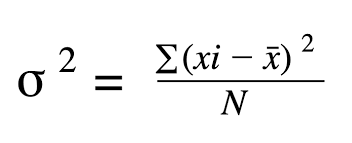
# **Measure Of Spread**

## **Variance**

How much data is spread. Average distance from each data point to the data’s mean.

1. Subtract each from its mean
2. Take square root of each
3. Sum number 2 result
4. Divide it by number of sample - 1



The higher the variance the higher the data spread.

## **Standard Deviation**

Square root of variance is standard deviation. The standard deviation is a widely used measure of how much the values in a dataset deviate from the mean.

## **Mean Absolute Deviation**

The mean absolute deviation measures the average absolute difference between each data point and the mean. Unlike the standard deviation, it uses the absolute values of the differences, which makes it less sensitive to extreme outliers.

1. Subtract sum from each
2. Calculate mean again

## **Quantile**

Split the data into some number of equal parts such as, 25%, 50%, 75%, 100%.

## **Interquartile Range (IQR)**

Is the distance between 25% and 75% quartile.

## **Outliers**

Are the data points which are substantially different from others.

Data which is ‘< Q1 – 1.5 x IQR’ or

‘> Q3 – 1.5 x IQR’ is outlier

# **Chances**

## **Samling**

How much probability of something will happen.

**P(event)=** # ways event can happen/ total number of outcomes.

**P(head)=** 1 way to get head/ 2 possible outcomes = 50%

**Two types of sampling, sampling with and without replacement.**

**Independent events vs dependent events**

If the probability of second event is not affected by the outcomes of first events is independent else dependent events.

## **Probability distribution**

It's the number of times each possible value of a variable occurs in the dataset. The number of times a value occurs in a sample is determined by its probability of occurrence.

1. Calculate probability of each group size and divide it by total sample in dataset, you will get each group size and its probability.
2. Multiplying number of sums of each group with its probability.

## **Continuous probability distribution**

A continuous distribution describes the probabilities of a continuous random variable's possible values. A continuous random variable has an infinite and uncountable set of possible values. The mapping of time can be considered as an example of the continuous probability distribution.

## **Binomial Distribution**

Probability distribution of the number of successes in a sequence of independent trails.

E.g., Number of heads in a sequence of coin flips.

**CDF (Cumulative Distribution Function):**

The Cumulative Distribution Function (CDF) gives you the cumulative probability of a random variable being less than or equal to a specific value. In other words, it provides the probability that a random variable takes on a value less than or equal to a given value. For discrete distributions, the CDF is the sum of the probabilities of all values up to and including the given value.

binom.cdf(k, n, p) calculates the probability that the random variable is less than or equal to k in a binomial distribution with parameters n (number of trials) and p (probability of success).

**rvs (Random Variates):**

Generating random variates means simulating random values from a probability distribution. In the context of scipy.stats, the rvs function generates random samples (values) that follow a specific distribution. It's a way to simulate outcomes based on the defined distribution.

binom.rvs(n, p, size) generates size random values that follow a binomial distribution with parameters n and p.

**PMF (Probability Mass Function):**

The Probability Mass Function (PMF) gives you the probability of a discrete random variable taking on a specific value. It describes the distribution of the random variable by assigning probabilities to each possible outcome.

binom.pmf(k, n, p) calculates the probability that a binomial-distributed random variable takes on the value k, given parameters n and p.