SHAPEAI

Population vs Sample

Populatip refers to the entire group of people or the objects on which we want to draw some conclusion using the data, while sample refers to the actual group of people or the objects on which observations are being made to draw conclusion for any population.

SHAPEAI

Let's say, we want to calculate average life for any mosquito, then in that case, all the mosquitos becomes our population. But it's impossible to conduct experimment on all the mosquitos, so we pick some mosquitos to conduct the experiment. That is our sample.

Population

Sample

SHAPEAI

- Sample is a subset of the population.
- Sample <= Population.
- Reasons for sampling:
    - Sometimes it's impossible to conduct observations on whole population.
    - Sampling is easier and faster
    - Cost for observing is less
    - Storing and running statistics on smaller dataset is easier and feasible.

SHAPEAI

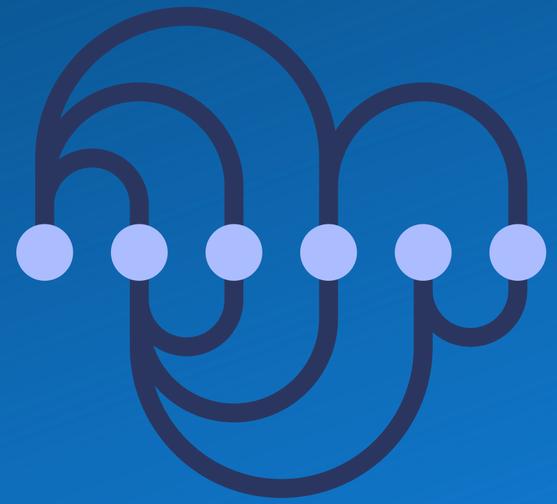
Correlation vs Casuation

In stataistics, two variables are termed to be realted, if the value of one changes if the value of another one increases or decreases. Correlatin and Casuation are smothing related to that but more often misunderstood.

SHAPEAI

Correlation

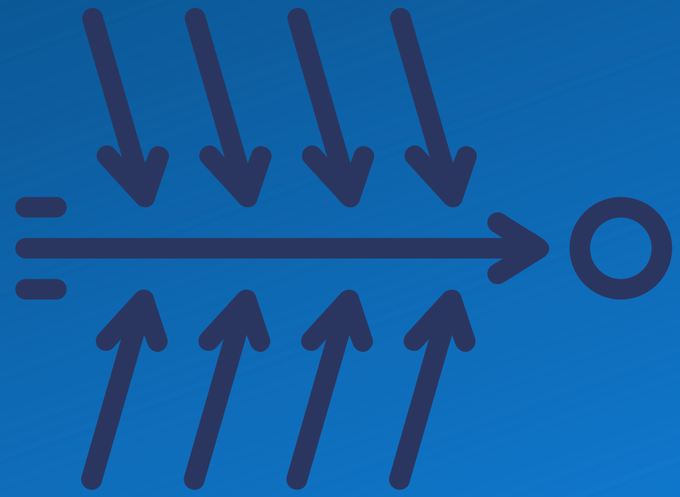
We go to buy a house, the two variables, the size of the house and the price of the house are corelated with each other. Bigger is the size of the house, larger is the price. So we may say the two vatiables are correlated.



SHAPEAI

Casuation

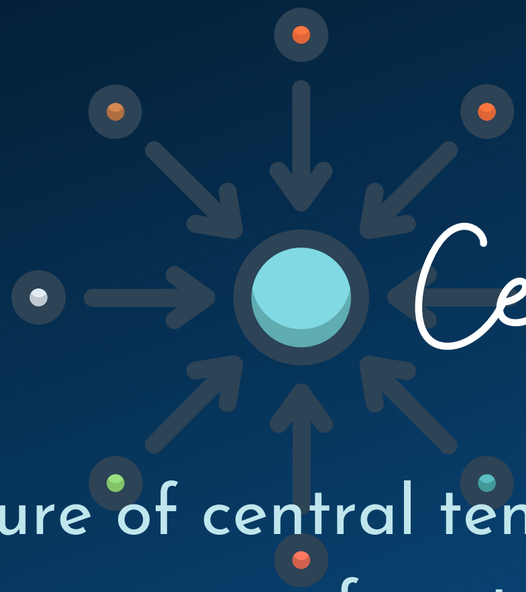
Casuation indicates, that event A is result of occurance of the event B. In the previous example, we may say that there is a correlation, but can we say, that there is casualtion between the variables as well.



SHAPEAI

Casuation Cotd...

No, we can't say there is a causation for sure. If the house A is expensive than the house B, we can't say that it's because the house A is bigger than house B. There might be some other reasons deciding the price of the house. So, the event ($A > $B), is not caused by (size(A) > size(B)).



SHAPEAI

Central Tendency

The measure of central tendency is a measure of the central value of the data. The measure of central tendency is also called a measure of central location of the data.

To measure the central tendency, we generally use either of the three measures: Mean, Median, Mode.

Mean

SHAPEAI

Mean reffers to the average of the data points.
It's a most popular measure to calculate the centeral tendency.

DISADVANTAGE:

- The mean value get's affected a lot by the presence of outliers.

Outliers are basically the exceptional points in the dataset, with either very large or very small value

SHAPEAI

Median

Mean is the middle value in the dataset arranged in order of the value (either ascending or descending).

SHAPEAI

Mode

- Mode reffers to the most frequent value in the dataset.
- Mode is geenerally used for the categorical data to find out the most frequent class of data.
- Mode ususally gives bad result in case of continious data.
- Mode fails to give good measure of centeral tendency when the most common value is far away from the rest of the values.

SHAPEAI

Variability

Variability for any data describes how dispersed the data is (or how spread out the data is). Variability sometimes termed as spread or dispersion. Generally we try to describe the variability of any data using these 4 measures:

- Range
- Interquartile range
- Variance
- Standard Deviation

SHAPEAI

Range

Range is the difference between the maximum value and the minumum value in the dataset. It's the simplest measure one can use to describe the variability.

- Presence of outliers makes the result worse.

SHAPEAI

Interquartile Range

It's similar to range, but the difference is, we are not just concerned with the difference between the max and min value in the dataset, but we are also concerned with where the middle values are lying.
The interquartile range is calculated using the formula:

$$IQR = Q3 - Q1$$

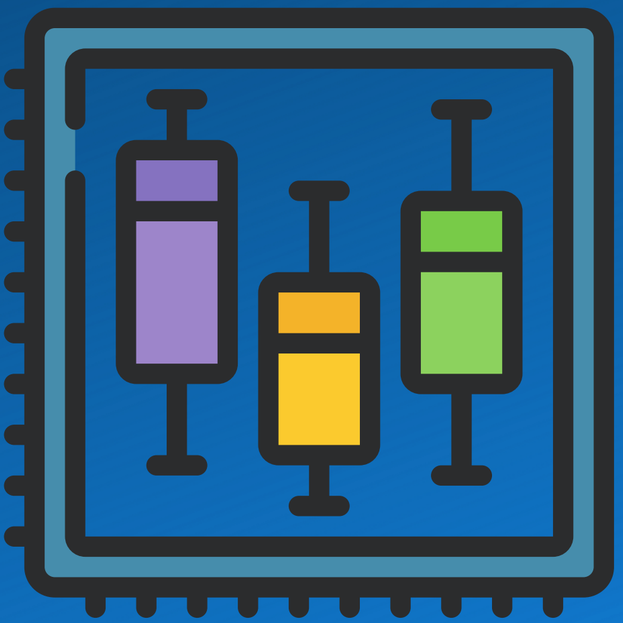
Q3: 3rd Quartile (75 Percentile)

Q1: 1st Quartile (25 Percentile)

SHAPEAI

Interquartile Range Contd...

Boxplot in python is quite useful when it comes to find the "Interquartile Range". It devides the data as outliers, and the quartiles.



SHAPEAI

Variance

Variance tries ton calculate the spread of the data using the squared of the distance of the data points(xi) from the mean($\bar{x}$) of the data points.

$$\text{Population Variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

img source: cuemath

Standard Deviation

Standard deviation is similar to variance, as it's just the sqaure root of the variance.

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

img source: cuemath

SHAPEAI

Scaling Data

Data scaling refers to change the range of the data. When we work with data with multiple features, it is important to have the different features within the same range of numbers so that one feature doesn't over power another one. There are two kinds of scaling that we normally use:

- Normalization
- Standardization

SHAPEAI

Normalization

Normalization means to scale all the values between 0 and 1.

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

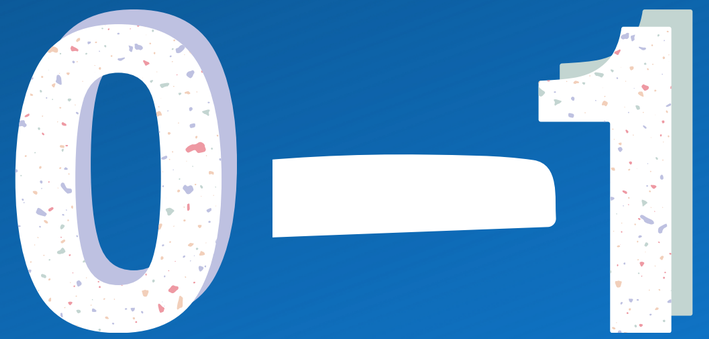
SHAPEAI

Standardization

Standardization means to scale the values such that the mean of the data is 0 and the standard deviation is 1.

$$Z = \frac{x - \mu}{\sigma}$$

z = *standard score*
x = *observed value*
μ = *mean of the sample*
σ = *standard deviation*

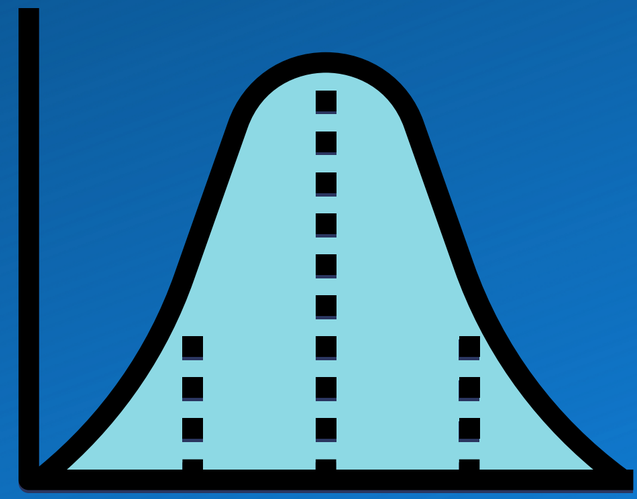


SHAPEAI

Normal Distribution

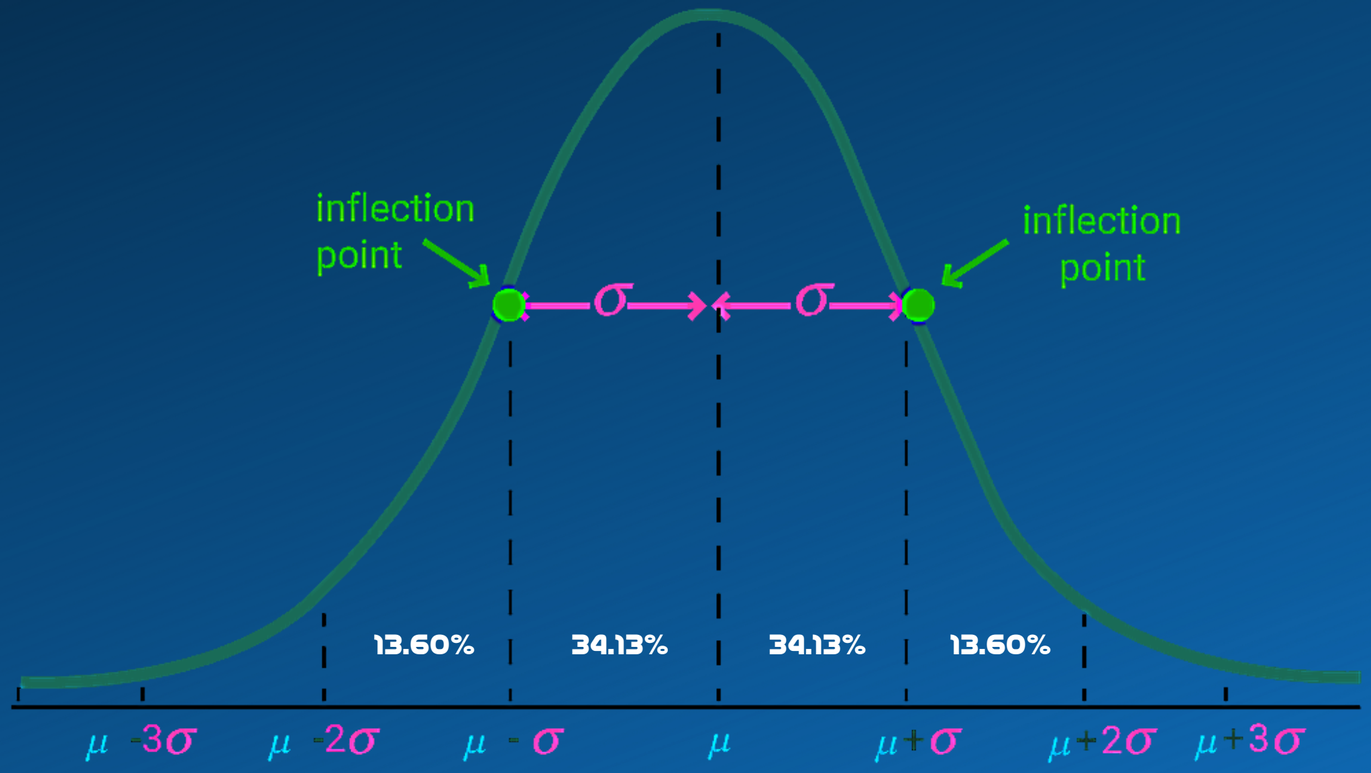
Normal distribution (also called as Gaussian Distribution) is a common probability distribution that tells us how closely are the values gathered around the mean value. The shape of the normal distribution depends on the standard deviation and the mean of the dataset.

- The normal distribution curve is bell shaped (also called as "bell curve" or "Gaussian curve").



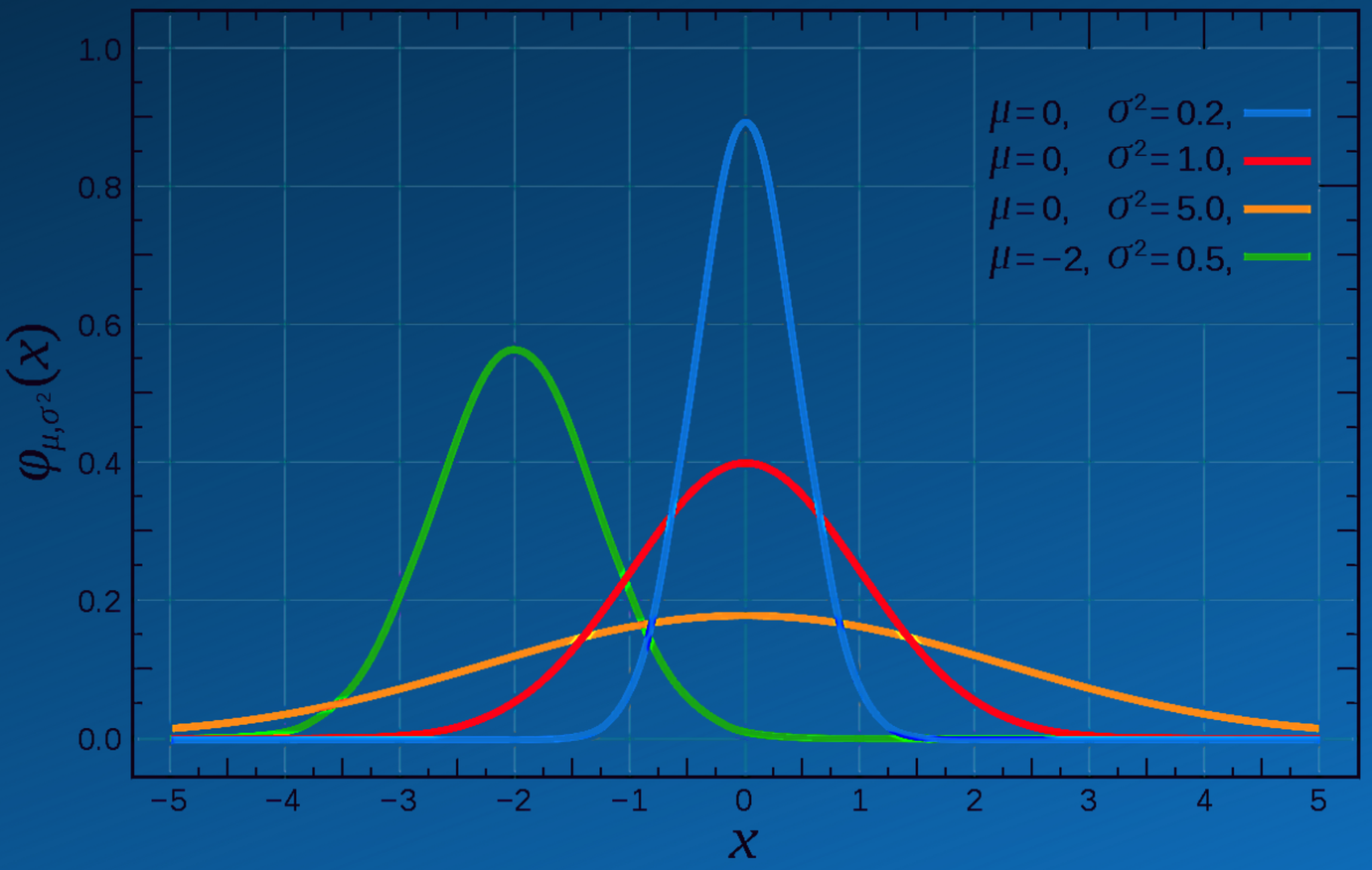
Normal Distribution Cotd...

SHAPEAI



img source: https://calcworkshop.com/wp-content/uploads/normal-distribution-curve-1.png

Normal Distribution Cotd...

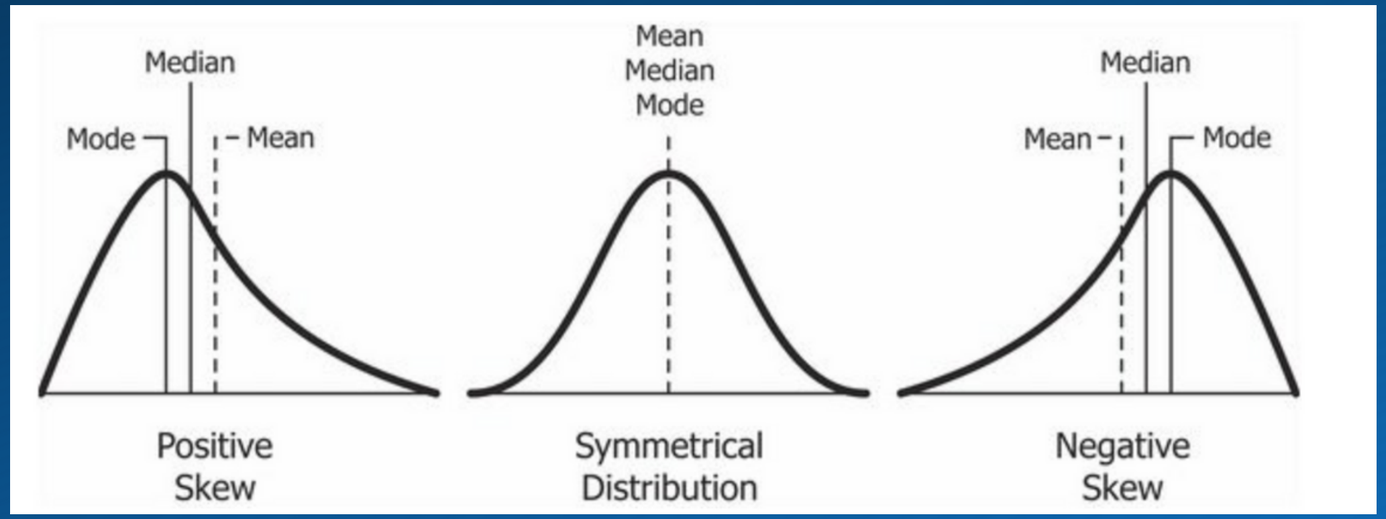


img source: https://upload.wikimedia.org/wikipedia/commons/thumb/7/74/Normal_Distribution_PDF.svg/1200px-Normal_Distribution_PDF.svg.pngpng

SHAPEAI

Skewness of Data

Any data is said to be skewed when the distribution curve of the data seemed to be distorted either towards right or towards left.

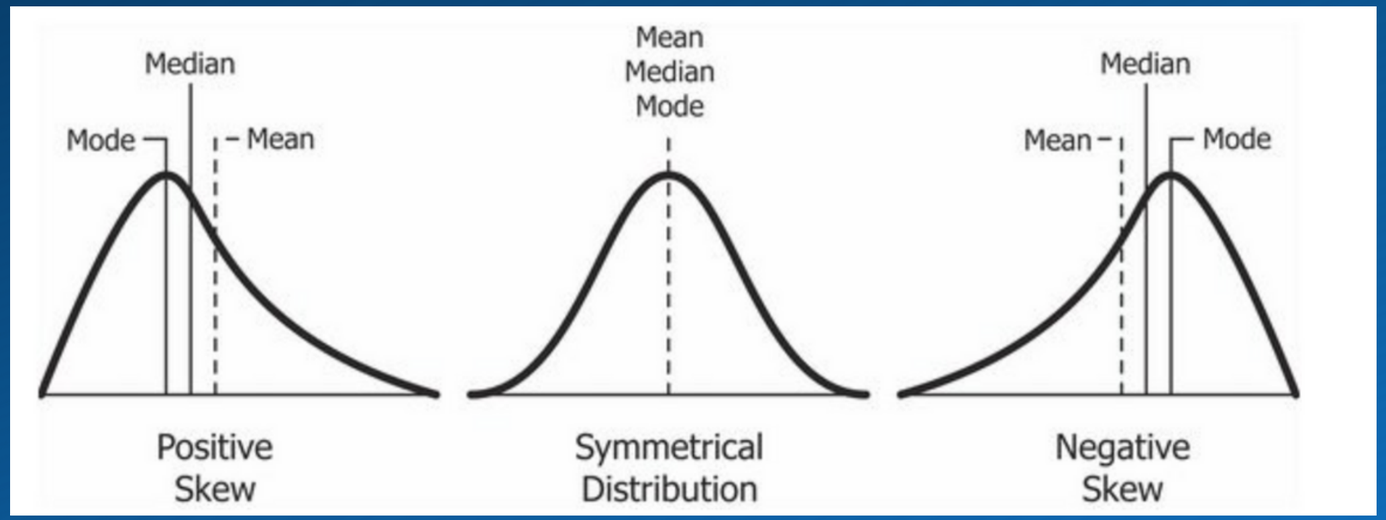


img source: https://cdn.analyticsvidhya.com/wp-content/uploads/2020/06/sk1.png

SHAPEAI

Skewness of Data Contd...

Positive skew is also called the left skew as the data points are concentrated (skewed) on the left. For the similar reason, the negative skew is also called the right skew.



img source: https://cdn.analyticsvidhya.com/wp-content/uploads/2020/06/sk1.png

Skewness of Data Contd...

As in the left skew, data is concentrated in the left, the Median is less than the mean (as outliers in the right shift the mean towards right). Similarly, in the right skew, the Median is greater than Mean.

Also data is highly concentrated in left in left skew, the mode exists in the left portion of the dataset and exists in the right portion of the dataset in the right skew.

SHAPEAI

Hypothesis Testing

Let's say we are given two distributions. Now we do some statistics on those distribution and try to make hypothesis, whether the two distributions belong to the same population or are they significantly different. The hypothesis which says the two belongs to the same population is termed as Null Hypothesis, while the other one is termed as the Alternate Hypothesis.

To check the correctness of the hypothesis, we use tests called:

- t-test
- z-test

t-test or z-test

SHAPEAI

Both t-test and z-test are similar and even share the similar formula. The difference is that when the sample size/distribution size (number of points in the dataset) is greater than 30, the test is called z-test and for the lesser number of data points, the test is called the t-test.

t-test or z-test

SHAPEAI

$$z \text{ or } t = \frac{(\bar{x1} - \bar{x2})}{\sigma_{x1x2}}$$

$$\sigma_{x1x2} \sqrt{\frac{\sigma_{x1}^2}{n1} + \frac{\sigma_{x1}^2}{n2}}$$

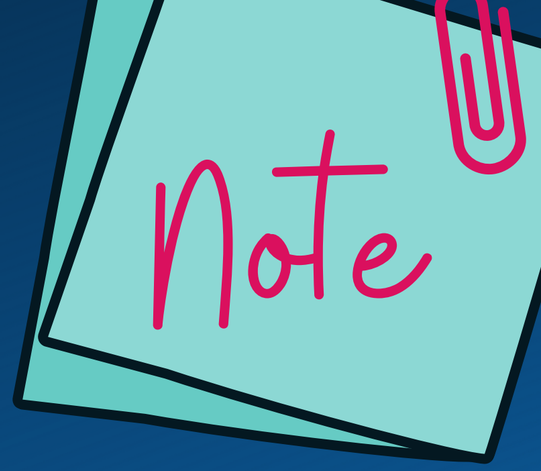
- sigma represents the standad deviation for the two samples x1 and x2
- x1 bar and x2 bar represents the mean for the sample x1 and x2 respectively

(depending on the sample size we will either use z or t)

Steps to check the hypothesis

- Check the assumption (make sure the distribution behaves similar to Normal Distribution)
- Check weather the testing is one-tail testing or two-tail testing.
- Set the significance level.
- Calculate the value of t or z (based on the number of samples)
- If the value of t or z lies outside the significant region then we can say the hypothesis is wrong else we can say that we can not prove the hypothesis wrong.

SHAPEAI



We always to try to prove the null hypothesis wrong. If the t or z value for the null hypothesis falls outside the significant region, then we say the null hypothesis is wrong so the alternate hypothesis is correct. But the t or z value falls inside the significant region, then we just say that we could not prove the null hypothesis wrong.

And also we are checking for the null hypothesis, so we assume that both sample comes from the same population.

SHAPEAI

One tailed test or Two tailed test

One tailed test only check the difference between the groups only in one direction, while two tailed test determine the differnce between the groups in both direction (i.e. it allows both positive and negative difference).

SHAPEAI

Example One tailed test

Two distributions are given, one is for the height of girls and other is for the height of boys.

Hypothesis: Boys are generally taller than the girls.

Here we are concerned with the difference in only one direction (i.e only the positive difference between the height of boy and girl)

SHAPEAI

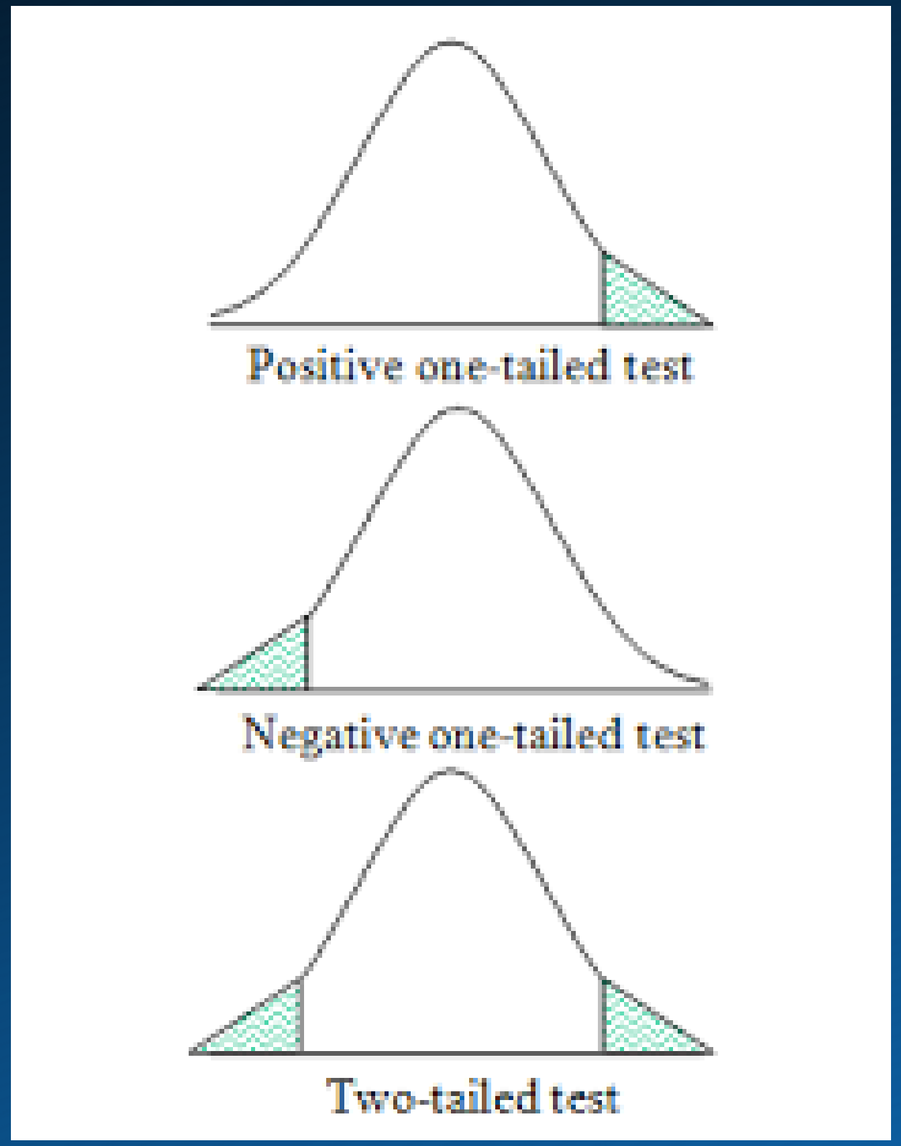
Example Two-tailed test

Two distributions are given, one is for the height of girls and other is for the height of boys.

Hypothesis: There is a significant difference in heights of boys and girls

Here we are concerned with the difference in any direction (i.e both the negative and the positive difference between the height of boy and girl)

SHAPEAI



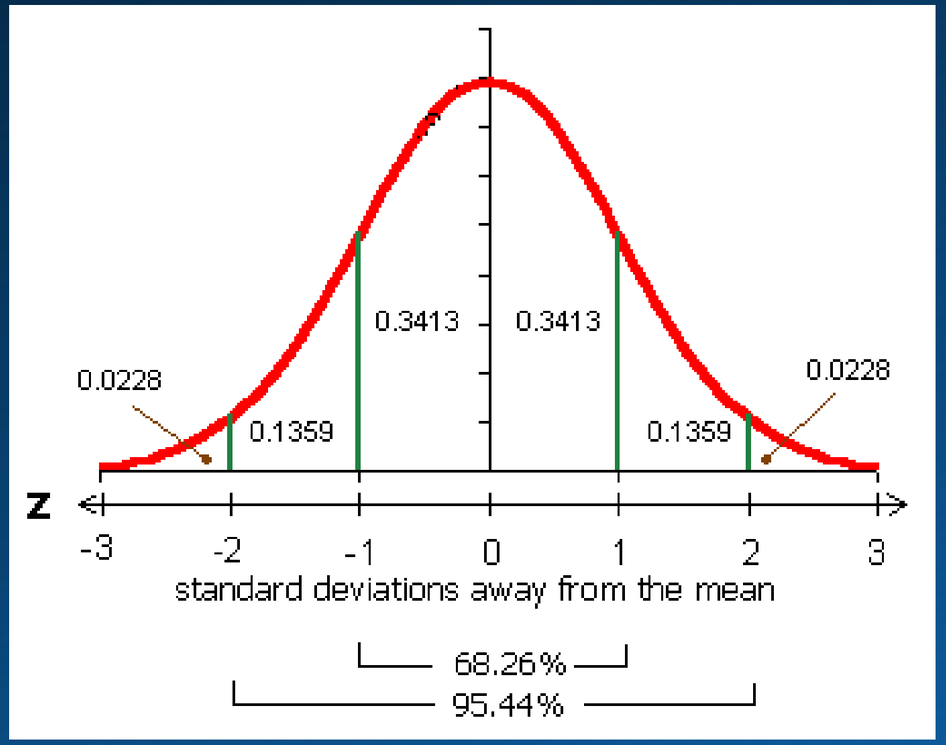
Difference in one direction is considered only

Difference in one direction is considered only

Difference in both the direction is considered

img src: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTt6QuWdTk1YYv40rRclCAVajqnro9uXAzOy1dtkfMyDWPIHA89EUmukDppTZtvE2zPyZk&usqp=CAU

SHAPEAI



The figure shows the percentage of data points lying within the marked range of standard deviation.

img src:https://i0.wp.com/www.comfsm.fm/~dleeling/statistics/standard_normal_curve.png

SHAPEAI

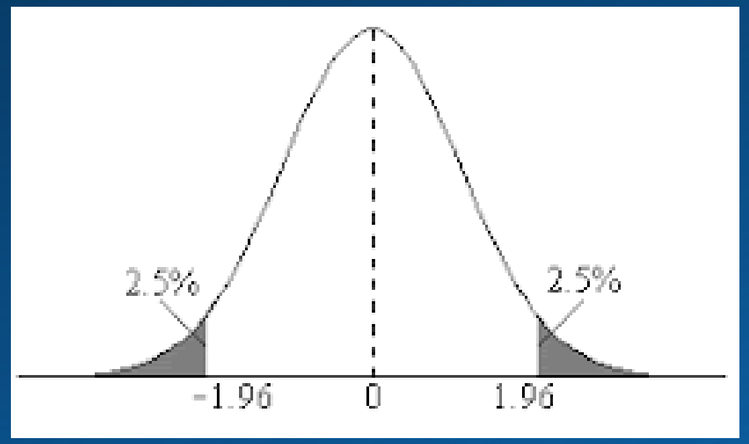
Example

Let's say for a company product the average price is $8.95 and the standard deviation is $0.40 and total number of samples is 200. For the company 2 the average price is $0.60 and the stndard deviation is $0.60 and total number of samples is 175. Is there a significance difference in the price of products of the two companies with the significance tolerance of 5%.

SHAPEAI

Solution

The problem is a two tailed problem as we are concerned with the difference in the price. The difference in the price can either be positive or negative.
So the outer 5% of the population will be considered as 2.5% on both sides of the normal distribution like shown in the picture.



SHAPEAI

Solution contd...

$x1 = 8.95$, $x2 = 9.10$, $\sigma 1 = 0.40$, $\sigma 2 = 0.60$

$n1 = 200$ and $n2 = 175$

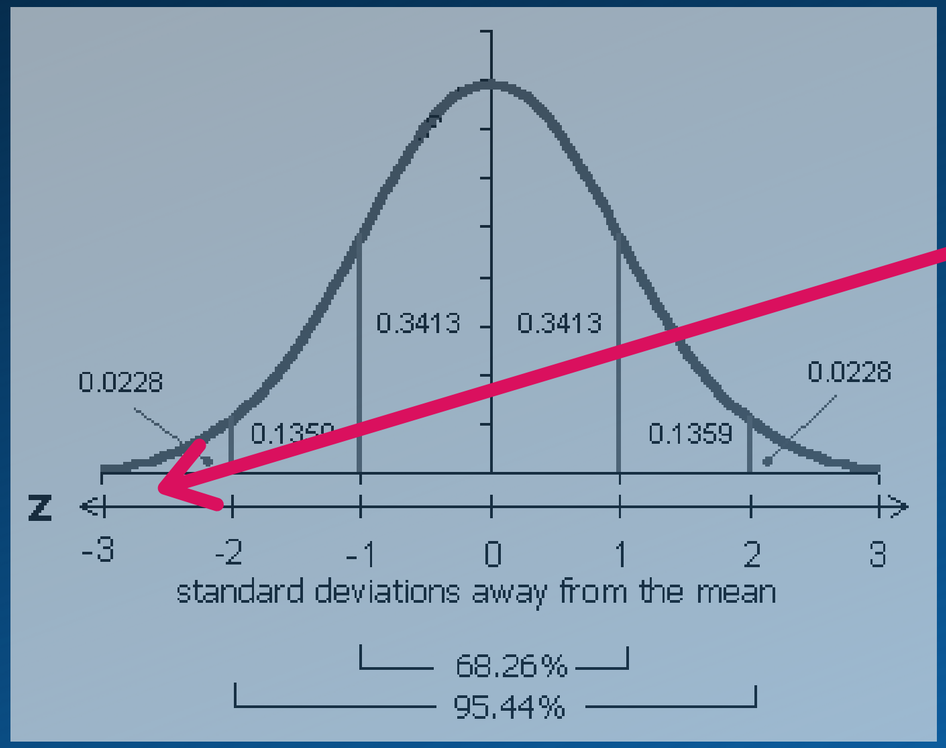
using

$$z = \frac{(\bar{x1} - \bar{x2})}{\sigma_{x1x2}}$$

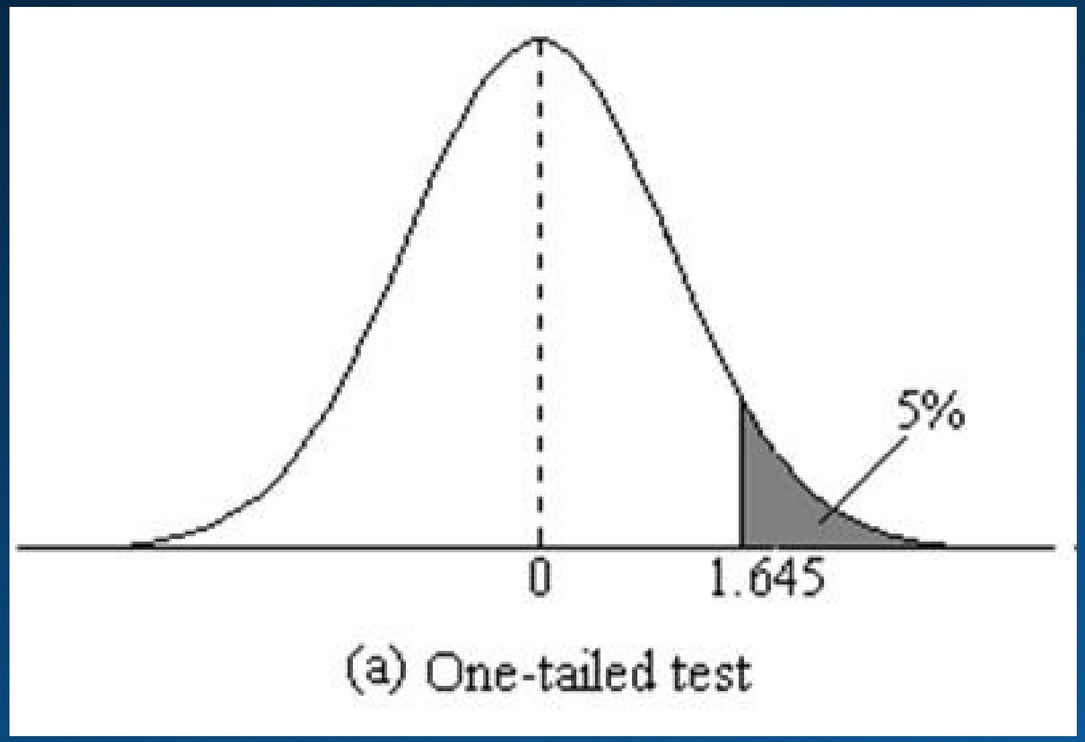
$$\sigma_{x1x2} \sqrt{\frac{\sigma_{x1}^2}{n1} + \frac{\sigma_{x1}^2}{n2}}$$

$$z = -2.83$$

Solution contd...



-2.83 lies there which is out of the 95% of significance population. So the null hypothesis (that there is no significant difference in the price of the sample) is wrong. So alternate hypothesis must be true that there is a significant difference.



(a) One-tailed test

For a positive one tailed test, outer 5% of the population lies in the shaded region of the normal distribution (i.e. standard deviation more than 1.645)