{mean} sal \ ; \ man_2 sal - man_{mean} sal \ ; ...; \ man_{40} sal - man_{mean} sal$$

Similarly,

$$woman_1 sal - woman_{mean} sal; \ woman_2 sal - woman_{mean} sal \ ; ...; \ woman_{60} sal - woman_{mean} sal$$

Then what do we do with these numbers? Some men (like the 5 ones who make 4.2) earn more than the average. Thus, their differences will be positive. Other men (like the 35 who make 0.8) earn less than the average. Thus, their differences will be negative. The same is true for women. If the goal is to know how far each man or woman salary is from their respective means, then the sign really does not matter. Whether you are 1km away from the left of the house or 1km away from the right of the house, you are still 1km away from the house. Thus, we can get rid of the sign by either taking the absolute value of each difference or raising it to the power of 2. We often use the power approach.

One of the reason is to amplify large differences and minimize small differences. Suppose we take two men, one makes 0.8k and the other makes 4.2k. Taking the difference gives us -0.4 for the first man and 3 for the seond man. If we take the absolute value, we get 0.4 and 3.35 but if we raise to the power of 2 we get 0.16 and 9. The second reason is, in some advanced statistical methods we can to differentiate the squared terms to find some coefficients. But we cannot differentiate an absolute value (You do not need to know about this but just keep it in mind).

Now we have computed the differences and remove the sign by raising them to the power of 2. Then what do we do ? Just like we did for the salaries, we are going to sum them up to find the total differences and then divide it by the number of people. We do that for each group separately. The value that we find is called the **variance**. (Think about it! The mean is found by calculating total salaries and dividing by the number of observations, and the variance is found by calculating total differences squared and dividing by the number of observations).

We find a variance of 1.3 for the men and 0.07 for women (try to do it yourself). But how do we interprete this? It is the difference squared in income per person. Hold on! What did we just say? Difference squared per person? This does not convey any meaningful information. Let cancel the power of 2 by taking the square root of the variance (square root is the opposite of power of 2 just like subtraction is the opposite of addition). Now we can simply call it the difference in income per person. This is also not very meaningful but at least it much better than the difference squared in income per person. We call it the **standard deviation**. Put simply, the standard deviation is the average distance from the mean. We find to be 1.14 for men and 0.26 for women. The mathematical expressions for the sample variance and standard deviaiton are shown below.

$$s^2 \ (variance) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$s \ (standard \ deviation) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Now we are able to say how far a given individual salary is from the mean salary that we found. Let say we find a man who makes 2k, then we can say that this man'salary is within the range of 1 standard deviation from the mean. That is mean < that man < mean + 1 standard deviation or simply 1.2 < 2 < 1.2 + 1.14. And if a we find a woman who makes 1.5k, then we can say that this woman'salary is within 3 standard deviation from the mean. That is 0.85 < 1.5 < 0.85 + 0.26 x 3.

2.3 Sampling distribution of sample means (SDSM)

We can expect that the means and variances of the two other samples to be different from the mean and variance of sample 1. Suppose for sample 2, we find a mean of 1.5k with a standard deviation of 0.5 for women and a mean of 1k with a standard deviation of 0.2 for men. And for sample 3, we find a mean of 2 with standard deviaiton of 0.8 for women and a mean of 1.5 with a standard deviation of 0.5 for men.

If it was possible, we could have taken even more samples and we would have gotten a lot of means and standard deviations. Suppose the world working population is 10k people and for every sample we only take 100 people. How many samples of 100 people can we take from this population with remplacement ? You might think about the notion of combination. However, the question here is quite different because we allow for the combination of two identical samples (that is what with replacement means). Let say we have 5 colors (red, black, green, pink, yellow) and we want to take a sample of 2 colors at a time. Without replacement, we can only make 10 samples of size 2 (**bolded** couples). But with remplacement, meaning a color can be taken again after it has been already taken, we can make 25 samples (both bolded and unbloded couples):

- **(r,b), (r,g), (r,p), (r,y)**, (r,r)
- **(b,g), (b,p), (b,y)**, (b,r), (b,b)
- **(g,p), (g,y)**, (g,r), (g,b), (g,g)
- **(p,y)**, (p,r), (p,b), (p,g), (p,y)
- (y,r), (y,b), (y,g), (y,p), (y,y)

So given a population of 10k people and a sample size of 100 people, we can draw $10k^{10}$ samples with replacement. Let generalize this. Thus, for a population of size N and sample of size n, we can make N^n samples of size n.

Think about it for a moment. We only have 3 samples out of $10k^{10}$ possible samples. Are our estimates of mean and standard deviation realistic? Most likely not. Fortunately, we are not the first one to face this problem. Some great statisticians out there have tried to answer the same question and proposed the **central limit theorem**. What this theorem says is that as long as the sample size is greater than 30, we can be sure that the **sampling distribution of sample means** follows a **normal distribution**.

What is the sampling distribution of sample means? Suppose that we actually collect $10k^{10}$ samples and find the mean for each sample. We will have $10k^{10}$ means. Then, we can take the average of those means and this gives us the mean of the means. In fact, the mean of the means should be the same as **the mean of the population** because we have collected the entire population. Furthermore, as usual we should also calculate the standard deviation. Put differently, we should find the distance between each sample mean and the mean of the population. Instead of calling it standard deviation of the mean of the population, we call it the **standard error** to avoid confusion. Likewise, the standard error is the same as the **standard deviation from the mean of the population**. The mathematical expressions for the population mean and standard error are shown below.

$$\mu\ (pop.mean) = \frac{1}{N}\sum_{i=1}^{N}\overline{x_i} = \overline{x_1}\ + \overline{x_2}\ ... + \overline{x_N}$$

$$\theta\ (pop.variance) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

$$\sigma\ (standard\ error) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{x_i} - \mu)^2} = \frac{\sqrt{\sum_{i=1}^{N}(\overline{x_i} - \mu)^2}}{\sqrt{\frac{1}{N}}} = \frac{\theta(pop.variance)}{\sqrt{\frac{1}{N}}}$$

where $\overline{x_1}\ + \overline{x_2}\ ... + \overline{x_N}$ is the sum of all sample means and N is the population size (number of men or women).

Did you notice the small difference between population variance θ and and sample variance s^2 ?

If you haven't, the difference is in the denominator. For the sample variance, we divide by the number of observations in the sample minus 1 whereas for the population variance we divide by the total number of observations.

- **Why do we divide the sample variance by n-1?**

The answer to this question is not very as it will require advanced mathemaical concepts to prove the point we are about to make. Simply note that, with a sample variance we divide by n-1 because we do not want to underestimate it relative to the population variance. The population variance will always be greater than the sample variance because we will sum up a larger number of differences squared ($N > n$). Since our goal is to try to get as to as to the population variance, we divide the difference squared of the sample by a smaller number to make it big or to avoid underestimating it too much. That is the intuition behind it. This is sometimes referred to as the **degree of freedom**.

Then what does a normal distribution means? To answer this question, we also need to explain the conception of probability.

2.4 Probability and distribution

Put simply, a probability is the chance that an event does not happen by pure hazard. That is, it is the level of certainty regarding the occurence of an event. When we say there is a high probability that it will rain tomorrow, we mean that the event of raining tomorrow is very certain or that the rain that we will fall tomorow will not happen by hasard. Of course, to make such claim we need evidence. This is where the notion of **probability distribution** comes in.

A distribution is simply how observations are distributed or divided. Then to talk about a dision, we need to find a dividing point. That is why we need a measure of **central tendancy**. In other words, we need to find the middle point or the point at the center of the data. Then, once we find that point we can see how other data are divided and spread around it. There exists several measures of central tendancy such as **the mean, the median and the mode**. We assume the readers are already familar with these basic concepts.

Finding the central point is just the begining. We also need to know how the data are spread. Thus, we need measures of spread such **standard deviation and interquartile range**. Again we assume the readers are already familar with these basic concepts. When both the mean and the standard deviation are found, we can situate any observation within the distribution. Notice that we have already completed these steps for sample 1 and we can do that for sample 2 and sample 3 as well but for the sake of necessity we will skip it.

Now we can say a person salary is within a distance of x standard deviation from the mean. But is it really meaningful to tell someone that his or her salary is 1 or 2 standard deviation from the mean? Not really. In addition, the problem that we encounter is that the mean and the standard deviation of the samples we found may not reflect the true mean and standard deviation of the population. And to answer this question, we look at the *central limit theorem* which lead us to this point. It says that we can assume a normal distribution as long as our sample size is greater than 30. But why the normal distribution at the first place?

Types of distribution

The choice of a distribution depends primarly on the type of the variables. We can generally classify variables into two types: continuous and discrete. In the case at hand, income is a continuous variable because it can take any value within a given range. Let say within a range of 0 to 10, income can be any value such as 2.45 or 7.999 or 5. However, gender on the other hand is a discrete variable because it cannot take on any value within a given range. It only takes specific value. An individual is either of the female or male sex.

It becomes obvious that the measure of central tendancy and spread for income cannot be the same as the measure of central tendancy and spread for gender. What does it mean to have the mean of gender? Really nothing. Instead for discrete variable, we might be interested in the proportion (i.e.,how many women we have relative to the entire sample). Therefore, their distributions are different. There exists several type of distributions such as the (1) normal distribution (coninous), (2) binomial distribution (two discretes), (3) poisson distribution (more than 2 discretes) and several others. Since we are only interested in the normal distribution in this article, we leave the revision of other distribution types to the readers.

Normal distribution

The normal distribution is first and foremost used for continuous variable. Normal here means symmetrical. It is called normal because data are equally spread around the measure of central tendancy or mean. To better illustrate the concept of normal distribution, Let us take for example the population age.

We know that age is between 0 and 100 for most people (only a handfull make it to 120). Thus, we can guess that the average age is about 50 but it is difficult to guess the standard deviation. Nevertheless, we can ask ourself the following question:

- What are the chances that if we go out on the street we can see people aged around 30 and 70 ? Chances are high.

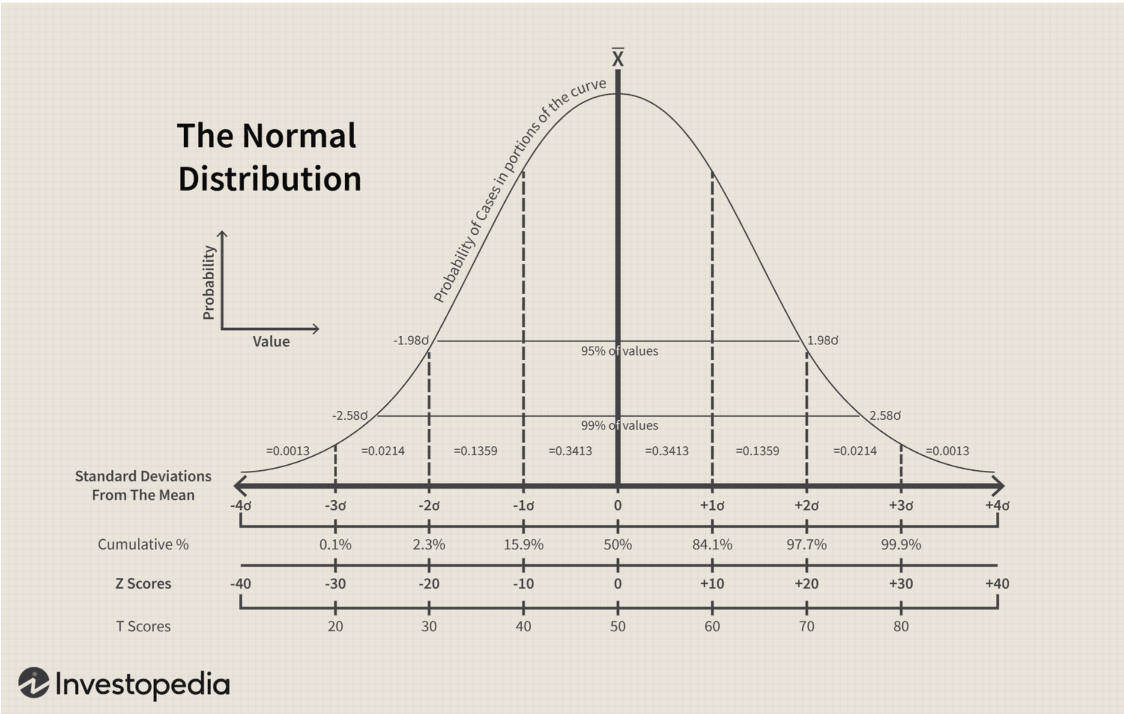
- What are the chances that if we go out on the street we can see people aged around 20 and 80 ? Chances are medium.
- What about the chances that we see people aged below 10 and above 80. Chances are low.
- Now, what about the chances that we see people aged above 90 and below 10. Chances are very low.

Notice that the further we move away from the mean, the lower the probability gets. This is the notion of normal distribution as shown in Figure 1. It is simply the distribution of probability that a given observation is within a certain distance from the mean. And in a normal distribution, most observations are close to mean and the futher we move away the less the observations. Referring back to our example of age, we can say that people aged between 30 and 70 are within 1σ (standard deviation) from $\overline{X}$ (mean). People who are over 90 or under 10 are likely to be within 4σ from $\overline{X}$. The area below the bel-shaped curve that is within each standard deviation distance reprensent the probability. This is shown in figure 2. We can see that the yellow are which represent all observations that are within 1σ from the mean is much bigger than the blue area which represent all observations that are within 4σ from the mean. You may ignore the other annotations such Z score or cumulative percentage for the moment. We will come back to them later.

[1]:
```
from IPython import display
print("\nFigure 1. Normal distribution")
display.Image("normdistr_.png")
```

```
Figure 1. Normal distribution
```

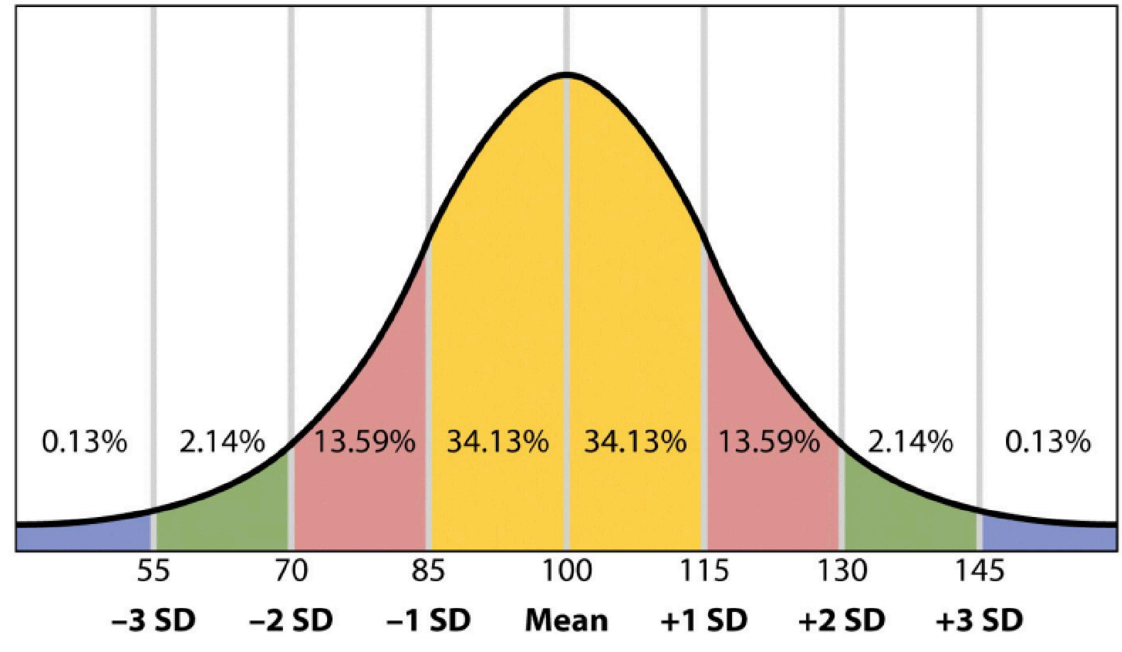
[1]:



[2]:
```
print("\nFigure 2. Normal distribution: area under the curve")
display.Image("area.png")
```

```
Figure 2. Normal distribution: area under the curve
```

[2]:



Let recap. We said that *the central limit theorem* can help us answer the question of whether our sample mean and sample standard deviation are close to that of the population. This central limit theorem assumes that *the sampling distribution of the sample means (SDSM)* follows a normal distribution. And we explain that to find the SDSM we will have to sample the entire population, compute its mean and its standard error.

However, this is IMPOSSIBLE as we already know. Statisticians have done several researches to come up with this central limit theorem. They say that to use *the central limit theorem*, we need to meet the following conditions:

- Our sample must be randomly selected
- Our sample size must be at least 30 not less
- Each sample must be independent from others; meaning sample A outcome should not depend on sample B in any way.

If we meet these conditions, then we can assume that the SDSM follows a normal distribution. Then, this implies that if we are given the mean of a sample, we will be able to say how far it is from the population mean. Of course, we have to know what the population mean is. Put differently, we should at least have an idea of what the population mean is.

- **Can we test for something we don't know?**

Until now we have said that we have no idea about what the population mean and standard deviations for income are. But we have collected 3 samples and found 3 different means and standard deviations. If we really do not know anything about the population, why are we doubting our results? Let think about this for a moment. If no one knows what the income of men and women are, but we have travelled in 3 different countries and get some numbers. We can assume that we now know what it is. After all, who can contradict us? We are the first one to have tried to estimate the salaries of men and women. Because we have 3 samples, we can chose 1 of them based on our own belief and reject the other 2. Or we can compute the average of the 3 sample and say that is the actual salary of men and women. All of these options are possible because nobody has ever done it before. This simply shows that it is difficult to test agaisnt unknown reality. In other words, if there is nothing to test against, then what do we need the test for?. However, in almost every statistical study you will encounter, there is always a known reality that is being tested.

From now on, assume that in our case, a newspaper said that the population average salaries of men is 1.8k and that the average salaries of women is 1.5k. But we have collected 3 different samples and found different means and standard deviations. Hence, we believe the newspaper is wrong. We do not know the true mean ourselves but we think it is different from what the newspaper said. Now we have something to test against. But how are we going to do test that?

First, we should acknowledge that difference in means between samples always occur. We have experienced it ourselves with our 3 samples. Hence, because such difference are very common, we should decide on which difference we can tolerate and which difference we cannot tolerate. In other words, we have to choose a minimum acceptable difference between the means that we are comparing. This minimum acceptable difference is called **the confidence interval**.

With a confidence interval, we are saying that if mean A is within x distance from mean B, then we can accept that they are just the same. However, if mean A is not within x distance from mean B, then they are different.

3. Confidence Interval: Statistical Testing

The discussion of confidence interval goes hand in hand with the notion of statistical testing. Recall we want to test whether what we found is different from what the newspaper found. If it is, then we will reject the claim made by the newspaper. The general approach to statistical testing is as follows:

1. Define the Null and Alternative hypothesis
2. Define a significance level
3. Compute the statistical test score
4. Compare the test score to the critical value
5. Reject or fail to reject the null hypothesis

(1.) For the case at hand, our null is that the mean of the population is equal to the mean of the sample. The alternative hypothesis is that the mean of the population is NOT equal to the mean of the sample. Let think of it this way. The alternative hypothesis is often what we believe. In this case, we believe the mean of the sample is not the same as the mean of the population that the newspaper claims (because if we think it is the same, we do not need to test it again) . Thus, this is our alternative hypothesis. This simultaneously implies that the null hypothesis is the opposite of the alternative hypothesis (meaning that it assumes they are identical).

(2.) A significance level is the minimum level (probability) that is required for the null hypothesis to be "true". That is, if the probability value (**p-value**) of the null hypothesis is higher than or equal to the significance level, we cannot reject the null hypothesis. Put differently, we should "accept" the fact that the null hypothesis does not occur by pure hazard. We are using the quotes because we don't usual accept the null hypothesis or say that the null hypothesis is true but let keep this discussion for later.

A common practice in social sciences is to take a significance level of 5% or 0.005. That is, we want to accept the fact that the null hypothesis is not a hazard if it has a probability of at least 5%. This makes sense if we look at it from a different angle. Consider this example. You flip a coin 100 times every day for 100 days and every time you get exactly 5 heads. Can this really be a hazard? No it cannot. There should be something with your coin that makes it return extacly 5 heads every 100 times you flip it. Likewise, finding a p-value of at least 5% means that there is something in the population that makes the null hypothesis "true". Here we say 5% because our significance level is 5%. If the significance level is 10%, then we would need a p-value of at least 10% or otherwise we should reject the null hypothesis because it does not meet the minimum. Thus, let summarize this as follows:

- If **p-value < sig. level: reject the null hypothesis**
- If **p-value sig. level: fail to reject (accept) the null hypothesis**

(3.) There exists several statistical tests. For a normal distribution, we use the Z-test or t-test. The difference between the two is simply a matter of sample size. If the sample size is less than 30, we use the t test. Else, we use the Z test. The Z test helps us determine how far the sample mean is from the population mean. It is computed as follows:

$$Z = \frac{\overline{x} - \mu}{s_x}$$

In essence, we are dividing the distance between our mean and the hypothesized population mean by the standard deviation of the sample mean. In other words, we are calculating the distance per standard deviation unit.

(4.) Z test: the critical value

The idea behind the Z test is simple. Some great statisticians have calculated the Z score for different types of continuos variables that follow a normal distribution. What happen is that almost every one found different Z scores. So they decided to do it for million of data and found that when the sample size was greater than 30, the result get very close to each other. Hence, they decide to put all of their findings together in a table called Z distribution table. Because it is the Z score from million of data, it is assumed to be the norm. Therefore, when someone compute a Z score, he or she must compare it to the norm. The norm is referred to as **critical value**.

When we choose a confidence level like we did in (2), we are simultaneously chosing a Z critical value. Then, we will simply compare our z value to the z critical value to determine whether to reject or fail to reject (accept) the null hypothesis. Enough theory, let be more practical.

Suppose that we decided to only keep sample 1 and dismiss the other 2 samples just for simplicity. Recall that we found a mean of 1.2 with a standard deviation of 1.14 for men and a mean of 0.85 with a standard deviation of 0.26 for women. And according to the newspaper, the the population average salaries of men is 1.8k and for women is 1.5k. To make thing even simpler, let assume that the standard deviation of the sample is the same as that of the population (because we don't know the population standard deviation. If we know it and they are different, the calculations are a bit different).

We compute the z score for both groups as:

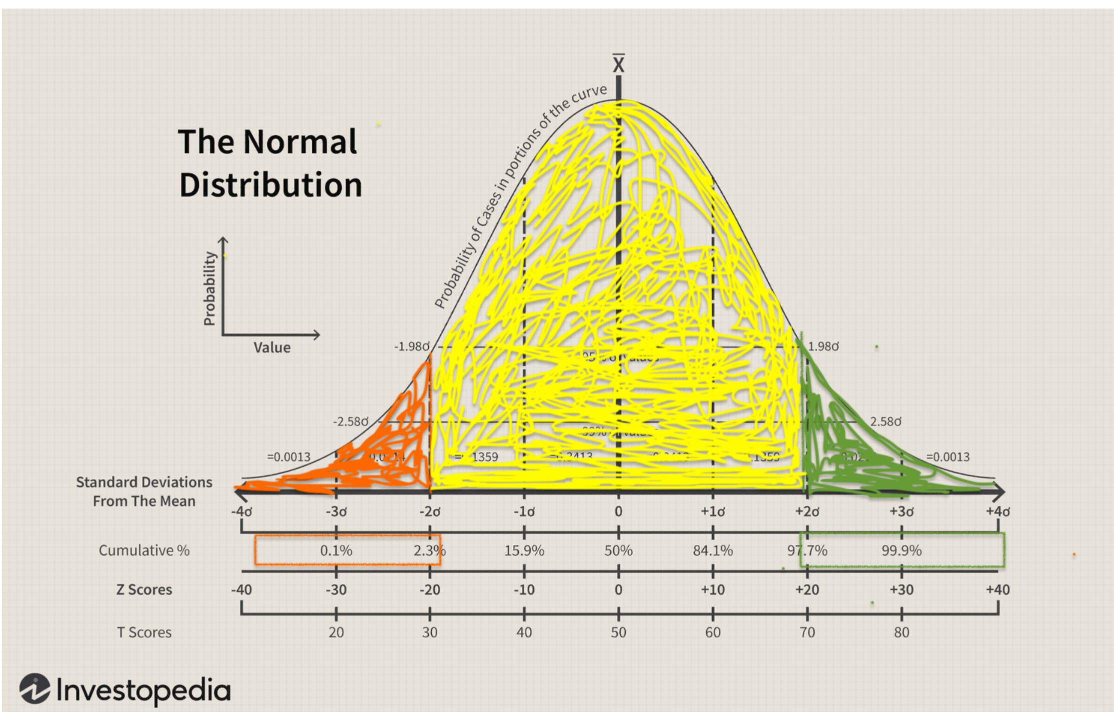
$$Men\ Z\ score = \frac{1.2 - 1.8}{1.14} = -0.526$$

$$Women\ Z\ score = \frac{0.85 - 1.15}{0.26} = -1.153$$

The negative or positive sign does not matter to us. It simply says that the sample mean is at a certain distance from the left of the population mean. Because we are just interested in the distance, we really do not care whether it is from the right or from the left. Then when we say we want a confidence level of 5% we should also specify whether we mean 5% from the left, or from the right or from both side. In this case, we mean 5% from both side so this implies 2.5% from the left and 2.5% from the right to be fair. In other words, we mean if the probability of the null hypothesis falls in the area after the cumulative 97.5% (green rayed area below) or before the cumulative 2.5% (orange rayed area), we should reject it as its probability is not meaningful. However, if it falls within the central area of 95% (yellow), we should not reject it.

[8]:
```
display.Image("normdistr.png")
```

[8]:



- **Comparing the z score to the z critical value**

Before using any Z table, we should always look the direction. In the Z table shown in figure 3, the direction is positive (the right side). So all the numbers given there are the cumulative probabilities from the left to the right-side. However, we are not interested in any direction and we want the cumulative probability from the center. Since we choose a confidence level of 5%, it means that if the probability of the null hypothesis (p-value) is at least 2.5% from the left and 2.5% from the right we cannot reject it. So we should look for a probability of 97.5% inside the table because it shows positive Z values. If the Z table shows the cumulative probability from the right to left or the negative Z values as shown in figure 4, then we should look for a probability of 2.5% in the table.

The column of the Z table should correspond to our confidence interval which is 0.05. Then, inside the column of 5% we shall look for the cumulative probability of at least 97% for the positive table or 2.5% for the negative table. You will notive that both table will lead to the same Z critical value which is 1.9 or -1.9 at the left most column. Since the direction matters less for us, we take 1.9 as the Z critical value. We should also do the same for the Z scores we found. Hence, instead of negative scores we simple take them as positive.

Now we can compare the Z critical value to our Z scores. If the Z score is greater than the Z critical value, we should reject the null hypothesis. Else, if the Z score is smaller than the Z critical value, we should not reject the null. Why? Because a Z score is directly linked to the confidence level. Recall we said that if the p-value we find is below the confidence level, then we will fail to reject the null. Both concepts are intertwined.

In our case, both Z scores are smaller than the Z value. For the men we find $0.526 < 1.9$ and for the women we found $1.153 < 1.9$. Therefore, we fail to reject the null hypothesis. This means that the newspaper's claim is not absolutely wrong. Put differently, it means that we did not find enough evidence to suggest that the newspaper is wrong. Yet, it does not mean that the newspaper is right (LOL! isn't it funny!). This is why we do not say "we accept the null hypothesis" but we rather say "we fail to reject the null hypothesis."

The concepts we have reviewed in this article are fundamental to statistics. Almost every statistical research will involve a statistical test. Here, we have tested the sample mean against some believed or hypothesis true population mean using the Z test for the mean. In other reasearches, different types of tests can be used depending on the question being asked, the distribution of the variable involved and many other factors. For instance, if our question was to compare the mean salary of men to the mean salary of women, then we would have used the Z test for the difference in means. Although this is a completely different test, the approach used here still hold. Hence, understanding the concepts explained in this article is crucial.

4. Conclusion

We started our discussion with a simple yet difficult question: how much do men earn and how much do women earn. The question is Not how much men earn compared to women because that would have implied a completely different approach. In that process, we collected 3 samples and computed their means and standard deviations. However, we observed that the means and standard deviations of the 3 samples are all different. This makes it hard to say which one of the 3 is the true mean of the population or the closest to the true mean of the population. Of course, we have to test them.

However, to test for anything there should be a certain standard or a starting point. But in our case, we assumed we do not know anything about the population. Then, there is no point where to start. In other words, how can we test that one of the 3 means we find is the true population mean if we do not have any idea or belief about the population mean? We really can't do that. Thus, we need a starting point and we supposed that a newspaper made a claim about the population mean salary. Because what we found in the 3 samples were different from what the newspaper claimed, this makes a good case for conducting a statisitical test.

To this end, we want to test if the mean we found is actually not that different from the mean that the newspaper found. By this we mean, if the mean we find is in fact very close to what the newspaper found based on some standard measures (not just comparing them side by side), then we would conclude that they are pretty much the same. Otherwise, we will conclude that they are different. Hence, we define the null and alternative hypothesis and choose a confidence level. Then, we compute the Z score and compare them to the Z critical values. Since our Z scores were below the the Z critical values, we fail to reject the null hypothesis and concluded that we do not have enough evidence to dispute the claim made by the newspaper.

[6]:
```
print("\nFigure 3. Z distribution table")
display.Image("Z.png")
```

```
Figure 3. Z distribution table
```

[6]:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999							

[7]:
```
print("\nFigure 4. Z distribution table")
display.Image("Z-.png")
```

```
Figure 4. Z distribution table
```

[7]:

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002

[ ]: