Statistics is a mathematical science including methods of collecting, organizing and analyzing data in such a way that meaningful conclusions can be drawn from them.

Examples for Statistical Way of Thinking

• The largest earthquake measured 9.2 on the Richter scale.
• Men are at least 10 times more likely than women to commit murder.
• One in every 8 Americans is COVID positive.

The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted. For example, consider some scenarios and the interpretations based upon the presented statistics.

1. A new advertisement for Amul's ice cream introduced in late May of last year resulted in a 30% increase in ice
cream sales for the following three months. Thus, the advertisement was effective.

Flaw: A major flaw is that ice cream consumption generally increases in the months of June, July, and August regardless of advertisements. This effect is called a history effect and leads people to interpret outcomes as the result of one variable when another variable (in this case, one having to do with the passage of time) is actually responsible.

2. The more liquor shops in a city, the more crime there is. Thus, liquor shops lead to crime.

Flaw: A major flaw is that both increased liquor shops and increased crime rates can be explained by larger populations. In bigger cities, there are both more liquor shops and more crime. This problem refers to the third-variable problem. Namely, a third variable can cause both situations; however, people erroneously believe that there is a causal relationship between the two primary variables rather than recognize that a third variable can cause both.

Hence, the correct Interpretation of the numbers is necessary!!!!

Types of Statistics:

1. **Descriptive statistics - Provides exact and accurate information.**
Deals with the processing of data without attempting to draw any
inferences from it. The characteristics of the data are described in simple terms. Events that are dealt with include everyday happenings such as accidents, prices of goods, business, incomes, epidemics, sports data, population data.

2. **Inferential statistics - Provides information of a sample and we need inferential statistics to reach to a conclusion about the population.**

A statistical method that deduces from a small but representative sample the characteristics of a bigger population. In other words, it allows the researcher to make assumptions about a wider group, using a smaller portion of that group as a guideline.

Basic Terminology in Statistics

Population - Refers to the total amount of things.
Sample - Small part of population that is used for study.
Sample Size - Total amount of things in a sample. (In case of several samples, average of all the sample sizes)
Variable - What we are studying. (Quantitative, Categorical)

Quantitative Variables - Data that is measured in numbers. It deals with numbers that make sense to perform arithmetic calculations with.

1. Discrete - Refer to variables that can be measured only in whole numbers
   Eg: Number of pets, Number of students
2. Continuous - Refers to variables that can take any numerical value
   Eg: Weight, Height,

Categorical variables - Refer to the values that place “things” into different groups or categories.

1. Ordinal - Logical ordering to the values of a categorical variable
   Eg: Workplace status: Entry Analyst, Analyst I, Analyst II, Lead Analyst
2. Nominal - No Logical Ordering to the values of a categorical variable.
   Eg: Blood type: O-, O+, A-, A+, B-, B+, AB-, AB+

Descriptive statistics

1. **Measures of Central Tendency - Mean, Median and Mode**

   a. **Mean** - Average value of the set of Numbers. Mean is a a number around which a whole data is spread out. Denoted by μ for population mean and for sample mean.

      Example: Find the mean of 5,5,2,6,3,8,9?
      A: Mean is (5+5+2+6+3+8+9) / 7 = 38/7 = 5.43

   b. **Median -** Median is the value which divides the data in 2 equal parts i.e. number of terms on right side of it is same as number of terms on left side of it when data is arranged in either ascending or descending order.

      Example: Find the Median of 5,5,2,6,3,8,9?

A: Putting it in ascending order = 2,3,5,5,6,8,9. Hence, Median = Mid Number = 5.
(Note: Median of a even set of numbers can be found by taking the average of the 2 middle numbers.
E.g. Median of 2,3,4,7 = average of (3 and 4) = 3.5)

c. **Mode** - Mode is the term appearing maximum time in data set i.e. term that has highest frequency.

Example: Find the Median of 5,5,2,6,3,8,9?
A: Mode = Maximum number of repetition in dataset = 5. Hence, Mode = 5.
(Note: If there is no repetition of data then mode is not present. E.g.: What is the mode of 1,2,3,5,6?
A: None i.e. mode is not present.)

1. What is Minimum and Maximum value?
It is the minimum and Maximum values of the dataset respectively.

2. What is 1st and 3rd Quartile? – Also called the lower and upper quartile respectively.

When we divide the dataset into two groups while calculating the median (sorted in ascending order), then the median of the first half is 1st Quartile and the median of the second half is 3rd Quartile.

3. Then where is the 2nd Quartile?
Your median is the 2nd Quartile.

Q. Given is the ages of people registered for a webinar, calculate the 5 point summary (5 number summary) of the ages of the participants?

19, 26, 25, 37, 32, 28, 22, 23, 29, 34, 39

When to Use the Mean, Median, and Mode?
Mean - It is best to use the mean when the distribution of the data is fairly symmetrical and there are no outliers.
Median - It is best to use the median when the distribution of the data is either skewed or there are outliers present.
Mode - It is best to use the mode when you are working with categorical data and you want to know which category occurs most frequently.

2. Measure of Spread / Dispersion -

a. **Standard Deviation -**

Standard deviation is the measurement of average distance between each quantity and mean. That is, how data is spread out from mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values. A useful property of standard deviation is that , unlike variance , it is expressed in the same units as the data.

Standard Deviation Formula	
Sample	**Population**
$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n-1}}$	$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$
X – The Value in the data distribution $\bar{x}$ – The Sample Mean n - Total Number of Observations	X – The Value in the data distribution μ – The population Mean N – Total Number of Observations

b. **Variance -**

Variance is a square of average distance between each quantity and mean. That is, it is a square of standard deviation.

Population	Sample
$\sigma^2 = \frac{\Sigma(x_i-\mu)^2}{n}$	$S^2 = \frac{\Sigma(x_i-\bar{x})^2}{n-1}$
μ - Population Average xi - Individual Population Value n - Total Number of Population σ^2 - Variance of Population	x - Sample Average x_i - Individual Population Value n - Total Number of Sample S^2 - Variance of Sample

What is Bessel's Correction?

Bessels' correction refers to the "n-1" found in several formulas, including the sample variance and sample standard deviation formulas. This correction is made to correct for the fact that these sample statistics tend to underestimate the actual parameters found in the population.

As we can see, SAMPLE formulas have n-1 in the denominator, where n is the sample size. So why do we subtract 1 when using these formulas?

The simple answer: the calculations for both the sample standard deviation and the sample variance both contain a little bias (that's the statistics way of saying "error"). Bessel's correction (i.e. subtracting 1 from your sample size) corrects this bias. In other words, you'll usually get a more accurate answer if you use n-1 instead of n.

c. **Range**

Range is one of the simplest techniques of descriptive statistics. It is the difference between lowest and highest value.

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

d. **IQR (InterQuantile Range)**

In statistics and probability, quartiles are values that divide your data into quarters provided data is sorted in an ascending order.

$$IQR = Q3 - Q1$$

Steps to find out the IQR

1. Order the data from least to greatest
2. Find the median
3. The left side of the median is the lower half and the right side of the data is the upper half.
4. Calculate the median of both the lower and upper half of the data (Called Q1 and Q3 respectively)
5. The IQR is the difference between the upper and lower medians

Transforming Data

1. Below are the weights of 5 persons. Calculate Mean, Standard Deviation : 105, 156, 145, 172, 100 2. Suppose each one of them gained extra 5 Kg. weight during winters. Can you calculate the new Mean and Standard deviation?

Mean = (105 + 156 + 145 + 172 + 100) / 5 = 678 / 5 = 135.6 kg

Standard deviation:
Step 1: Calculate the squared differences from the mean for each weight.
Step 2: Find the average of those squared differences.
Step 3: Take the square root of that average.
Standard Deviation: 31.75

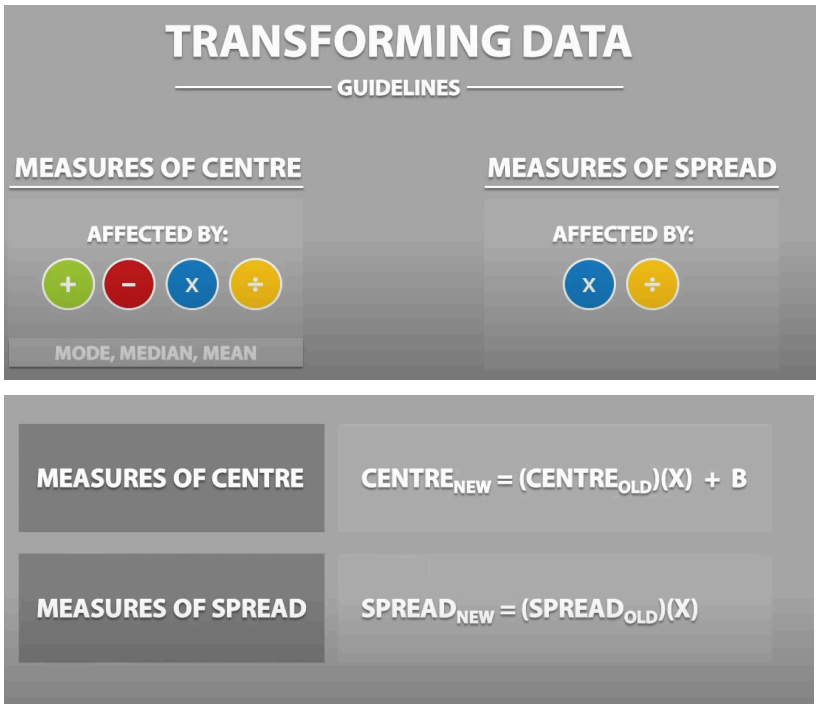
New mean: Mean = 135.6 + 5 = 140.6 kg
New standard deviation: The standard deviation remains the same. Adding a constant value to each data point does not affect the standard deviation, only the mean shifts.

Example 2: Considering the same set of people from the previous example, Suppose that these persons are advised to drink 2.5 ml of water for every Kg they weigh plus 750 ml of water everyday. 105, 156, 145, 172, 100 What is the mean and STD for the amount of water consumed everyday?

To find the mean and standard deviation for the amount of water consumed every day by each person, we'll first calculate the amount of water consumed by each individual according to the given formula and then find the mean and standard deviation.

Given: Weight of persons: 105, 156, 145, 172, 100 (in kg)

Formula for water consumption: 2.5 ml/kg + 750 ml 2.5ml/kg+750ml



X -> is the Multiplicative term ie 2,5
B -> is the additive term ie 750

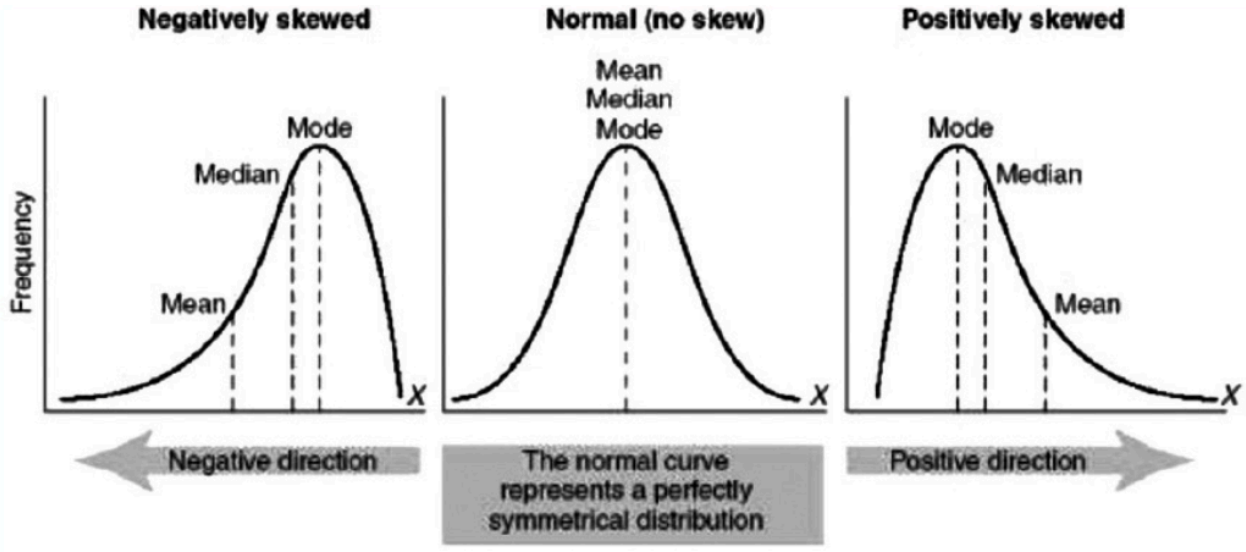
Mean = 135.6*2.5 + 250 = 140.6 kg = 1089
Standard Deviation = 31.75 x 2.5 = 79.38

3. Measure of symmetricity – Skewness and Kurtosis

a. **Skewness** is usually described as a measure of a dataset's symmetry – or lack of symmetry. A perfectly symmetrical data set will have a skewness of 0. The normal distribution has a skewness of 0. Example for negative skew - Age of Dea

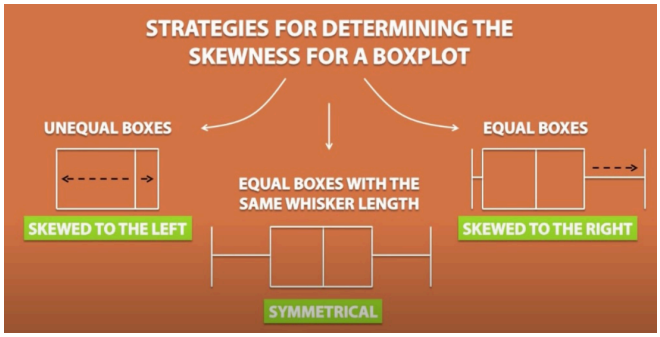
So, when is the skewness too much? The rule of thumb seems to be:

1. If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
2. If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed.
3. If the skewness is less than -1 or greater than 1, the data are highly skewed.



Formula:

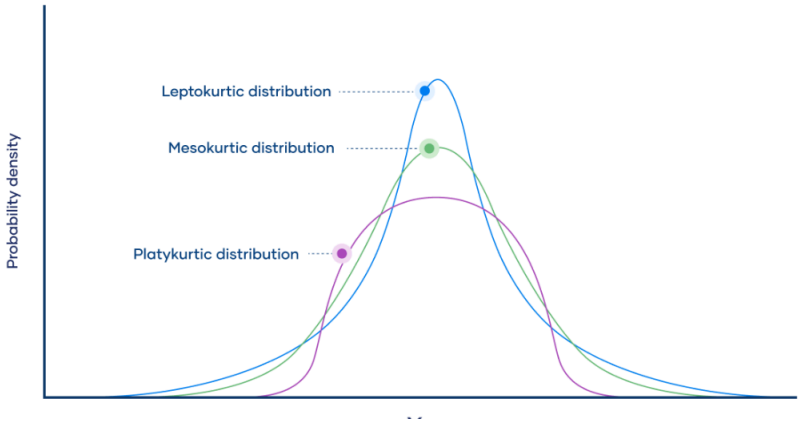
$$a_3 = \sum \frac{(X_i - \bar{X})^3}{ns^3}$$



b. **Kurtosis -** Kurtosis is all about the tails of the distribution – not the peakness or flatness. It measures the tail-heaviness of the distribution.
Outlier Detection : Large Kurtosis suggests there could be outliers in the data.

Formula:

$$a_4 = \sum \frac{(X_i - \bar{X})^4}{ns^4}$$



	Category		
	Mesokurtic	**Platykurtic**	**Leptokurtic**
Tailedness	Medium-tailed	Thin-tailed	Fat-tailed
Outlier frequency	Medium	Low	High
Kurtosis	Moderate (3)	Low (< 3)	High (> 3)
Excess kurtosis	0	Negative	Positive
Example distribution	Normal	Uniform	Laplace