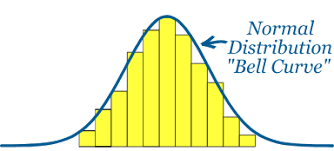
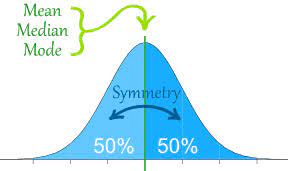
**Statistics Assignment 3**

**Name = Khushal tondon**

**Ans no. 1**

If we have distribution and that are following Bell curve distribution

The blue curve is a Normal Distribution. The yellow histogram shows some data that follows it closely, but not perfectly (which is usual) . It is often called a "Bell Curve" because it looks like a bell.



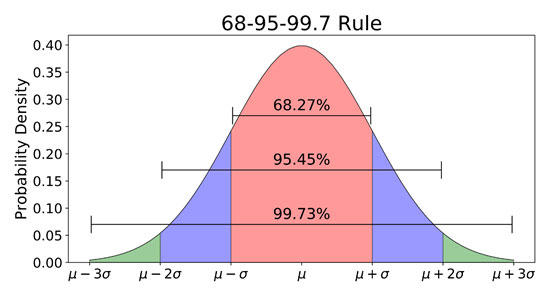
Empirical formula

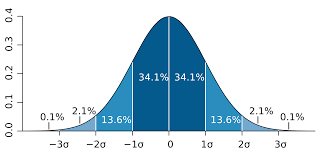
68 – 95 – 99.7% rule , which basically tells that if the dataset is Gaussian distribution then

* 68 - In first region around 68 % of the entire distribution is present there

And basically In first standard deviation , one S.D towards right and one S.D. towards left

* 95 - within two S.D region around 95% of entire distribution lies in this region
* 99.7% - In the third S.D. around 99.7 % of entire distribution lies in this region .

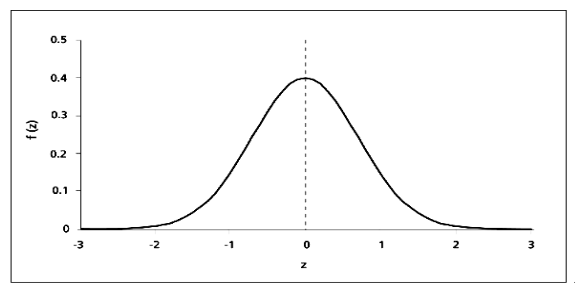




**Ans no. 2**

A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.

It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution.



The formula for calculating a z-score is is

z = (x-μ)/σ,

where x is the raw score, μ is the population mean, and σ is the population standard deviation.

It is useful to standardized the values (raw scores) of a normal distribution by converting them into z-scores because:

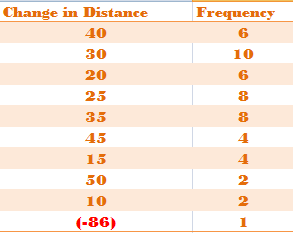
(a) it allows researchers to calculate the probability of a score occurring within a standard normal distribution;

(b) and enables us to compare two scores that are from different samples (which may have different means and standard deviations).

**Ans no. 3**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

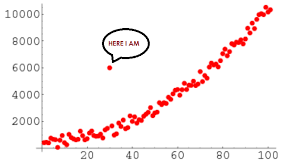
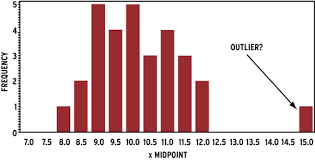
In this example, -86 is an outlier. An outlier is any value that is numerically distant from most of the other data points in a set of data. We know that -86 is far below any of the other values in our data set. It is not uncommon to find an outlier in a data set.



The most common source of outliers is measurement error.

Other sources of outliers include:

* Human error (i.e. errors in data entry or data collection)
* Participants intentionally reporting incorrect data (This is most common in self-reported measures and measures that involve sensitive data. )
* Sampling error (i.e. including high school basketball players in the sample even though the research study was only supposed to be about high school track runners)



**Ans no . 4**

**Dealing with Outliers**

1. **Deleting the values:** we can delete the outliers if we know that the outliers are wrong or if the reason the outlier was created is never going to happen in the future.

For example, there is a data set of peoples ages and the usual ages lie between 0 to 90 but there is data entry off the age 150 which is nearly impossible. So, we can safely drop the value that is 150.

For finding out an outlier we can use Box plot too.

1. **Changing the values :** We can also change the values in the cases when we know the reason for the outliers.

For example , where we had 10 voltmeters out of which one voltmeter was faulty. Here what we can do is that we can take another set of readings using a correct voltmeter and replace them with the readings that were taken by the faulty voltmeter.

1. **Data transformation** : Data transformation is useful when we are dealing with highly skewed data sets. By transforming the variables, we can eliminate the outliers for example taking the natural log of a value reduces the variation caused by the extreme values. This can also be done for data sets that do not have negative values.

1. **Using different analysis methods :** We could also use different statistical tests that are not as much impacted by the presence of outliers – for example using median to compare data sets as opposed to mean or use of equivalent nonparametric tests etc.
2. **Valuing the outliers :** In case there is a valid reason for the outlier to exist and it is a part of our natural process, we should investigate the cause of the outlier as it can provide valuable clues that can help you better understand your process performance. Outliers may be hiding precious information that could be invaluable to improve your process performance. You need to take the time to understand the special causes that contributed to these outliers. Fixing these special causes can give you significant boost in your process performance and improve customer satisfaction. For example, normal delivery of orders takes 1-2 days, but a few orders took more than a month to complete. Understanding the reason why it took a month and fixing this process can help future customers as they would not be impacted by such large wait times.

**Ans no. 5**

The variance is a way to measure the spread of values in a dataset.

The formula to calculate population variance is:

σ2= Σ (xi – μ)2 / N

where:

Σ: A symbol that means “sum”

μ: Population mean

xi: The ith element from the population

N: Population size

The formula to calculate sample variance is:

S2 = Σ (xi – x)2 / (n-1)

where:

x: Sample mean

xi: The ith element from the sample

n: Sample size

Notice that there’s only one tiny difference between the two formulas:

* + 1. When we calculate population variance, we divide by N (the population size).
    2. When we calculate sample variance, we divide by n-1 (the sample size – 1).

When calculating the sample variance, we apply something known as Bessel’s correction – which is the act of dividing by n-1. Also known as degree of freedom .

So ,

Bessel correction refers to the n-1 part used as the denominator in the formula of sample variance because by this it reduces the gap between the sample mean and population mean . it is not underestimating the true population variance .