REVISTA CIE

Revista de Investigación Educativa | Vol. 2, No. 2
ISSN: 3061-8657 | Julio-Diciembre 2026

STATISTICS

Normal Distribution

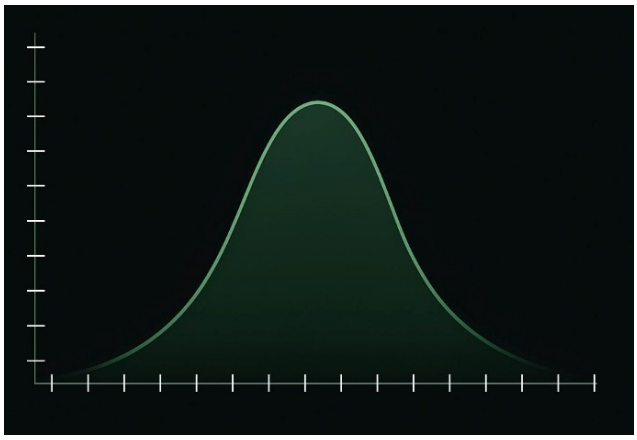
Dr. Gerardo García Maldonado
Full-Time Professor and Researcher
National System of Researchers and Innovators (SNII), Level 1
Autonomous University of Tamaulipas
Faculty of Medicine of Tampico "Dr. Alberto Romo Caballero"
gmaldonado@docentes.uat.edu.mx

Dr. Raúl de León Escobedo
Full-Time Professor and Researcher
SNII Candidate
Autonomous University of Tamaulipas
Faculty of Medicine of Tampico "Dr. Alberto Romo Caballero"
raul.deleon@docentes.uat.edu.mx

Dr. José Eugenio Guerra Cárdenas
Full-Time Professor and Researcher
Autonomous University of Tamaulipas
Faculty of Medicine of Tampico "Dr. Alberto Romo Caballero"
jguerra@docentes.uat.edu.mx

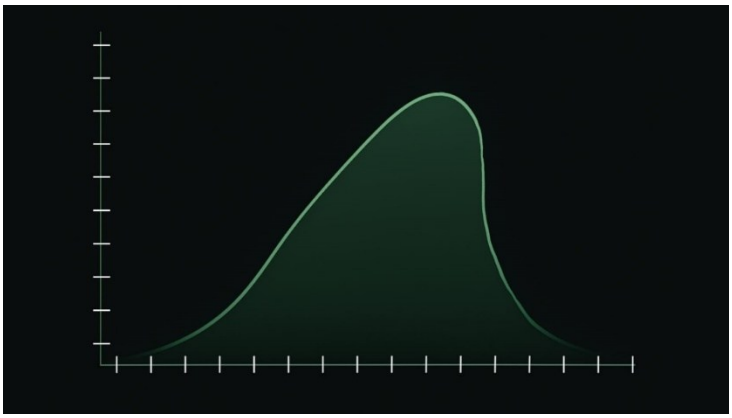
INTRODUCTION

The normal distribution is a widely used theoretical model that describes how the numerical values of a variable are distributed within a dataset. It is commonly referred to as the bell-shaped curve because the graph of its probability density function resembles the shape of a bell.



Theoretically, it is assumed that data derived from any random variable, which by definition are always numerical and measured on either an interval or ratio scale, can generally be approximated by a normal distribution. However, this assumption is not always met in practice. Therefore, data transformation procedures and, in some cases, modifications to the measurement scale are recommended to achieve an approximately normal distribution, particularly because this assumption is required for the application of parametric statistical methods.

For example, if serum creatinine levels were measured in a group of individuals between 40 and 50 years of age, the resulting data distribution might be skewed, as creatinine values tend to increase with age. This pattern would produce an asymmetric distribution curve with a positive skew, characterized by a longer tail extending to the right.



To solve this drawback, we could implement, for example, a statistical procedure known as a log transformation, which would make it possible to normalize the data or reduce skewness.

The normal distribution is a model defined by a mathematical algorithm, which is calculated from a formula:

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

Where:

f(x)= Function of the variable x

σ= Population standard deviation

σ^2= Population variance

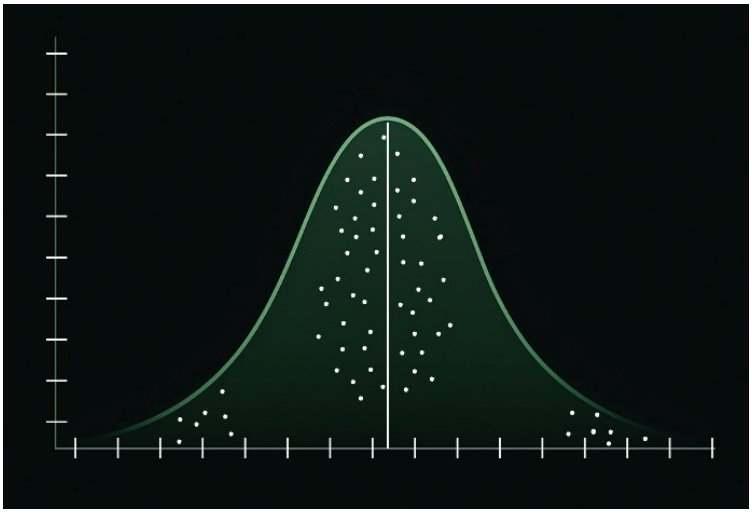
π= 3.1416

e= Base of the natural logarithm= 2.7182

μ= Population mean of the random variable X

This type of distribution, also known as the Gaussian distribution or bell curve, is one of the most important and widely used probability distributions in mathematics, statistics, and numerous scientific disciplines.

It is characterized by its symmetrical shape, in which most observations are concentrated around the mean. This property also reflects the fact that the probability of observing atypical or extreme values (outliers) decreases exponentially as the distance from the mean increases.



The German mathematician Carl Friedrich Gauss, from whom the term Gaussian curve or Gaussian distribution derives, introduced this model through his work and observations in fields such as astronomy, physics, and mathematics. Its development was particularly influenced by his studies of the method of least squares, number theory, and the distribution of prime numbers.

It is important to emphasize that this model enables statistical inference and, more importantly, that data interpretation depends on the context of the problem under investigation, as it describes the behavior of a numerical random variable under specific natural or experimental conditions.

In the analysis and interpretation of a normal distribution curve, the mean and variance are key parameters that determine the shape and location of the distribution.

- The mean defines the central point around which values are distributed, thereby determining the symmetry of the curve.
- The variance influences the dispersion of the data, thereby determining the width of the curve.

The standard deviation also plays a fundamental role in determining the shape of the distribution. A larger standard deviation and variance result in a wider and flatter curve, indicating that the values are more dispersed. Conversely, smaller values imply that the data are more tightly clustered around the mean, producing a narrower and more peaked distribution. The width of the curve refers to its visual appearance, specifically how broad or narrow it appears.

Variability refers to the degree of dispersion of values around the mean. It is worth noting that variance is more commonly used because its mathematical properties facilitate:

- Modeling of random processes
- Comparisons of variability
- Risk analysis
- Multivariate distributions

In fact, in the formula of the normal distribution, the variance appears in the denominator of the exponent.

Applications of the Normal Distribution

- Analysis of data exhibiting a symmetric and unimodal distribution.
- Inferential statistical procedures, particularly hypothesis testing.
- Construction and interpretation of confidence intervals.

Limitations

- Assumption of normally distributed data, which is not always realistic.
- Sensitivity to outliers, which can affect accuracy.

Other important concepts are the probability density function (PDF) and the area under the curve, which, although related, are not exactly the same.

The PDF specifically describes the probability that a random variable takes a particular value. It is a mathematical function that assigns a probability to each possible value of a variable.

The PDF has several properties, among which the following are particularly noteworthy:

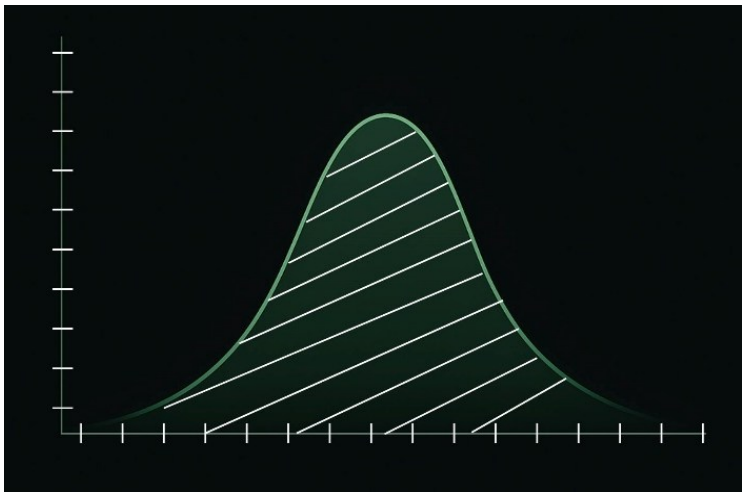
- Non-negativity: the function is non-negative for all values of x.
- Integral equal to 1: the integral of the function over all possible values of x is equal to 1.

Examples of Interpretation

- Maximum density: The highest point of the PDF indicates the most probable value of the random variable.
- Distribution tails: The tails of the distribution indicate the probability of extreme values.
- Symmetry: The symmetry of the PDF indicates whether the distribution is symmetric or asymmetric.

On the other hand, the area under the probability density function (PDF) curve represents the cumulative probability that the variable assumes a value within a specific interval:

- The total area under the curve is equal to 1, representing the total probability.



In other words, the probability density function (PDF) represents the height of the curve at each point, whereas the area under the curve corresponds to the integral of the PDF over a specified interval.

A normally distributed random variable with population mean (μ) and population standard deviation (σ) is commonly denoted as N(μ, σ).

Example

If the fasting blood glucose levels of a group of individuals have an arithmetic mean of 110 mg/dL and a standard deviation of 10 mg/dL, and the data follow a normal distribution, this can be expressed as N(110, 10).

The normal distribution is fundamental across many fields and is particularly useful when:

- The data exhibit a central tendency, meaning that most values cluster around a mean (the center of the distribution).
- Statistical models are required to make inferences about large populations based on sample data.
- A large number of observations are available. According to the Central Limit Theorem, if sufficiently large random samples are drawn from any distribution with a finite mean and variance, the distribution of the sample means will tend toward a normal distribution.

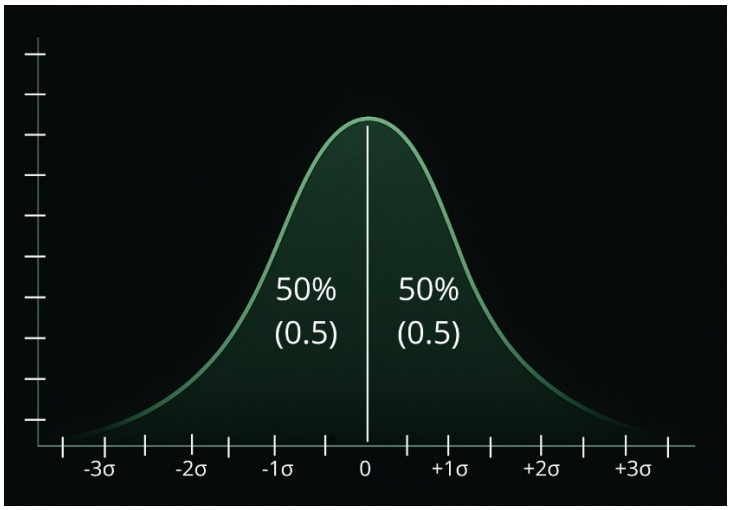
Examples of use

- Human height: In a given population, individuals' heights typically follow a normal distribution, with most people having heights close to the population mean and relatively few exhibiting extremely tall or short statures.
- Standardized academic test scores: The results of many standardized examinations tend to approximate a normal distribution. Most students achieve scores near the average, whereas exceptionally high or low scores occur less frequently

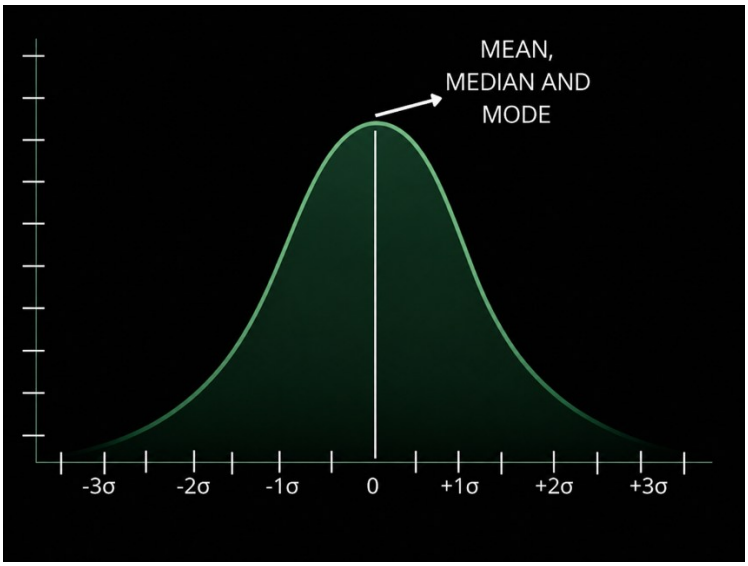
The normal curve, given its importance, has been studied extensively, and several of its properties are particularly useful for probability calculations.

To interpret the area under the Gaussian curve, the following should be considered:

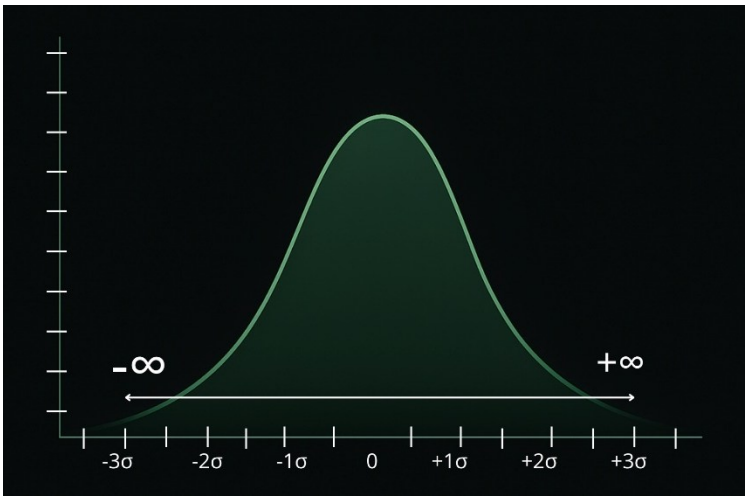
- The area under the curve is the same on both sides of the mean (50%, or 0.5 when expressed as a decimal), such that the right side is essentially a mirror image of the left side.



- A distinctive characteristic of the normal distribution is that the mean, median, and mode are equal. This property makes it a symmetric distribution, meaning that all three measures have the same value and are represented at the peak of the curve.

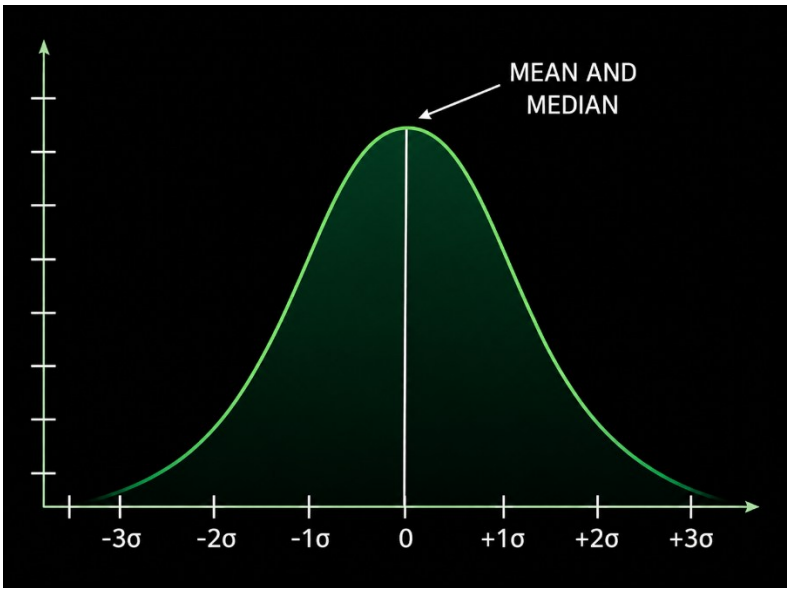


- The area between the curve and the x-axis represents the probability that the random variable assumes any value between $-\infty$ and $+\infty$; this probability is equal to 1.

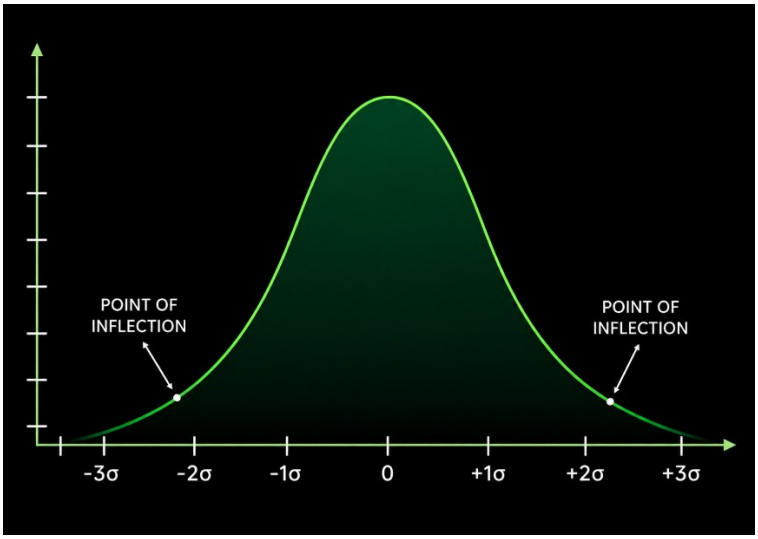


- Values located more than three standard deviations (σ) from the arithmetic mean have a substantially lower probability of occurrence.

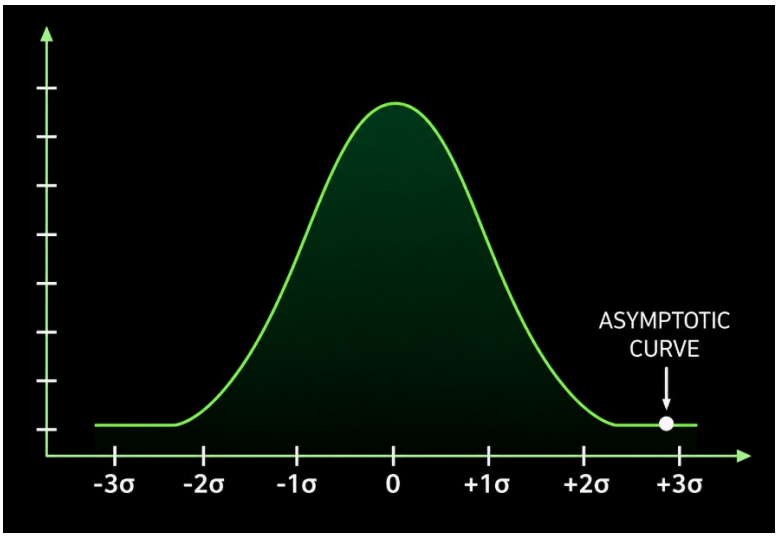
Any value located far from the mean is possible, although its probability of occurrence is low.



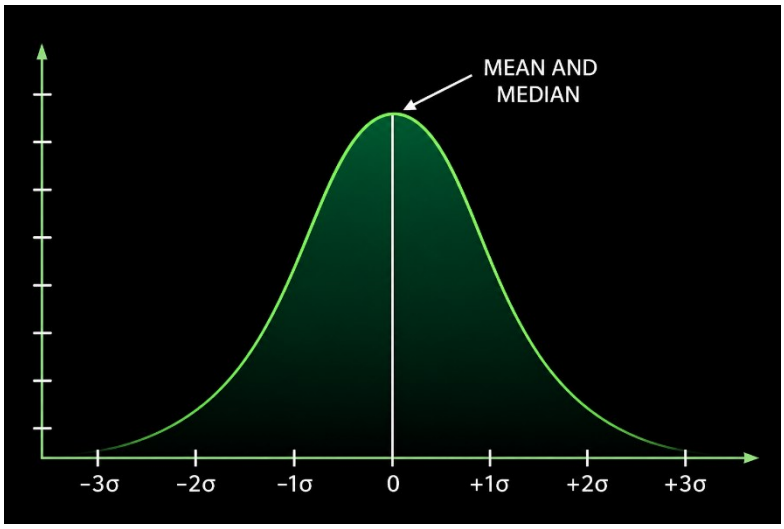
- The inflection points, where the curve changes from concave to convex, occur at values of X equal to (μ − σ) and (μ + σ).



- The tails of the distribution are asymptotic; that is, they approach but never actually intersect the horizontal axis (also known as the x-axis or axis of abscissas).



- The median divides the data into two equal parts in terms of the number of observations. However, when the data follow a normal distribution and the mean and median coincide, the arithmetic mean also divides the data into two equal halves with respect to the number of observations.



Z SCORES

Also known as standard scores, Z scores are standardized measures used to calculate the probability that a given score occurs within a normal distribution. They are expressed in terms of standard deviations from the mean. Z scores are also useful for comparing two scores that belong to different normal distributions.

To further understand the interpretation of the normal distribution, the concept of the area under the curve, and Z scores, it is first necessary to consider the concept of the standard deviation, which is a measure of dispersion for numerical variables and serves as the basis for calculating Z values.

Standard Deviation or Measure

- First, it should be pointed out that it is a measure of variability whose function is to assess the dispersion of data with respect to its mean, which makes it possible to measure how much individual values deviate from the mean of a dataset.
- Second, it identifies atypical values (outliers).
- Finally, it allows the comparison of the dispersion of different datasets and the calculation of confidence intervals.

The formula is:

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

Where:

σ= Standard deviation

x= Individual values

μ= Population mean

N= Number of observations

$\sum$ = Summation symbol (sum of values)

This measure of variability is particularly useful because it enables comparisons across variables expressed in different units of measurement.

Example: Comparison of ages Between two groups.

Suppose we have two groups of students:

Group A

- Ages: 18, 19, 20, 21, 22
- Mean: 20 years
- Standard deviation: 1.41 years

Group B

- Ages: 18, 22, 25, 28, 30
- Mean: 24.6 years
- Standard deviation: 4.76 years

Analysis

Although the mean age of Group B is higher, the key aspect is the comparison of the standard deviations. The standard deviation of Group A is low (1.41 years), indicating that the ages are clustered closely around the mean (20 years). In contrast, the standard deviation of Group B is high (4.76 years), indicating that the ages are more widely dispersed and farther from the mean (24.6 years).

Conclusion

Standard deviation enables the comparison of data variability between groups. In this example, Group A exhibits greater homogeneity in ages, whereas Group B shows greater diversity. Thus, standard deviation provides information beyond that offered by a measure of central tendency such as the mean alone. This can be useful in a variety of applications, including the planning of educational programs, preventive health initiatives, and the identification of health-related conditions within populations, to mention only a few examples.

In summary

- A high standard deviation indicates that the data are widely dispersed and far from the mean.
- A low standard deviation indicates that the data are clustered close to the mean and exhibit little variability.

It is important to note that raw numerical scores are frequently transformed into other types of scores that facilitate more meaningful comparisons and more intuitive interpretations.

One method of transforming data is through the use of Z scores, also known as standard scores or z-values.

This approach is considered a method of standardizing data within a normal distribution, allowing the determination of how many standard deviations a particular value lies from the mean of the dataset.

Interpretation of the Z Score

$Z = 0$ the value is exactly at the mean.
$Z > 0$ the value is above the mean.
$Z < 0$ the value is below the mean.

Formula for calculating a z Score:

$$z = \frac{x - \mu}{\sigma}$$

Where:

x = individual value

μ = mean of the dataset

σ = population standard deviation

The greater the absolute value of the z-score, the farther the value is from the mean.

Example:

Suppose we are analyzing examination scores..

- The mean is μ=70
- The standard deviation is σ=10

We want to determine the z-score of a student who obtained X=85

Applying the formula:

$$z = \frac{x - \mu}{\sigma} = \frac{85 - 70}{10} = \frac{15}{10} = 1.5$$

Interpretation of the result

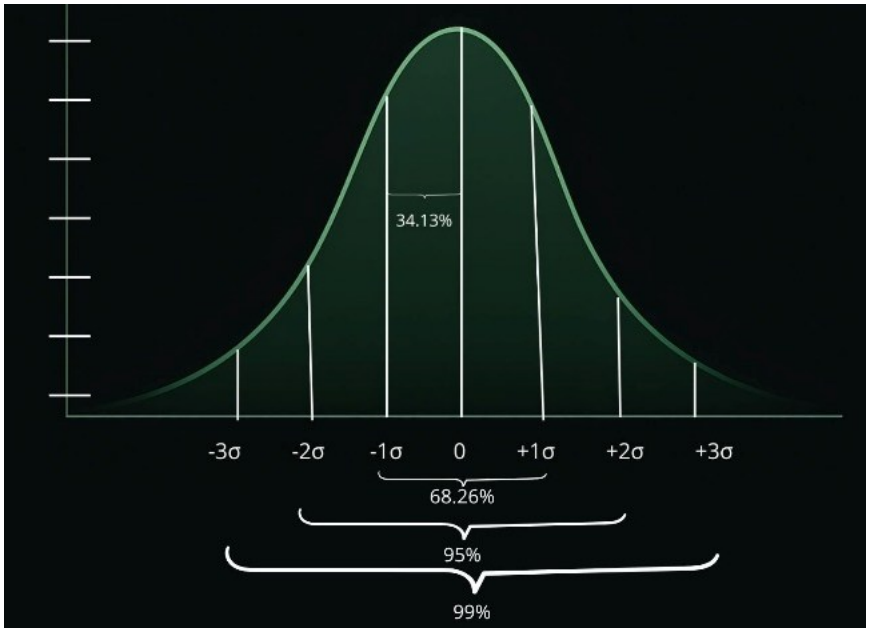
The student who scored 85 points is 1.5 standard deviations above the mean, indicating performance that is notably higher than the average. A z-score is obtained by subtracting the mean from an individual value and dividing the result by the standard deviation.

This procedure is particularly useful for identifying outliers and determining how many standard deviations a given value lies above or below the arithmetic mean within a distribution. When the distribution is strictly normal, the z-score distribution has a mean of 0 and a standard deviation of 1.

Additional considerations:

- The area under the curve represents 100% of the data; one-half of the area corresponds to 50% on each side of the distribution.
- In proportional terms, the total area under the curve is equal to 1.

When a set of numerically expressed values measured on an interval scale follows a normal distribution, their representation as Z scores is as follows:



- Within one standard deviation to the right and left of the mean, 68.26% of all observations are contained (therefore, each side of the curve contains 34.13%).
- Within two standard deviations to the right and left of the mean, 95.44% of all observations are contained (each side of the curve contains 47.72%). In practice, this value is commonly rounded to 95%.
- Within three standard deviations to the right and left of the mean, 99.74% of all observations are contained (each side of the curve contains 49.87%), that is, only 2.15% more than that contained within two standard deviations. In practice, this value is commonly rounded to 99%.

For greater precision, it should be noted that, in practice, 95% of the data fall between -1.96 and +1.96 standard deviations, 99% fall between -2.58 and +2.58 standard deviations, and 99.9% fall between -3.90 and +3.90 standard deviations. This clarification is fundamental for inferential statistical procedures.

In summary, z-scores represent a statistical measure that indicates how many standard deviations a value lies from the mean of a dataset. This tool is useful for comparing observations from different distributions and for assessing their relative position with respect to the mean.

Example A:

Suppose we have a set of examination scores obtained by university students, and we wish to calculate the Z score for a particular student. (For this example, assume that the standard deviation is not initially available)

Scores of 10 students: 70, 85, 90, 60, 75, 80, 95, 55, 85, 100

Step 1: Calculate the mean (average)

The mean (μ) is calculated by summing all values and dividing the result by the total number of observations:

$$\mu = \frac{70 + 85 + 90 + 60 + 75 + 80 + 95 + 55 + 85 + 100}{10} = \frac{835}{10} = 83.5$$

Step 2: Calculate the standard deviation (σ)

The standard deviation was calculated using the following formula:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

Where Xi represents the values in the dataset, μ is the mean, and N is the total number of observations. For this dataset, the calculation would be as follows:

$$\sigma = \sqrt{\frac{(70 - 83.5)^2 + (85 - 83.5)^2 + (90 - 83.5)^2 + \cdots + (100 - 83.5)^2}{10}} = \frac{\sqrt{4402.5}}{10}$$

$$= \sqrt{440.25} \approx 20.98$$

Step 3: Calculate the Z score for a specific score

The Z score is calculated using the following formula:

$$z = \frac{x - \mu}{\sigma}$$

Where X is the student's score. If a student obtained a score of 95 points, then:

$$z = \frac{95 - 83.5}{20.98} \approx \frac{11.5}{20.98} \approx 0.55$$

Interpretation

A z-score of 0.55 means that the raw score of 95 points is 0.55 standard deviations above the mean (a value within one standard deviation). In other words, the student has a score that is slightly above the class average (the score lies between the mean and one standard deviation above it).

APPLICATION IN MEDICINE

Example B:

Child Growth

Z scores allow the comparison of a child's weight, height, and body mass index (BMI) with the expected reference values for children of the same age and sex.

A 5-year-old child weighs 20 kg.

The mean weight for children of the same age is 18 kg, with a standard deviation of 2 kg (in this case, the standard deviation is available).

The Z score is calculated as follows:

$$z = \frac{x - \mu}{\sigma} = \frac{20 - 18}{2} = \frac{2}{2} = 1.0$$

Interpretation

A Z = 1.0 means that, in terms of weight, the child is 1 standard deviation above the mean, which suggests that the child's weight is above average, but within the normal range.

If Z were greater than +2 or less than −2, it would indicate significant overweight or underweight.

Example C:

Bone Densitometry

In studies of bone mineral density (BMD), such as dual-energy X-ray absorptiometry (DXA), the Z score is used to compare a patient's bone density with that of individuals of the same age and sex.

A 60-year-old woman has a BMD of 0.85 g/cm².

The mean BMD for her age group is 1.0 g/cm², with a standard deviation of 0.1 g/cm².

Z Score:

$$z = \frac{0.85 - 1.0}{0.1} = \frac{-0.15}{0.1} = -1.5$$

Interpretation

A Z = -1.5 suggests that her bone density is 1.5 standard deviations below the mean, that is, lower than average, which could indicate osteopenia; even lower values would suggest osteoporosis.

Example D:

Laboratory tests and clinical values

Z scores can also be used to analyze biomarkers in laboratory tests, such as glucose and cholesterol levels, or in the analysis of vital signs such as blood pressure, comparing them with population reference values.

A patient has a systolic blood pressure of 140 mmHg.

The population mean is 120 mmHg with a standard deviation of 10 mmHg.

Z Score:

$$z = \frac{140 - 120}{10} = \frac{20}{10} = 2.0$$

Interpretation

A Z = 2.0 indicates that the patient's blood pressure is 2 standard deviations above the mean, higher than the average, which could suggest systolic arterial hypertension.

REFERENCES

Altman DG. Practical Statistics for Medical Research. Boca Raton (FL): CRC Press; 1991.

Bland M. An Introduction to Medical Statistics. 4th ed. Oxford: Oxford University Press; 2015.

Daniel WW, Cross CL. Biostatistics: A Foundation for Analysis in the Health Sciences. 10th ed. Hoboken (NJ): Wiley; 2013.

Devore JL. Probability and Statistics for Engineering and the Sciences. 9th ed. Boston (MA): Cengage Learning; 2015.

Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Hoboken (NJ): Wiley; 2013.

Kirk RE. Statistics: An Introduction. 5th ed. Belmont (CA): Wadsworth Publishing; 2007.

Martínez González MA, Sánchez-Villegas A, Toledo Atucha E. Bioestadística amigable. 2ª ed. Barcelona: Elsevier; 2020.

Montgomery DC, Runger GC. Applied Statistics and Probability for Engineers. 7th ed. Hoboken (NJ): Wiley; 2018.

Pagano M, Gauvreau K. Principles of Biostatistics. 2nd ed. Belmont (CA): Duxbury Press; 2000.

Sokal RR, Rohlf FJ. Biometry: The Principles and Practice of Statistics in Biological Research. 4th ed. New York: W.H. Freeman and Company; 2012.

Triola MF. Estadística. 13ª ed. Madrid: Pearson Educación; 2019.

Zar JH. Biostatistical Analysis. 5th ed. Upper Saddle River (NJ): Pearson Prentice Hall; 2010.