SKEWNESS AND NORMAL DISTRIBUTION

12 October 2023 21:22

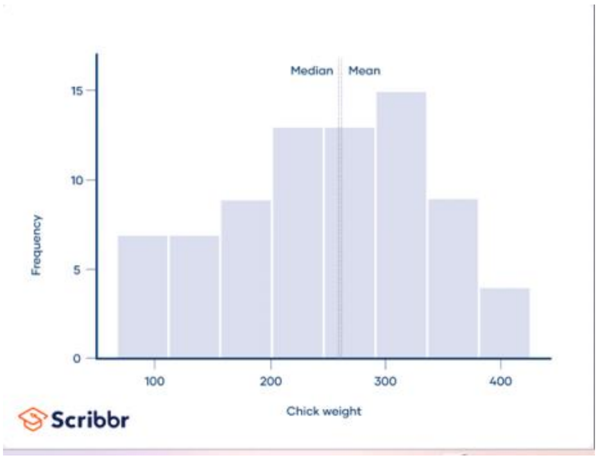
Skewness | Definition, Examples:

Skewness is a measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images.
A distribution can have right (or positive), left (or negative), or zero skewness.

A right-skewed distribution is longer on the right side of its peak, and a left-skewed distribution is longer on the left side of its peak:

What is zero skew?

When a distribution has zero skew, it is symmetrical. Its left and right sides are mirror images.
Normal distributions have zero skew, but they're not the only distributions with zero skew. Any symmetrical distribution, such as a uniform distribution or some bimodal (two-peak) distributions, will also have zero skew.
The easiest way to check if a variable has a skewed distribution is to plot it in a histogram. For example, the weights of six-week-old chicks are shown in the histogram below.
The distribution is approximately symmetrical, with the observations distributed similarly on the left and right sides of its peak. Therefore, the distribution has approximately zero skew.



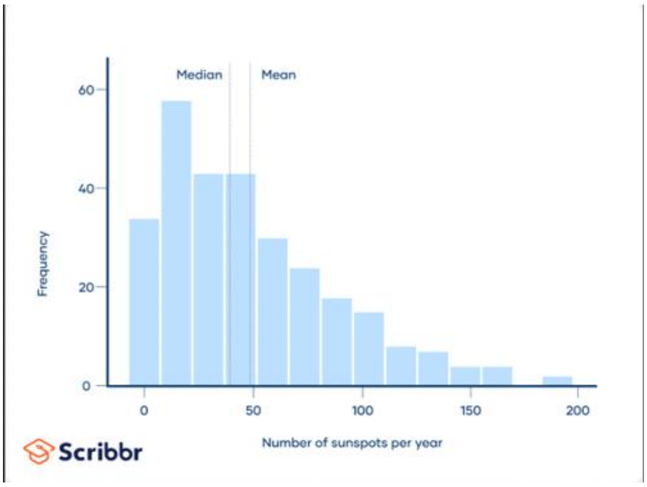
In a distribution with zero skew, the mean and median are equal.

Zero skew: mean = median

For example, the mean chick weight is 261.3 g, and the median is 258 g. The mean and median are almost equal. They aren't perfectly equal because the sample distribution has a very small skew. Real Data have at least bit of skew.

What is right skew (positive skew)?

A right-skewed distribution is longer on the right side of its peak than on its left. Right skew is also referred to as positive skew. You can think of skewness in terms of tails. A tail is a long, tapering end of a distribution. It indicates that there are observations at one of the extreme ends of the distribution, but that they're relatively infrequent. A right-skewed distribution has a long tail on its right side.
The number of sunspots observed per year, shown in the histogram below, is an example of a right-skewed distribution. The sunspots, which are dark, cooler areas on the surface of the sun, were observed by astronomers between 1749 and 1983.
The distribution is right-skewed because it's longer on the right side of its peak. There is a long tail on the right, meaning that every few decades there is a year when the number of sunspots observed is a lot higher than average.



The mean of a right-skewed distribution is almost always greater than its median. That's because extreme values (the values in the tail) affect the mean more than the median.

Right skew: mean > median

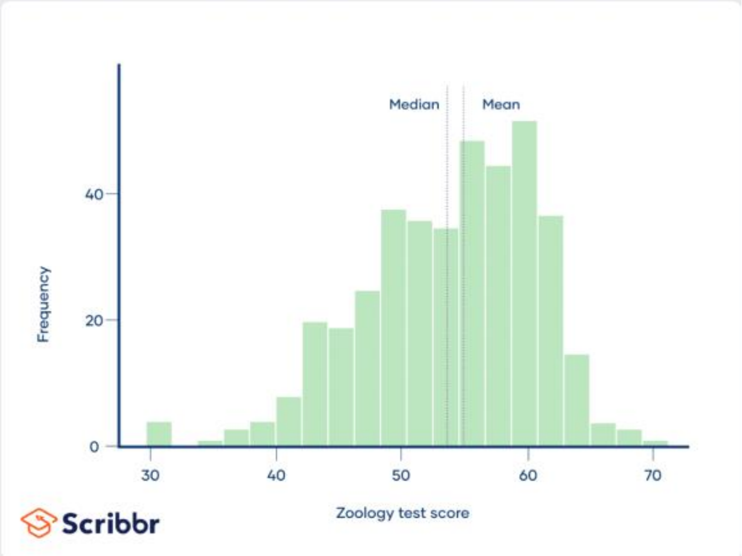
For example, the mean number of sunspots observed per year was 48.6, which is greater than the median of 39.

What is left skew (negative skew)?

A left-skewed distribution is longer on the left side of its peak than on its right. In other words, a left-skewed distribution has a long tail on its left side. Left skew is also referred to as negative skew.

Test scores often follow a left-skewed distribution, with most students performing relatively well and a few students performing far below average. The histogram below shows scores for the zoology portion of a standardized test taken by Indian students at the end of high school.

The distribution is left-skewed because it's longer on the left side of its peak. The long tail on its left represents the small proportion of students who received very low scores.



The mean of a left-skewed distribution is almost always less than its median.

Left skew: mean < median

For example, the mean zoology test score was 53.7, which is less than the median of 55.

How to calculate skewness

There are several formulas to measure skewness. One of the simplest is Pearson's median skewness. It takes advantage of the fact that the mean and median are unequal in a skewed distribution.

Pearson's median skewness =

$$3 \times \frac{(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

Pearson's median skewness tells you how many standard deviations separate the mean and median.

Real observations rarely have a Pearson's median skewness of exactly 0. If your data has a value close to 0, you can consider it to have zero skew. There's no standard convention for what counts as "close enough" to 0 (although this research suggests that 0.4 and −0.4 are reasonable cutoffs for large samples).

]

Pearson's median skewness of the number of sunspots observed per year:

- Mean = 48.6
- Median = 39
- Standard deviation = 39.5

Calculation

Pearson's median skewness =

$$3 \times \frac{(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

Pearson's median skewness =

$3 \times \frac{(48.6 - 39)}{30.5}$

Pearson’s median skewness =

0.73

NORMAL DISTRIBUTION:

In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.
Normal distributions are also called Gaussian distributions or bell curves because of their shape.

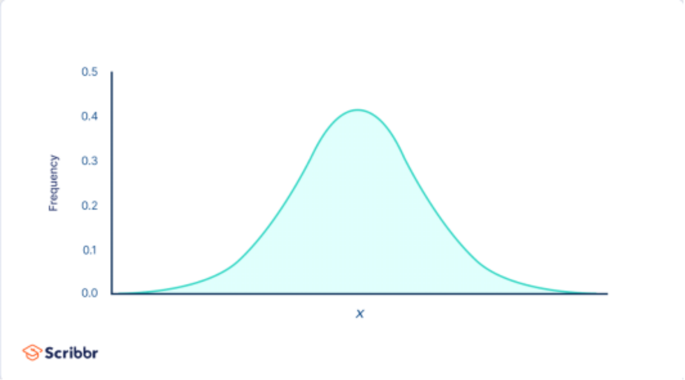
Why do normal distributions matter?

All kinds of variables in natural and social sciences are normally or approximately normally distributed. Height, birth weight, reading ability, job satisfaction, or SAT scores are just a few examples of such variables.
Because normally distributed variables are so common, many statistical tests are designed for normally distributed populations.
Understanding the properties of normal distributions means you can use inferential statistics to compare different groups and make estimates about populations using samples.

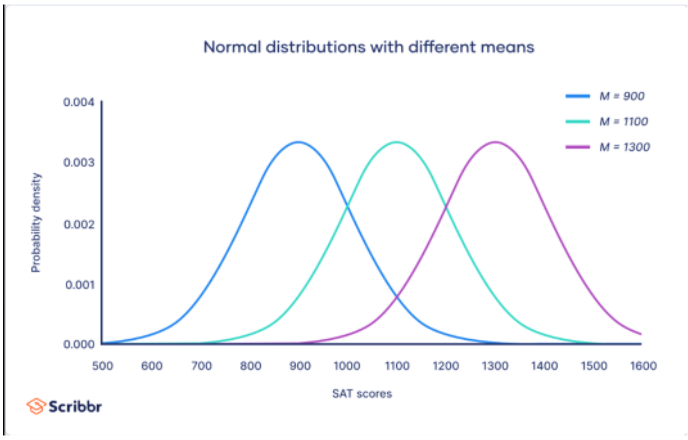
What are the properties of normal distributions?

Normal distributions have key characteristics that are easy to spot in graphs:

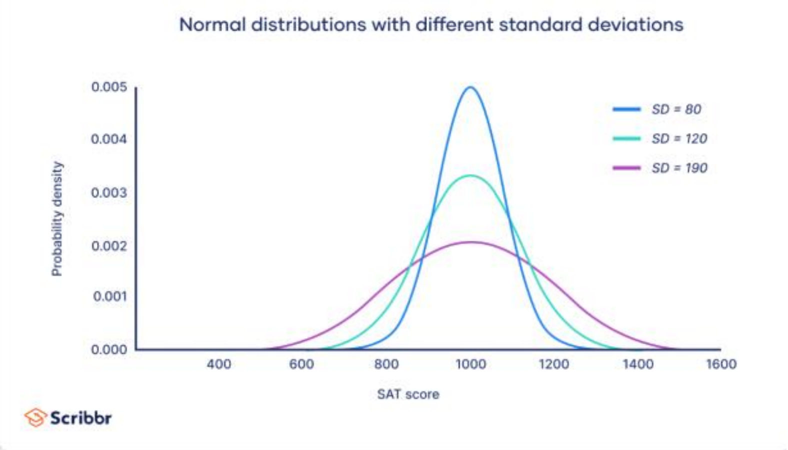
- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.



The mean is the location parameter while the standard deviation is the scale parameter.
The mean determines where the peak of the curve is centered. Increasing the mean moves the curve right, while decreasing it moves the curve left.



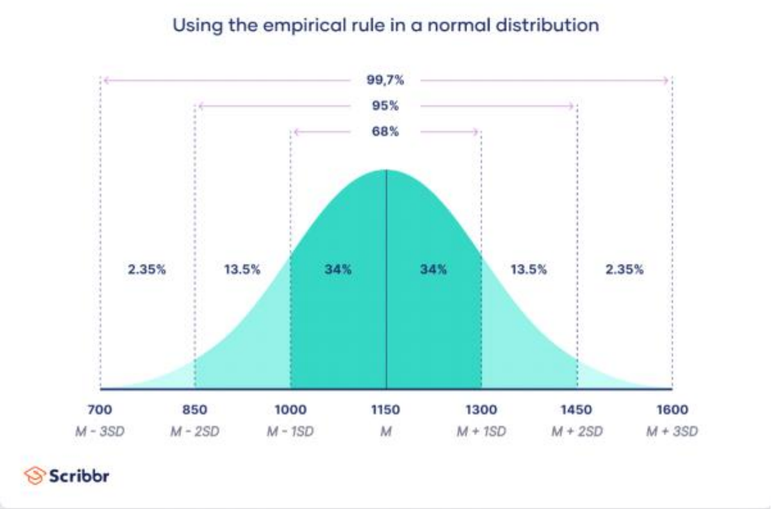
The standard deviation stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.



Empirical rule

The **empirical rule**, or the 68-95-99.7 rule, tells you where most of your values lie in a normal distribution:

- Around 68% of values are within 1 standard deviation from the mean.
- Around 95% of values are within 2 standard deviations from the mean.
- Around 99.7% of values are within 3 standard deviations from the mean.



The empirical rule is a quick way to get an overview of your data and check for any outliers or extreme values that don't follow this pattern.

If data from small samples do not closely follow this pattern, then other distributions like the t-distribution may be more appropriate. Once you identify the distribution of your variable, you can apply appropriate statistical tests.

Central limit theorem

The Central Limit Theorem (CLT) is a fundamental concept in statistics. In simple words, it states that if you take a large enough sample from any population, and you calculate the average (or mean) of those samples, the distribution of those averages will be approximately normal, regardless of the shape of the original population's distribution.

Here's a breakdown of the central ideas:

1. Large Enough Sample: You need a reasonably large sample size for the CLT to work effectively. The rule of thumb is often that the sample size should be at least 30. However, the larger the sample, the better the approximation to a normal distribution.

2. Averages Become Normal: When you calculate the average of the values in these samples, you'll notice that as you take more and more samples and calculate their averages, the distribution of these averages will tend to look like a bell-shaped, normal distribution. This is true even if the original population's distribution is not normal.

3. Useful for Inference: The CLT is incredibly useful because it allows statisticians to make inferences about a population, even when they don't know the population's distribution. You can use the properties of the normal distribution to make estimates and perform hypothesis tests.

   In practical terms, the CLT is often used in situations where you have data from a large, diverse population, and you want to make general statements about that population.

The central limit theorem shows the following:

- Law of Large Numbers: As you increase sample size (or the number of samples), then the sample mean will approach the population mean.
- With multiple large samples, the sampling distribution of the mean is normally distributed, even if your original variable is not normally distributed.

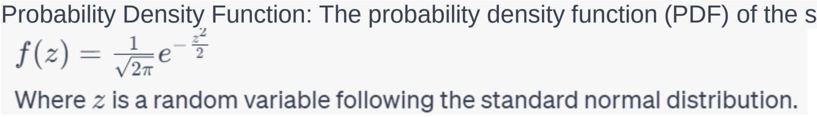
What is the standard normal distribution?

A standard normal distribution, also known as the Z-distribution or the standard Gaussian distribution, is a specific type of normal distribution with a mean (average) of 0 and a standard deviation of 1. In statistics, the normal distribution is often called the "bell curve" because of its characteristic shape, which is symmetrical and resembles a bell.
The standard normal distribution is used as a reference or baseline for many statistical analyses. Here are some key characteristics and uses:

Characteristics:

- Mean (μ) = 0: The center of the distribution is at zero.
- Standard Deviation (σ) = 1: The spread or variability of the data is standardized to one unit.

Probability Density Function: The probability density function (PDF) of the standard normal distribution is given by the formula:



$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Where z is a random variable following the standard normal distribution.

Standard Scores (Z-Scores): The Z-score for a data point in a normal distribution represents how many standard deviations it is away from the mean. It's calculated as: �=�−��$Z=\sigma X-\mu$Here, �$X$is the data point, �$\mu$is the mean, and �$\sigma$is the standard deviation. A Z-score allows you to compare and interpret data points in terms of their position relative to the mean.

Use in Statistical Testing: The standard normal distribution is fundamental in hypothesis testing and confidence intervals. Many statistical tests, such as the Z-test, T-test, and chi-squared test, rely on the properties of the standard normal distribution for making inferences about populations.

Z-Table: To find probabilities associated with Z-scores, statisticians use a Z-table (or a calculator or software). The Z-table provides the probability that a Z-score falls below a certain value, allowing you to determine the likelihood of observing a particular value or range of values in a standard normal distribution.

Data Standardization: In data analysis, you can transform data from any normal distribution to the standard normal distribution by using the Z-score formula. This transformation allows you to compare and analyse data from different sources or with different scales.