Learn Statistics

MathopediaForAI Tutorial 4.2

This is a tutorial that develops a firm foundation of statistics by covering topics such as central tendency, expected values, gaussians, skewness, kurtosis, and joint and conditional probability.

1. Fundamentals of statistics

There are five key terms which form the basis of statistical analysis - population, sample, parameter, statistic and variable. In this section, we'll start by defining each of these and then proceed to the introduction to the different branches of statistics.

- **Population:** the group of people for which the conclusion needs to be drawn (for example, all women in the United States who shop at Target stores)
- **Sample:** a subset of the population which is being surveyed (for example, randomly chosen 100 women from each state in the USA who shop at Target stores)
- **Parameter:** a characteristic of the population which makes it unique (for example, the % of women who visit Target only on weekends)
- **Statistic and variable:** variables are specific characteristics related to the behavior being observed (for example, the number of women who bought at least one vest over a period of one month)

Now that we understand the basic terminology, let's look at some of the commonly used formulae in statistics. Note that we will be discussing these formulas in the later sections - for now let's just list them out.

$$\text{Mean} = \frac{\sum \text{All observations}}{\text{Number of observations}}$$

$$\text{Median for even values of } n = \frac{\left(\frac{n}{2}\right)^{th} \text{ term } \left(\frac{n}{2}+1\right)^{th} \text{ term}}{2}$$

$$\text{Median for odd values of } n = \left(\frac{n+1}{2}\right)^{th} \text{ term}$$

$$\text{Mode} = L + h\frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)}$$

$$\text{Standard deviation} = \sqrt{\frac{\sum (x - x_{average})^2}{n-1}}$$

2. Measures of central tendency

Central tendency is used to observe the overall distribution of the data and determine a specific value, or a small range of values, which can represent the entire dataset. The three most common methods of obtaining the Central tendencies are mean, median, and mode.

Mean:

There are three main types of mean with the most common one being arithmetic mean. The other two types are Geometric mean and Harmonic mean. You may read about GM and HM in more depth in the algebra course with reference to inequalities. In this section, we'll be mainly discussing arithmetic mean.

Calculating the arithmetic mean for ungrouped data is the same as the simple mean. For grouped data, the formula becomes:

$$\text{Arithmetic mean} = \frac{\sum_{i=1}^{n} X_{ai} \cdot X_{bi}}{n}$$

Median:

The formulas for median with ungrouped data were mentioned in the previous section. For grouped datasets, the formula becomes:

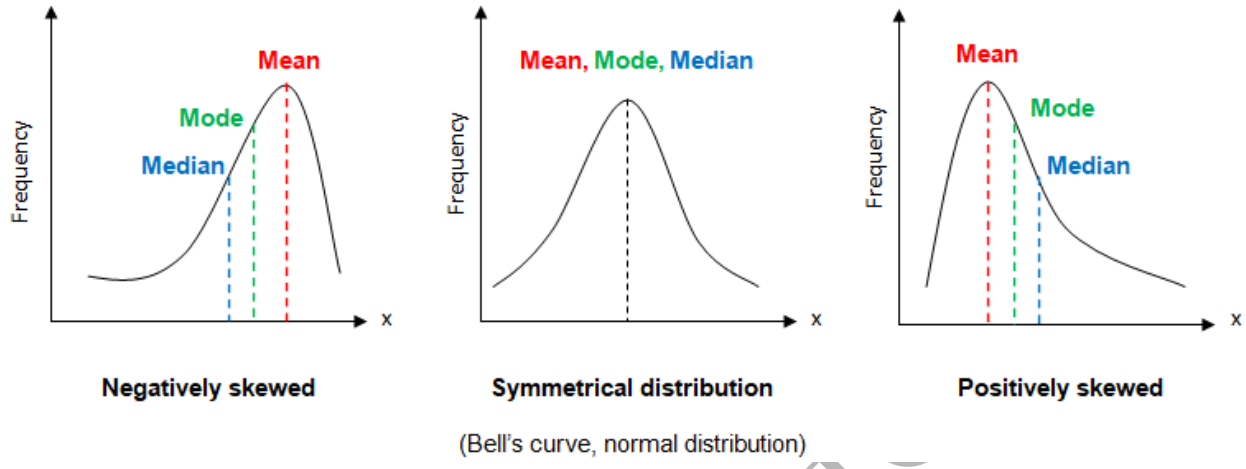
$$1 + \frac{N/2 - c_r}{f} \times h$$

Mode:

For ungrouped data, finding the mode is as simple as identifying which value is repeated the maximum number of times within the dataset. For grouped data, the mode can be calculated with the formula below.

$$\text{Mode for grouped data} = 1 + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right] h$$

But this raises an important question: when and why do we use each of these tendency identifiers? Have a look at the graph below.



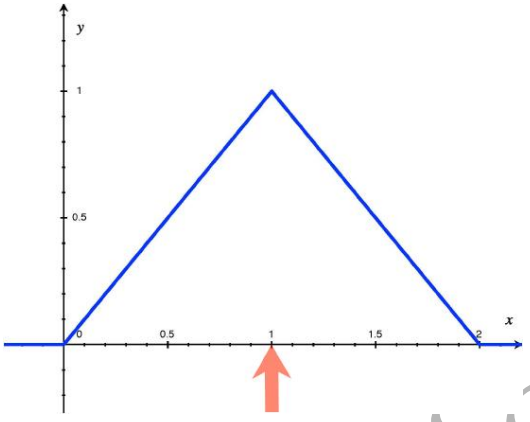
3. Expected value of functions

In general, the expected value is defined as the most probable average of a particular variable given a set of constraints in the long run. In most cases, the expected value is used to make predictions about behavioral changes or results to particular changes. A simple example could be defining a list of teams for a tournament and predicting which team is expected to win. This value is classically denoted as $E[X]$ and it can be calculated using the formula below.

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

Expected values are most commonly used in multivariate models or scenario based analysis in order to determine the optimal conditions required to achieve a certain outcome.

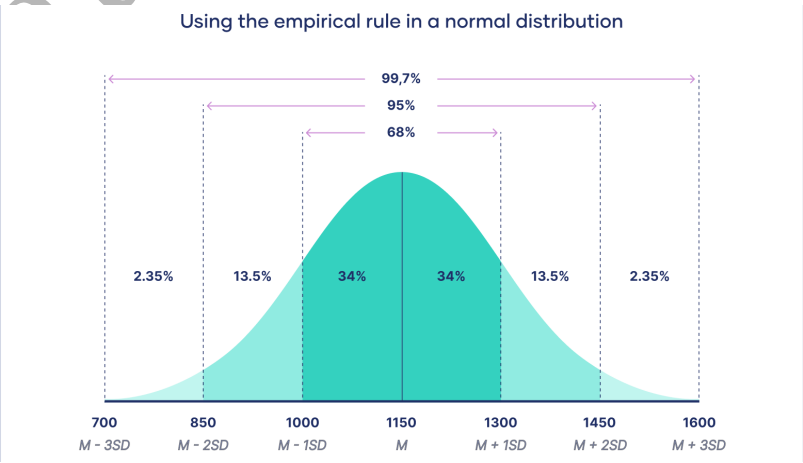
Note that the expected value of a function is typically its average value over the long run. The



As an example, in the diagram above, the expected value would lie at the peak of the graph (pointed by the red arrow). Remember that the expected does not have to equal either the median or the mode in cases when the obtained graphs are skewed.

4. Gaussians

Gaussian distribution (also known as normal distribution) is characteristically represented with a bell shaped curve as shown below.

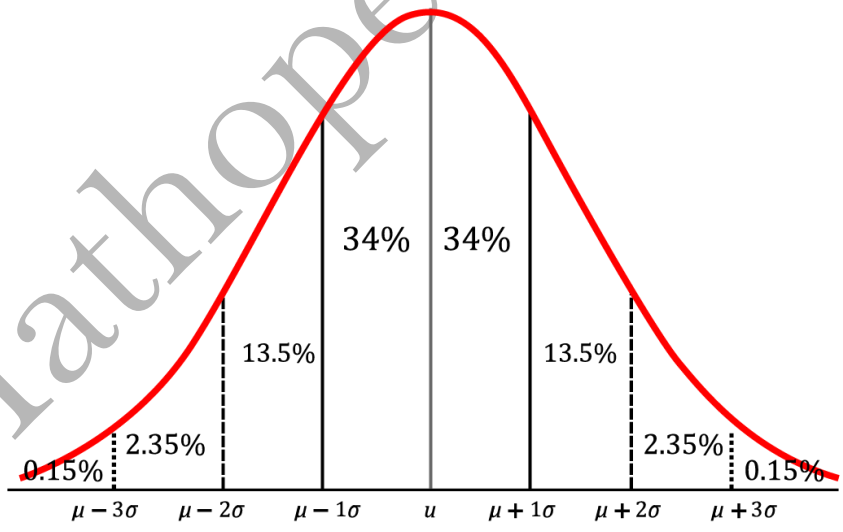


The general formula used to represent the Gaussian distribution function is given as:

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

Properties of the Gaussian distribution:

- Zero skew, kurtosis of 3
- Symmetrical distribution in the form of the bell curve
- Mean = Mode = Median
- Standard distribution = 1
- Follows the empirical rule which states that 68.2% of observations appear in ± 1 standard deviation, 95.4% appear in ± 2 standard deviation 99.7% occur in ± 3 standard deviation.



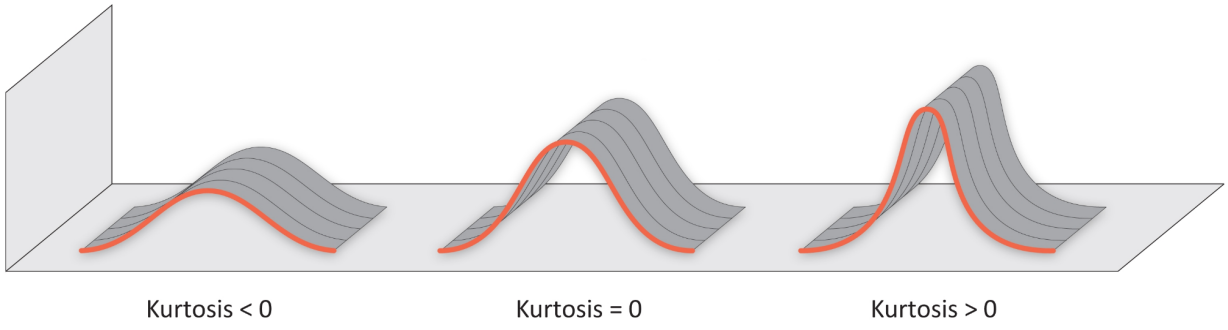
Graphic representation of the empirical rule

5. Distribution (Skewness and Kurtosis)

We've used these terms quite a number of times before - skewness and kurtosis. But what exactly do they represent and how are they useful?

Skewness: To put simply, skewness measures how symmetric the data of frequency is throughout the graph. So, if a graph is higher towards the left as compared to the right, we say that it is negatively skewed. Similarly, if it is pushed more towards the right then the data is positively skewed. Recall that the data in a Gaussian distribution is peaked exactly at the center - meaning that the data is not skewed at all.

Kurtosis: This is a measure of the elongation of data curves. Typically, the range of acceptable kurtosis lies between $$-10$$ to $$+10$$. The diagram below shows how different kurtosis values can be seen graphically.

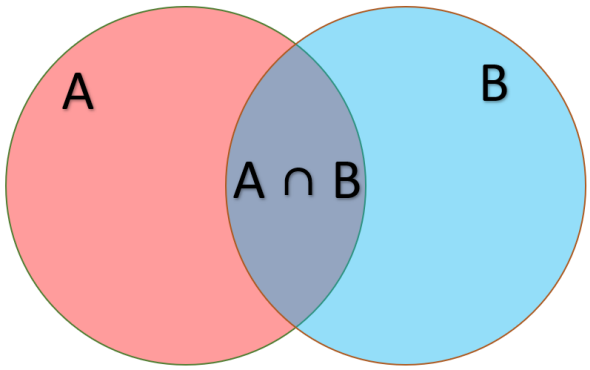


Note that a negative kurtosis stands for lack of elongation.

You can also think of skewness as the horizontal pull or stretch and kurtosis as the vertical stretch (so when the kurtosis is negative, it implies that the graph has been compressed)

6. Joint and conditional probability

A joint probability distribution represents a probability distribution for two or more random variables.



Classically, the joint probability is given as $$P(A) \times P(B)$$ as it considers the likeliness of event 1 occurring and then considers the probability of B occurring when A is assumed to have occurred.

The concept of joint probability is very often confused with that of conditional probability. Joint probability simply determines the probability of multiple events happening whereas conditional probability relies upon an interdependence between individual events.

For example, the probability of a person buying avocados AND visiting a missing school would be given as a joint probability as these events must occur together within a single timeframe. However, the probability of a person wearing black vest and black pants would be given by conditional probability because if one wears a black vest then they are more likely to wear black pants (i.e. the color of the second garment is somehow dependent on the first one).