Introduction to Descriptive Statistics and Probability for Data Science

Abhishek
Kumar

Table of Contents:

- Introduction
- Measure of Central Tendency (Mean, Mode, Median)
- Measures of Variability (Range, IQR, Variance, Standard Deviation)
- Probability (Bernoulli Trials, Normal Distribution)
- Central Limit Theorem
- Z scores

Introduction:

In Descriptive statistics you are describing, presenting, summarizing, and organizing your data, either through numerical calculations or graphs or tables. Some of the common measurements in descriptive statistics are central tendency and others the variability of the dataset.

Descriptive statistical analysis helps us to understand our data and is very important part of Machine Learning. Doing a descriptive statistical analysis of our dataset is absolutely crucial. A lot of people skip this part and therefore lose a lot of valuable insight about their data, which often leads to wrong conclusions.

Measure of Central Tendency:

It describes a whole set of data with a single value that represents the centre of its distribution. There are three main measures of central tendency:

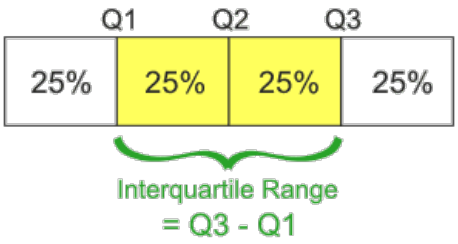
1. **Mean**: It is the sum of the observation divided by the sample size. It is not a robust statistics as it is affected by extreme values. So, very large or very low value(i.e. Outliers) can distort the answer.
2. **Median**: It is the middle value of data. It splits the data in half and also called 50th percentile. It is much less affected by the outliers and skewed data than mean. If the no. of elements in the dataset is odd, the middle most element is the median. If the no. of elements in the dataset is even, the median would be the average of two central elements.
3. **Mode**: It is the value that occurs more frequently in a dataset. Therefore a dataset has no mode, if no category is the same and also possible that a dataset has more

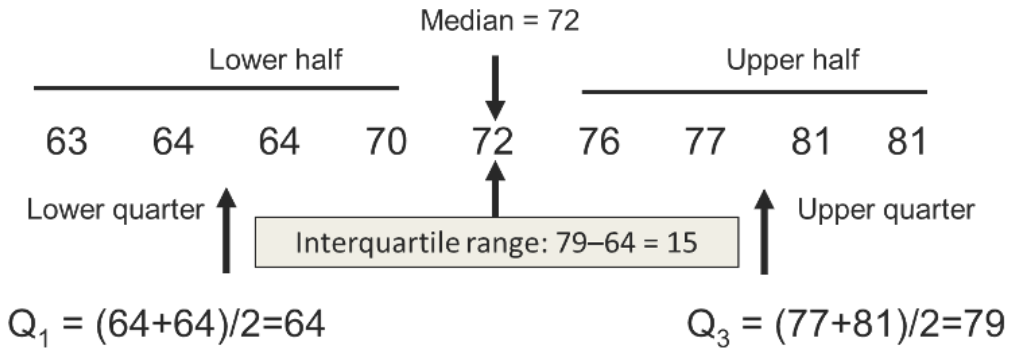
has no mode, if no category is the same and also possible that a dataset has more than one mode. It is the only measure of central tendency that can be used for categorical variables.

Measures of Variability

Measures of Variability also known as spread of the data describes how similar or varied are the set of observations. The most popular variability measures are the range, interquartile range (IQR), variance, and standard deviation.

1. **Range**: The range describes the difference between the largest and the smallest points in your data. The bigger the range the more spread out is the data.
2. **IQR**: The interquartile range (IQR) is a measure of statistical dispersion between upper (75th) quartiles i.e Q3 and lower (25th) quartiles i.e Q1. You can understand this by below example.





While the range measures where the beginning and end of your datapoint are, the interquartile range is a measure of where the majority of the values lie.

3. **Variance:** It is the average squared deviation from mean. The variance is computed by finding the difference between every data point and the mean, squaring them, summing them up and then taking the average of those numbers.

The problem with Variance is that because of the squaring, it is not in the same unit of measurement as the original data.

$$\text{variance} = \sigma^2 = \frac{\Sigma (x_r - \mu)^2}{n}$$

$$\text{standard deviation} \quad \sigma = \sqrt{\frac{\Sigma (x_r - \mu)^2}{n}}$$

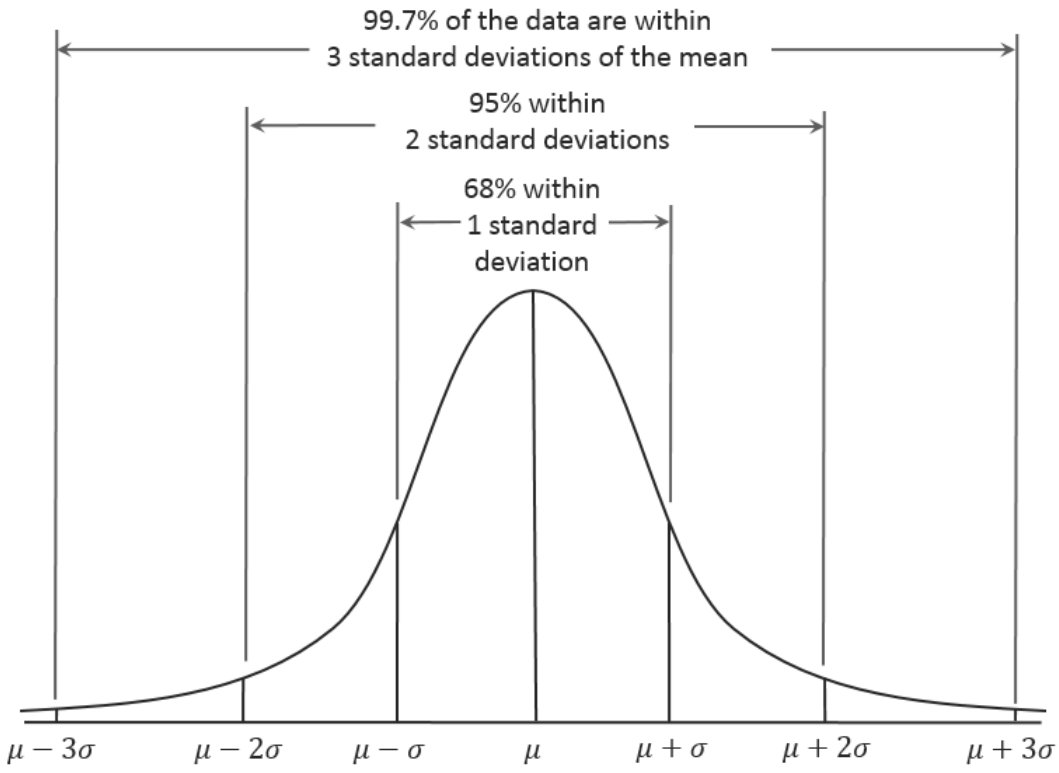
μ = mean

4. **Standard Deviation:** Standard Deviation is used more often because it is in the original unit. It is simply the square root of the variance and because of that, it is returned to the original unit of measurement.

When you have a low standard deviation, your data points tend to be close to the mean. A high standard deviation means that your data points are spread out over a wide range.

Standard deviation is best used when data is unimodal. In a normal distribution, approximately 34% of the data points are lying between the mean and one standard deviation above or below the mean. Since a normal distribution is symmetrical, 68% of the data points fall between one standard deviation above and one standard deviation below the mean. Approximately 95% fall between two standard deviations below the mean and two standard deviations above the mean. And approximately 99.7% fall between three standard deviations above and three standard deviations below the mean.

The picture below illustrates that perfectly.



With the so-called „Z-Score“, you can check how many standard deviations below (or above) the mean, a specific data point lies.

Probability

I will just give a brief indroduction of probability. Before going to the actual definition of Probability let’s look at some terminologies.

- **Experiment**: An experiment could be something like—whether it rains in Delhi on a daily basis or not.

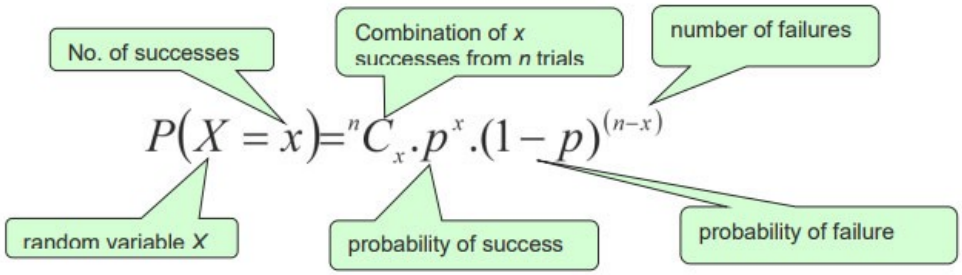
- **Outcome**: Outcome is the result of a single trial. If it rains today, the outcome of today's trial is "it rained".
- **Event**: An event is one or more outcomes of an experiment. For the experiment of whether it rains in Delhi every day the event could be "it rained" or it didn't rain.
- **Probability**: This simply the likelihood of an event. So it there's a 60% chance of it raining today, the probability of raining is 0.6.

Bernoulli Trials

An experiment which has exactly two outcomes like coin toss is called Bernoulli Trials.

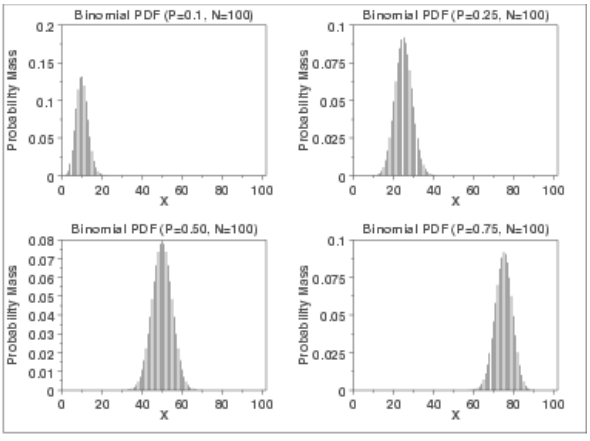
Probability distribution of the number of successes in n Bernoulli trials is known as a **Binomial distribution**.

Formula for Binomial distribution is given below.



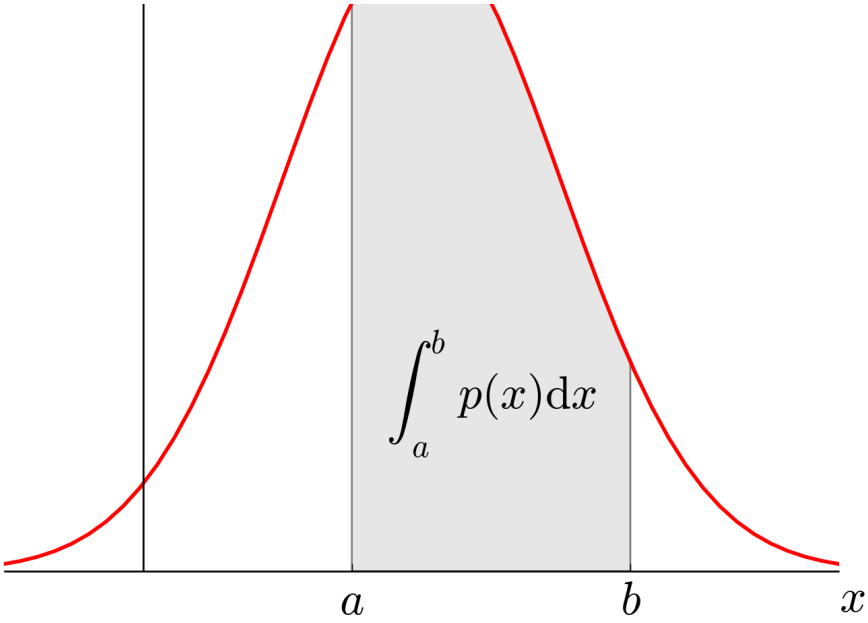
Binomial distribution Formula

Probability Mass Function for Binomial distribution with different probability of success and 100 random variables.



Probability Mass Function

Probability distribution of **continuous random variable (**variable that can assume any possible value between two points) is known as **Probability Density Function**. There will be infinite no. of trials in case of continuous random variable.



Area under a probability density function gives the probability for the random variable to be in that range.

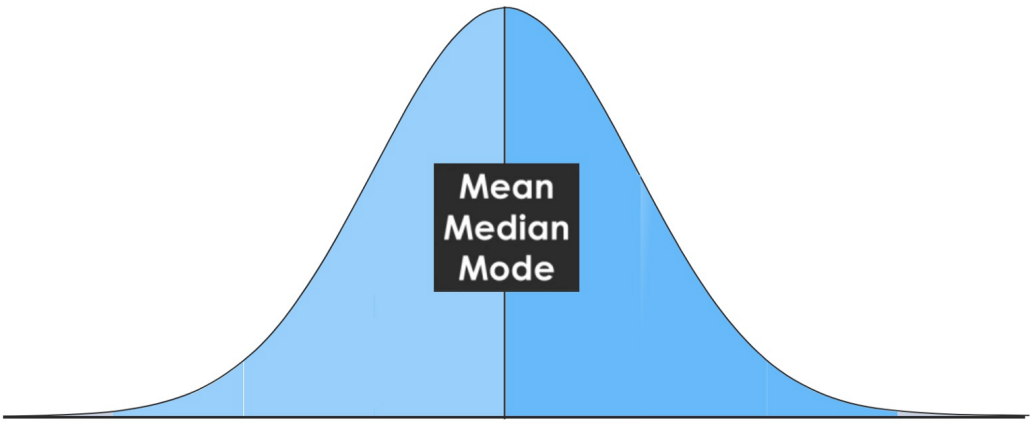
If I have a population data and I take random samples of equal size from the data, the sample means are approximately normally distributed.

Normal Distribution

It basically describes how large samples of data look like when they are plotted. It is sometimes called the “bell curve“ or the “Gaussian curve“.

Inferential statistics and the calculation of probabilities require that a normal distribution is given. This basically means, that if your data is not normally distributed, you need to be very careful what statistical tests you apply to it since they could lead to wrong conclusions.

In a perfect normal distribution, each side is an exact mirror of the other. It should look like the distribution on the picture below:



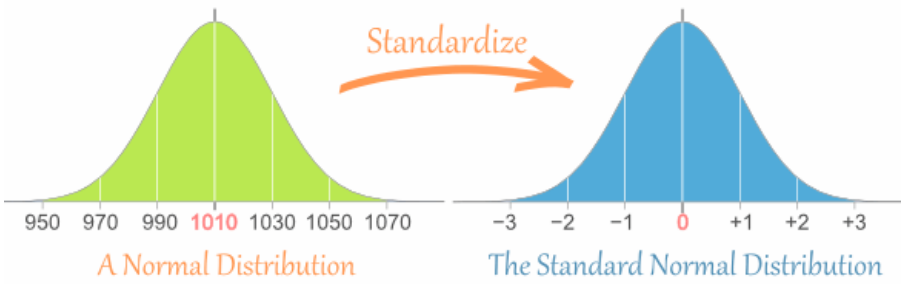
Normal Distribution

In a normal distribution, the mean, mode and median are all equal and fall at the same midline point.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{(2\sigma^2)}}$$

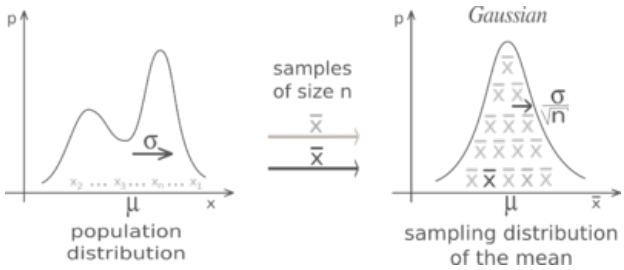
Normal Distribution Function

A normal distribution with a mean of 0 and a standard deviation of 1 is called a **standard normal distribution**. Area under the standard normal distribution curve would be 1.

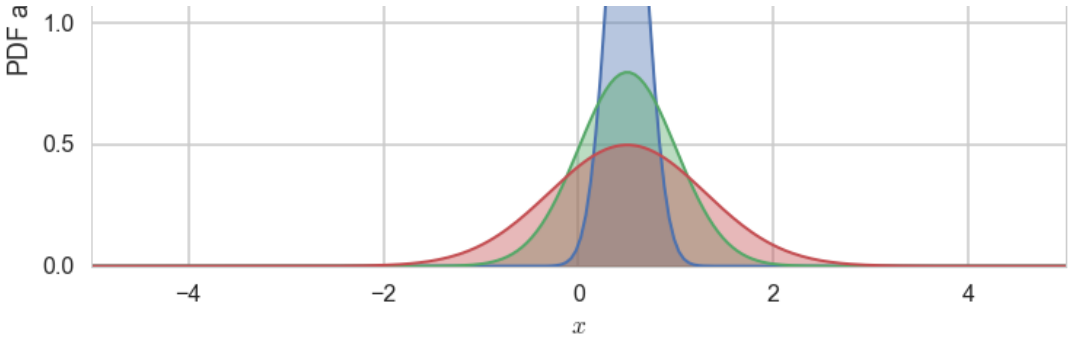


Central Limit Theorem

- If we take means of random samples from a distribution and we plot the means, the graph approaches to a normal distribution when we have taken sufficiently large number of such samples.
- The theorem also says that the mean of means will be approximately equal to the mean of sample means i.e. population mean.



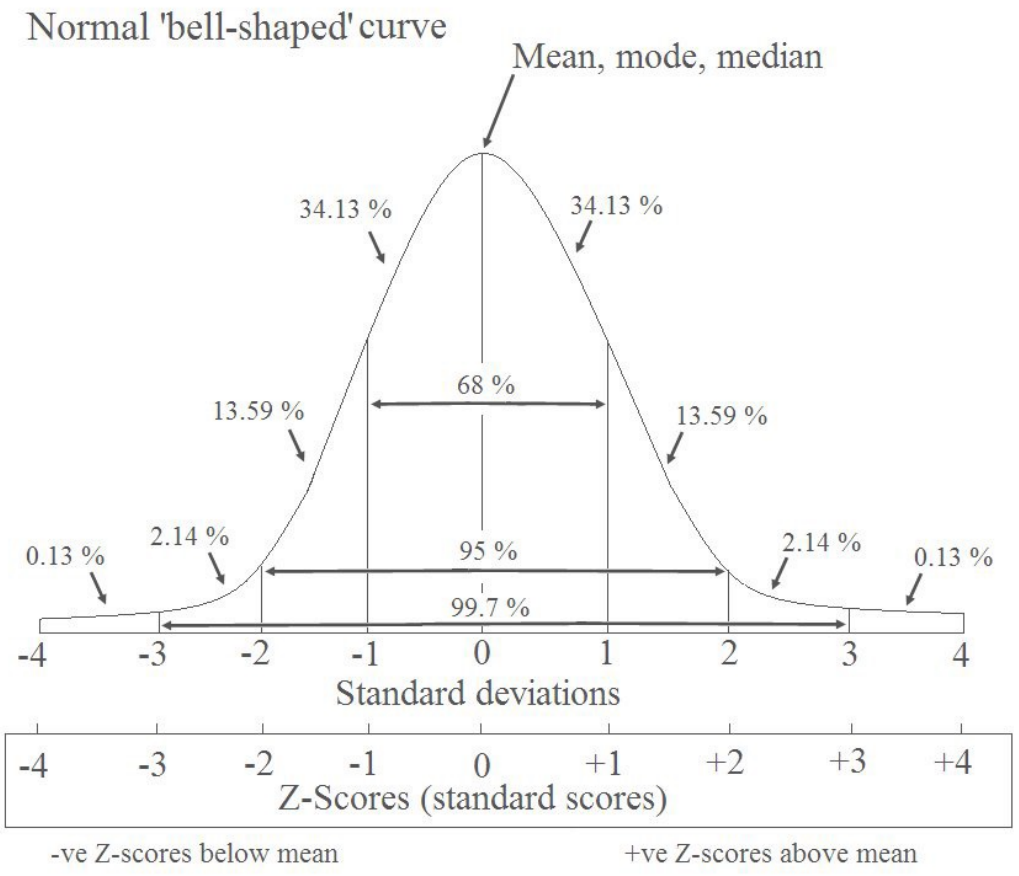
Normal distributions for higher standard deviations are flatter i.e. more spread as compared to those for lower standard deviations.



Z scores

The distance in terms of number of standard deviations, the observed value is away from the mean, is the standard score or the Z score.

A positive Z score indicates that the observed value is Z standard deviations above the mean. Negative Z score indicates that the value is below the mean.



Observed value = μ+zσ [μ is the mean and σ is the standard deviation]

From above graph area around 2 standard deviation around the mean is 0.95, that means 0.95 probability of data lying within that range.

For a particular z score, we can look into the Z-table to find the probability for values to fall less than that particular z value.

So, I hope this post gave you a proper introduction to descriptive statistics.

Data Science | Machine Learning | Statistics | Probability | Descriptive Statistics