

# Data distribution analysis – a preliminary approach to quantitative data in biomedical research

Przemysław Guzik

Department of Cardiology – Intensive Therapy,  
Poznan University of Medical Sciences, Poland  
University Centre for Sports and Medical Studies,  
Poznan University of Medical Sciences, Poland

 <https://orcid.org/0000-0001-9052-5027>

Corresponding author: [pguzik@ptkardio.pl](mailto:pguzik@ptkardio.pl)

Barbara Więckowska

Department of Computer Science and Statistics,  
Poznan University of Medical Sciences, Poland

 <https://orcid.org/0000-0002-1811-2583>


**Keywords:** statistical analysis, medical research, quantitative data, normal distribution, parametric tests

**Received** 2023-06-12

**Accepted** 2023-06-23

**Published** 2023-06-27

**How to Cite:** Guzik P, Więckowska B. Data distribution analysis – a preliminary approach to quantitative data in biomedical research. *Journal of Medical Science*. 2023;92(2);e869. doi:10.20883/medical.e869

 DOI: <https://doi.org/10.20883/medical.e869>



© 2023 by the author(s). This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC) licence. Published by Poznan University of Medical Sciences

## ABSTRACT

Statistical analysis is an integral part of medical research. It helps transform raw data into meaningful insights, supports hypothesis testing, optimises study design, assesses risk and prognosis, and facilitates evidence-based decision-making. The statistical analysis increases research findings' reliability, validity and generalisability, ultimately advancing medical knowledge and improving patient care. Without it, meaningful analysis of the data collected would be impossible. The conclusions drawn would be unsubstantiated and misleading.

Many health professionals are unfamiliar with statistical analysis and its basic concepts. The analysis of clinical data is an integral part of medical research. Identifying the data type (continuous, quasi-continuous or discrete) and detecting outliers are the first and most important steps. When analysing the data distribution for normality, graphical and numerical methods are recommended. Depending on the type of data distribution, appropriate non-parametric or parametric tests can be used for further analysis. Data that are not normally distributed can be normalised using various mathematical methods (e.g., square root or logarithm) and analysed using parametric tests in the next steps.

This review provides essential explanations of these concepts without using complex mathematical or statistical equations but with several graphical examples of various statistical terms.

## Introduction

Statistical analysis is essential in medical research. It transforms raw data into meaningful

insights, supports hypothesis testing, optimises study design, assesses risk and prognosis, and facilitates evidence-based decision-making. The statistical analysis increases research findings'

reliability, validity and generalisability, ultimately advancing medical knowledge and improving patient care. Without it, meaningful analysis of the data collected would be impossible. The conclusions drawn would be unsubstantiated and misleading.

Many medical professionals are unfamiliar with statistical analysis and its basic concepts, starting with the types of quantitative data, such as continuous and discrete data, or their distribution analysis. This review provides essential explanations of these concepts without using complex mathematical or statistical equations but with several graphical examples of various statistical terms.

## Data types

Different types of data are collected in biomedical research. The most common are quantitative, qualitative and descriptive (textual).

Quantitative or numerical data can take any numerical value and are represented as numbers. Some values are less than, equal to, or greater than others, for example, age 15, 21, and 35 years; length 22, 19, and 10 cm; area 2, 2, and 2.5 cm<sup>2</sup>; weight 78, 82, and 95.3 kg; power 224, 248, and 301 watts; or a ratio of two variables such as serum triglycerides to HDL concentrations of 1.2, 3.5, and 4.9. Quantitative data may or may not have units.

Qualitative data are usually non-numerical and are described by labels or qualitative characteristics. Many qualitative variables can only be categorised, labelled, but never ranked, ordered or graded, such as gender (male, female), ethnicity (e.g. African American, European, Latin American) or colour (red, yellow, green, black). For example, red is not bigger or smaller than blue. However, other qualitative observations can be ranked in a natural order based on qualitative analysis. However, the distances between the categories are unknown. One object may be larger than another. One person may be nicer than another. One heart failure patient in New York Heart Association (NYHA) functional class 2 has less severe symptoms than another in NYHA 3. Some examples of ordinal data for which relative, subjective or arbitrary scales should be used are warm, warmer and warmest, or primary, high

school, college, graduate and postgraduate for educational level. Similarly, the effect of pharmacotherapy on a patient's symptoms can be subjectively rated as no change, slightly better, really better compared to previous treatment, or best of all medications taken before. For qualitative data, signs (+, ++, +++ or -, --, ---) and letter codes (A, B, C) are often used instead of longer words. As some statistical software does not accept text, numbers are used as codes in such cases. The numerical codes entered should be treated as nominal (preferably) or ordinal (if they can be ranked) data. Otherwise, numbers replacing text may be treated as continuous and become a source of error.

Descriptive data are typically textual and consist of words, abbreviations, phrases and sentences, e.g. medical notes, observations, test summaries, open-text comments and opinions. Specialised analysis tools are required to quantify and/or describe such data. These tools can be natural language processing techniques, Qualitative Text Analysis (QCA) or other methods such as the Generative Pre-trained Transformer (GPT), which is part of the family of Large Language Models (LLMs) analysed by artificial intelligence (ChatGPT).

Regardless of the type, all data are collected in databases. Data can be stored in spreadsheets or dedicated databases. Spreadsheets provide a tabular format with rows and columns to store and manage data. Most people find them easier to use for entering, manipulating and analysing data effectively, performing calculations, applying formulas, formatting and exporting to external statistical software. Unfortunately, spreadsheets have limitations compared to dedicated database management systems for large and more complex data sets. Dedicated and specially designed database management systems such as MySQL, Oracle, and Microsoft SQL Server, the Redcap Project are better solutions for such tasks. They offer features such as data indexing, data relationships, integrity constraints, normalisation and transaction management.

Regardless of the form of the database, various characteristics or parameters (variables) describing people, objects, animals, samples, etc., are stored and prepared for further statistical analysis.

This review focuses on quantitative data. We will explain the most common terms, how to analyse and interpret their distribution and present graphical examples.

## Basic definitions

Several basic terms related to the types of quantitative data, their characterisation and the analysis of the distribution of data are presented in **Table 1**.

## Types of quantitative data

Continuous data can have infinite possible values within a given range, including fractions, decimals or integers. Between any two values, there is always another. The reporting of each value depends on the precision of the measurement, which may determine whether the data are continuous. For example, a precision of 1 in 100 is considered continuous data, as opposed to 1 in 10, which makes the data quasi-continuous (almost continuous) or sometimes discrete because it appears to be stepped.

Medical examples of continuous data include cardiac cycle duration (910.9, 920.0, 920.1 ms), age (31.85, 31.86, 31.87 years), body temperature (36.58, 36.59, 36.60 °C), body mass index (BMI) (27.27, 27.28, 27.29 kg/m<sup>2</sup>), blood glucose concentration (11.64, 11.65, 11.66 mmol/L).

Quasi-continuous data represent values that have been rounded or grouped into intervals. Using the same examples above, the rounded values will be 920 ms for the cardiac cycle, 32 years for age, 36.6°C for body temperature, 27.3 kg/m<sup>2</sup> for body mass index and 11.7 mmol/L for blood glucose concentration. Some values in clinical practice are always rounded and given in whole numbers, such as heart rate – 63, 86, 105 beats/minute, blood pressure – 122/78, 124/84, 152/95 mmHg, body mass – 56, 78, 113 kg, etc. Quasi-continuous data, however, should be considered continuous for statistical analysis.

Discrete data can only take specific values and have no value between two adjacent values. Typical examples are the number of pregnancies (there cannot be 3.35 pregnancies) and the number of hospitalisations (it is impossible

Possible data transformation at a cost of precision loss →

BMI as			
Original data	Transformed data		
continuous	quasi-continuous	discrete	ordinal categories
17.63148	17.6	18	1 – underweight
21.24743	21.2	21	2 – normal weight
23.22671	23.2	23	2 – normal weight
25.37863	25.4	25	3 – overweight
26.04873	26	26	3 – overweight
27.24712	27.2	27	3 – overweight
33.56914	33.6	34	4 – class 1 obesity
35.14975	35.1	35	5 – class 2 obesity
41.38937	41.4	41	6 – class 3 obesity

← Impossible data transformation

**Figure 1.** A sample of different original body mass indices derived from height and mass. These data are transformed from continuous through quasi-continuous, discrete to ordinal. Each further step involves a loss of accuracy. Transforming data to a less precise category often involves grouping observations into predefined ranges or bins. This results in a loss of information and granularity. Subtle differences between individuals may be obscured, making it harder to see fine patterns or relationships in the data or showing weak or no associations between variables. The reverse process of recovering the original information (continuous data) from all the transformed data is mathematically unfeasible.

to be hospitalised 5.173 times). The main difference between continuous and discrete data is that continuous data can take any value within a specific range, whereas discrete data can take only certain values. Continuous data are measured and expressed more accurately than discrete data.

Mathematical manipulation with continuous and discrete data is possible; e.g., measuring height and weight makes it easy to calculate BMI. Similarly, converting continuous or discrete data into categorical data is also straightforward. All such data belong to an interval or ratio scale.

Based on BMI and known criteria, a person is categorised as underweight, normal weight, overweight or obese category 1, 2 or 3. However, this is a one-way process. Retrieving backward information on BMI from one of these categories is impossible (see **Figure 1**).

## Outliers

An outlier is a value significantly different from other values in the dataset. Measurement inaccuracies, data entry errors, natural variation, or truly unusual observations are the leading causes of

**Table 1.** Basic terms related to the types of quantitative data, characterisation and analysis of distribution.

Term	Definition
Descriptive statistics	Analyses designed to describe and summarise the data set.
Continuous data	Data that can take on any value within a range, e.g., body temperature, serum sodium concentration, white blood cells count, and time.
Quasi-continuous data	Data nearly continuous or continuous data that were rounded for some purposes, e.g. age in years, body weight in kilograms, blood pressure in mmHg, heart rate in beats/minute.
Discrete data	Data that can only take on specific values, e.g., the number of children in a family, the number of fingers and toes, and the number of epilepsy attacks.
Distribution	It displays the rate or probability of different values occurring in a given data set.
Histogram	Graphical representation of the distribution of numerical data binned into neighbouring bars.
Density plot	Graphical visualisation of the distribution of continuous data as a smooth curve with continuous data representing the data distribution.
Q-Q plot	Short term for the quantile-quantile plot. Graphical visualisation of assessing whether data follow a normal distribution.
Outlier	An observation or data point with an extreme value that lies far away from most data points.
Minimum	The smallest value in a dataset.
Maximum	The largest value in a dataset.
Percentile	A measure used to indicate the value below which a given percentage of observations in a group of observations falls. For example, the 5th percentile indicates that 5% of the values in a dataset are less than or equal to that value.
Lower quartile (Q1)	25 <sup>th</sup> percentile, a value below or equal to which 25% of the values in the dataset are located.
Upper quartile (Q3)	75 <sup>th</sup> percentile, a value below or equal to which 75% of the values in the dataset are located.
Interquartile range (IQR)	The range between a dataset's first quartile (25 <sup>th</sup> percentile) and the third quartile (75 <sup>th</sup> percentile).
Range	Distance between the maximum and minimum values of a data set.
Central Tendency	Various measures indicating the middle or centre of a distribution.
Median	Middle value (50 <sup>th</sup> percentile) in a dataset ranked from minimum to maximum values.
Mode	The most common value in a dataset.
Mean	The average value of a dataset calculated by adding up all the values and dividing by the number of values.
Trimmed mean	A statistical measure calculated by removing a certain percentage of the largest and smallest values in a dataset and then calculating the mean of the remaining values. It is done to reduce the impact of outliers on the mean, for instance, removing 5% of the measurements reduces 2.5% of the largest and 2.5% of the smallest values.
Normal distribution	A bell-shaped curve represents the distribution of many biological phenomena and data.
Skewness	A measure of how asymmetrical a distribution is.
Kurtosis	A measure of how peaked or flat a distribution is compared to a normal distribution.
Deviation	The distance between the mean and a particular data point in a given distribution.
Standard deviation (SD)	A measure of how much the data deviates from the mean.
Variance	A measure of how spread out the data is from the mean.
Coefficient of variation (CV)	A relative and unitless measure of the dispersion of data points around the mean. It allows comparing variability between disparate groups and characteristics. A smaller CV indicates that the data points are more tightly clustered around the mean, while a larger coefficient of variation indicates that the data points are more spread out.
Standard error of the mean (SEM)	A measure of how much the sample mean is likely to differ from the true population mean to assess the precision of the sample mean. It is derived by dividing the standard deviation by the root of the sample size.
Confidence Interval (CI)	A range of values likely to contain the true population parameter with a certain confidence level. CI is usually expressed as a percentage, such as 95% or 99%
Z-score	A statistical measure that determines the relative distance of a given value from the mean, using standard deviation as the measure of that distance. In other words, the Z-score represents the number of standard deviations a data point is from the mean of the distribution. It is calculated as the difference between the given value and the mean divided by the standard deviation. By using the Z-score, data points from different distributions can be standardised and compared on the same scale. A positive z-score indicates that the data point is above the mean, while a negative z-score indicates that it is below the mean. A Z-score of 0 means that the data point is exactly at the mean.

these extreme values. Outliers can significantly impact statistical analyses and distort the results or conclusions drawn from the data. They can affect some measures of central tendency, such as the mean, and estimates of variability, such as the standard deviation. Outliers can violate statistical methods assuming normal data distribution. In contrast, non-parametric methods and descriptors such as median, interquartile range (IQR) or mode are insensitive to outliers.

Identifying outliers is essential to ensure the integrity and validity of data analysis. It involves examining the data distribution and looking for values unusually far from most observations. Outliers can be detected using various methods, including graphical techniques (e.g. box plots, scatter plots, violin plots), statistical tests and computational algorithms. The decision to deal with outliers depends on the specific research context, the nature of the outliers and the analysis objectives. In research, it is essential to document all procedures for identifying and dealing with outliers. This ensures transparency and reproducibility.

## Types of data distribution

The normal distribution is very common in biomedical research. It is also known as the Gaussian distribution or the bell curve. The normal distribution is symmetrical about the mean and has a characteristic bell shape. Many biological and physical phenomena follow a normal distribution. For example, the heights of adult humans follow a normal distribution.

Skewed distributions are another type of distribution. They occur when the data do not have symmetrical distribution around the mean. There are two types of skewed distribution: positively skewed (to the right) and negatively skewed (to the left). In a positively skewed distribution, the curve's tail is longer on the right than on the left. In a negatively skewed distribution, the curve's tail is longer on the left than on the right. A common example of a positively skewed distribution is income data, where many people have low incomes, and a few have very high incomes. The age distribution of patients admitted to a hospital with neonatal and paediatric wards is different and skewed to the right compared to another hos-

pital where only adults, especially older people, are admitted. The mean serum creatine concentration is higher, and the distribution is skewed to the left in nephrology patients compared to general medical ward patients.

Bimodal distributions occur when there are two peaks in the data. This happens when two different subgroups in the same data differ and emerge. For example, in a combined data set, the average muscle mass for men and women differs. It naturally separates – such data distributions show two peaks.

Multimodal distributions occur when there are more than two peaks in the data. They occur when more than two data groups have different characteristics. For example, the distribution of the height of men, women and children in the same database shows three peaks in the data. Often this represents unbalanced data collection, such as more young people or more women than men.

## Distribution analysis

Examining distributions is an integral part of data analysis. It involves comparing the characteristics of two or more distributions to determine whether they are different or similar. Several graphical and numerical methods can be used to compare distributions.

### Graphical data visualisation used to analyse data distributions

In medical research, data visualisation is invaluable for analysing all forms of quantitative data. Various data features can be identified using appropriate visualisations, such as central tendency, dispersion, minimum and maximum values, outliers, data distribution and shape. These visualisations make interpreting data, communicating findings, and drawing conclusions easier. Presenting complex data in visual formats simplifies identifying differences, associations, trends and patterns.

1. A *histogram* is used to estimate the probability distribution of all forms of quantitative data, preferably continuous and quasi-continuous. Although they are not best suited to discrete data, histograms can provide insight into the number of cases with certain discrete values.

To construct a histogram, the data values are first binned into ranges, i.e. the whole range is divided into a series of narrower intervals. The number of values falling within each bin is then counted (see **Figure 2**). The bins represent successive, adjacent, non-overlapping intervals of a variable and are often (but not necessarily) of equal size. Histograms show the empirical shape of the distribution, central tendency (mean, median, mode), dispersion (variance, standard deviation) and outliers.

2. A *density plot* also shows the distribution of continuous data, similar to a histogram. However, the density plot uses a smooth theoretical curve instead of bars to represent the distribution. While they may not be the perfect way to visualise discrete data, density plots can give some insight into their distribution.

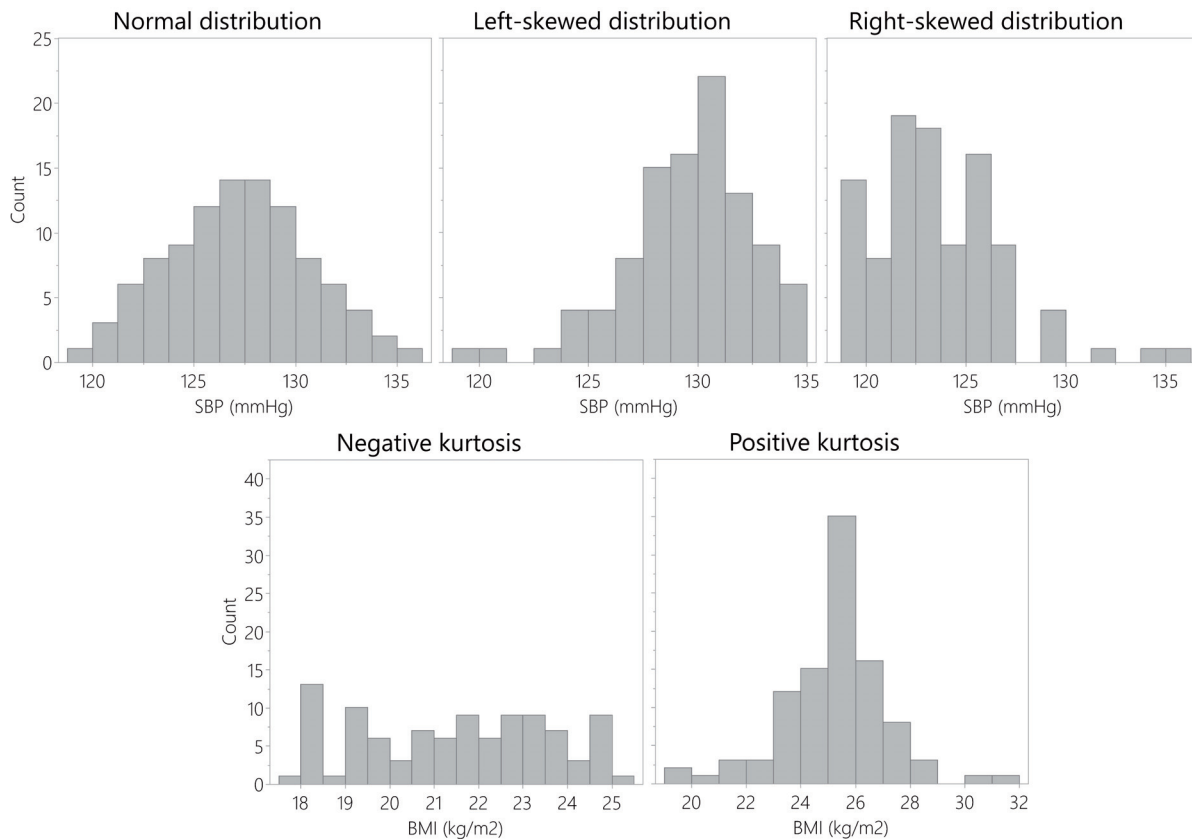
The smoothness of density plots refers to the degree of smoothing applied to the density curve drawn over the real data. With low smoothness, usually, more than one peak is

visible. Higher smoothing degrees provide only one peak and may resemble a normal or skewed distribution curve (see **Figure 3**).

These plots are particularly useful for identifying the shape of the distribution, including whether it is symmetric, skewed or bimodal. They also provide information about the central tendency, dispersion and outliers present in the data.

Using a theoretical plot fitted to real data has its consequences. Very often, such a plot crosses the real data at the lower and upper limits, and sometimes the density plots show values that are not possible. There are no negative values for concentration, length or weight. A height of 300 cm is humanly impossible. These are artificial effects of the smoothing algorithms, which can stretch the estimated density curve to values that do not make sense for a particular dataset.

To deal with such a problem, a density plot can be truncated at zero to avoid negative val-



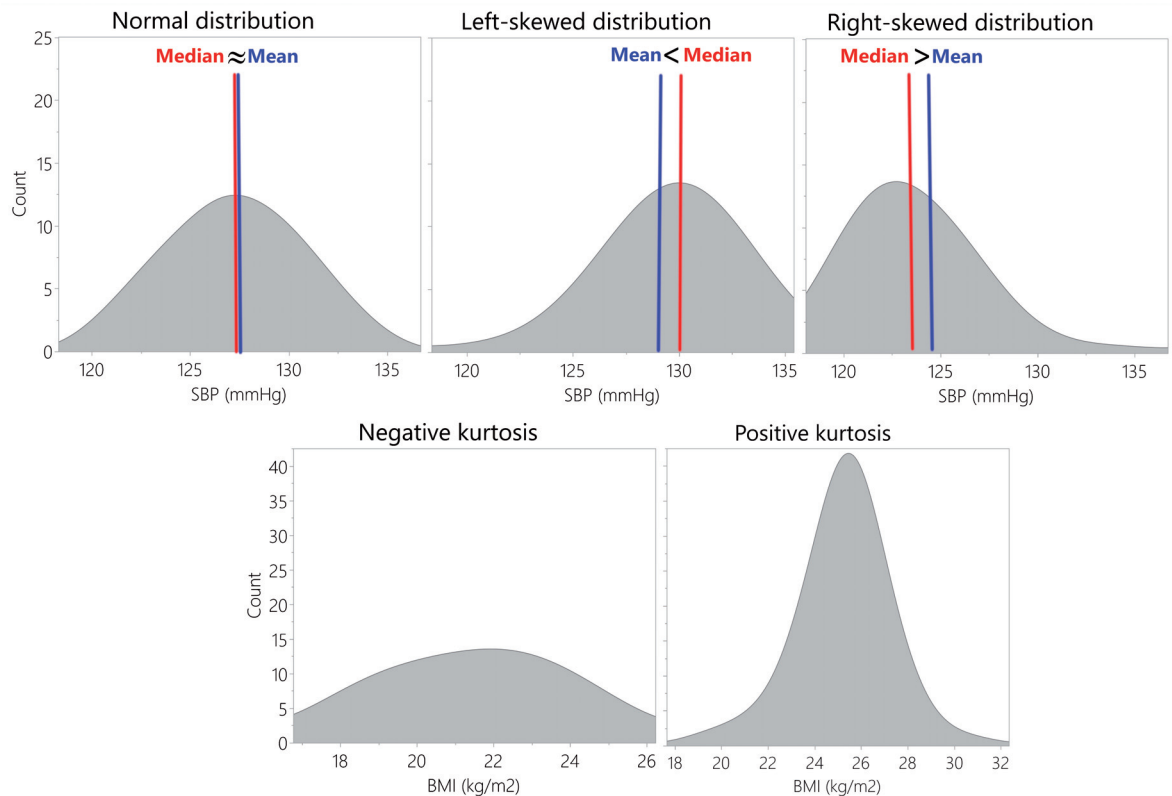
**Figure 2.** Examples of different histograms with systolic blood pressure (SBP) results in the upper panels and body mass index (BMI) in the lower panels. The first histogram with SBP shows a normal data distribution. The next two examples present data skewed to the left and right. The two BMI examples at the bottom display distributions with negative kurtosis (flattened shape) and positive kurtosis (narrower and higher shape).

ues – all negative values are set to zero. The upper range of these values should be defined for extremely high values.

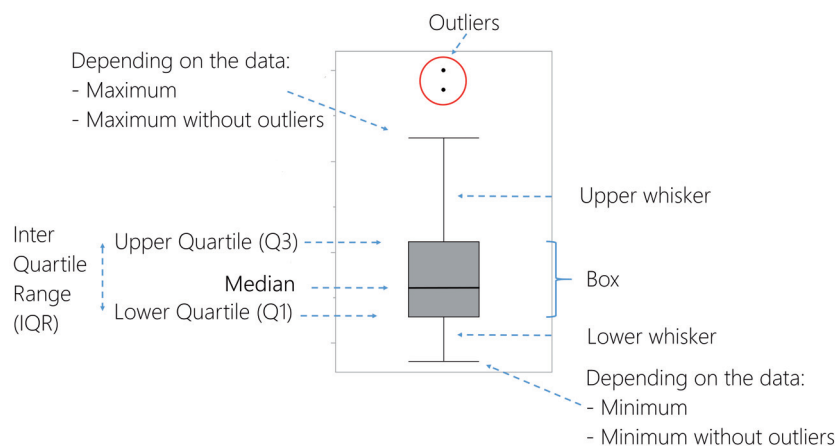
3. A box plot or box-and-whisker plot shows the distribution of all types of quantitative data. It summarises vital statistical characteristics

and highlights potential outliers in the data set (see **Figure 4**).

The following statistical measures are typically used to construct a box plot. – Median, represented by a horizontal line inside a box, dividing it in half. – Quartiles, i.e. the lower



**Figure 3.** For this figure, the same data were used as in Figure 2. The histograms are replaced by density plots showing the data with a normal distribution, skewed to the right and the left, and then with positive and negative kurtosis (leptokurtic and platykurtic distribution). The mean and median are usually overlapping or very close to each other for the normal distribution. In contrast, for skewed data, the mean and median are separated. Negative or positive kurtosis does not affect the position of the mean and median.



**Figure 4.** A general explanation of the box-whisker plot. The median represents the central tendency, while minima, maxima, outliers, whiskers and quartiles are different ways of expressing the dispersion of the data. The unequal distances between the median and Q1 and Q3, or the top and bottom whiskers, reflect whether the data are skewed or not. In this example, the data are right-skewed.

quartile (Q1) for the 25th percentile and the upper quartile (Q3) for the 75th percentile. The distance between Q1 and Q3 helps identify the data's spread. In the box plot, Q1 and Q3 are represented by the lower and upper boundaries of the box, respectively. This distance is called the interquartile range (IQR) and covers the middle half (50%) of all values in the data set. – Whiskers that extend from the box indicate the dataset's range. The lower whisker typically represents the minimum non-outlying value within 1.5 times the IQR below Q1, while the upper whisker represents the maximum non-outlying value within 1.5 times the IQR above Q3. Values outside the whiskers are considered outliers and are plotted individually. – Outliers, shown as individual data points or asterisks, are outside the whiskers (more on outliers in a separate section). They are considered to be potential anomalies in the data set.

Box plots are a flexible way of presenting data and may display the mean, SD, SEM or 95% CI. In this situation, the statistical analysis uses the Z-score to identify outliers or unusual values.

4. A *violin plot* combines features of a box plot and a kernel density plot to provide a comprehensive representation of the shape, central tendency, dispersion and multimodality of the data.

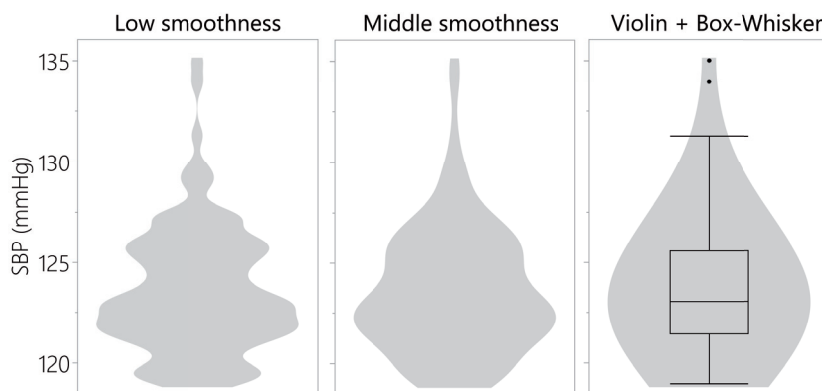
The width of the violin at each point represents the density of the data at that value.

In contrast, the body of the violin plot shows the density distribution, indicating the relative concentration of data at different values along the x-axis. Wider sections indicate higher density, while narrower sections indicate lower density (see **Figure 5**). Similarly to the box plot, the violin plot can include lines representing the data's median, Q1 and Q3 (IQR) and the outliers. Unlike box plots, violin plots show the shape and distribution of the data, indicating whether it is symmetrical, skewed, unimodal, bimodal or multimodal, with multiple peaks or modes representing different subgroups or patterns within the data.

5. A *scatter plot* displays individual data points as dots along a number line or axis. It shows the data's distribution and helps identify patterns or outliers.

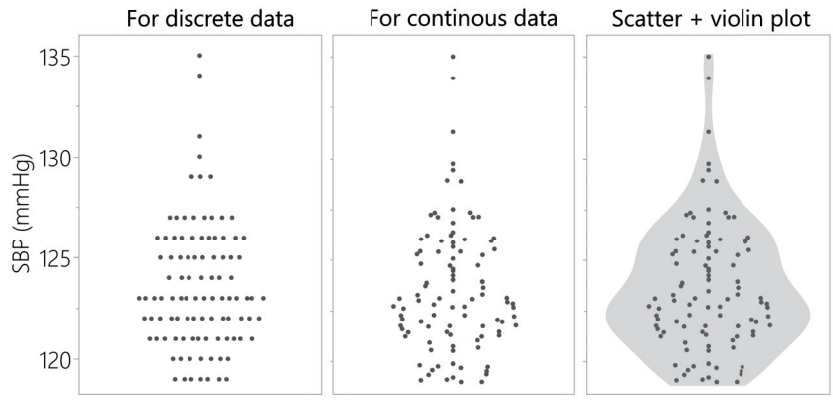
To create a scatter plot, each data point is plotted as a dot at its corresponding value on the number line. The dots are stacked vertically for multiple data points with the same value (non-unique or tied values). This stacking shows the frequency or density of data at each unique value (see **Figure 6**).

Unlike other plots that aggregate data, scatter plots show each data point. This allows the entire raw data set to be observed and specific values or patterns of interest to be identified. This makes it easy to follow the spread and concentration of data, with gaps or clusters of dots indicating areas of high or low density, uneven data distribution. Scatter



**Figure 5.** Identical systolic blood pressure (SBP) values are presented in three violin plots with different degrees of smoothness (low, medium and high). Low smoothness gives more information about the number of local peaks. With a more aggressive high level of smoothness, the violin is unimodal. Medians, Q1, Q3 or outliers can be added to all charts. Violin plots help to see if the data is skewed – the plots shown are right-skewed.





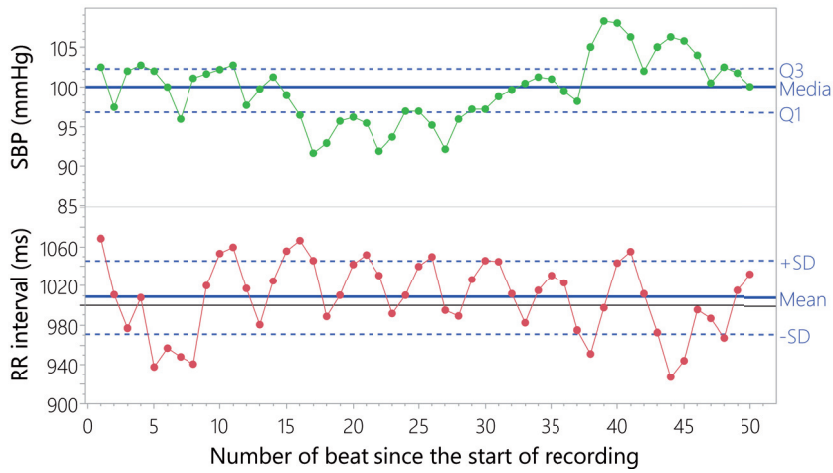
**Figure 6.** Systolic blood pressure (SBP) values presented as discrete data (left panel) and next (middle panel) as continuous data. Finally, a scatter plot is combined with a violin plot (right panel). In all cases, the data are randomly distributed around the centre of the scatterplot, but the shape of the scatterplot follows the data distribution. All forms of scatterplot can be supplemented with additional graphs, such as violin or box-whiskers plots, or measures of central tendency (Median, Mode, Mean) and dispersion (SD, Q1, Q3). Skewness can also be visualised using scatter plots.

plots can show measures of central tendency, such as the mean or median, SD or Q1 and Q3, and outliers.

6. A *line plot* presents quantitative data by connecting successive points that change over time. Many measurements are taken repeatedly to study their changes, e.g. blood glucose concentration before and after meals, blood pressure each morning and evening, and body weight during a weight loss programme. They show trends, patterns, and fluctuations over the observed period. The line plot is an example of a time series plot.

To construct a line plot, data points are plotted on the y-axis, representing the studied variable against time. Connected data with straight lines highlight the changes and trends over the observed period (see **Figure 7**).

By checking line plots, it is possible to reveal overall trends or patterns in the data, and the slope provides information about the direction and magnitude of the change, whether it is increasing, decreasing, or staying relatively constant over time. These plots help identify seasonal or periodic patterns, recurring fluctuations or cycles. As outliers deviate



**Figure 7.** Two samples of line plots of synchronised beat-to-beat recordings of the duration of each cardiac cycle (RR intervals from ECG) and systolic blood pressure (SBP) from the finger arterial pressure waveform from a 25-year-old healthy woman in a supine position. For SBP (the upper panel), the median and Q1 and Q3 values are shown, while the mean and +/- SD values are displayed for the RR intervals.

significantly from the general trend, it is easy to spot them. Line plots can help make predictions or forecasts based on historical data. The line plots can be accompanied by shaded areas or error bars around the lines representing, for instance, the 95% confidence intervals and showing the dispersion of values around the trend lines.

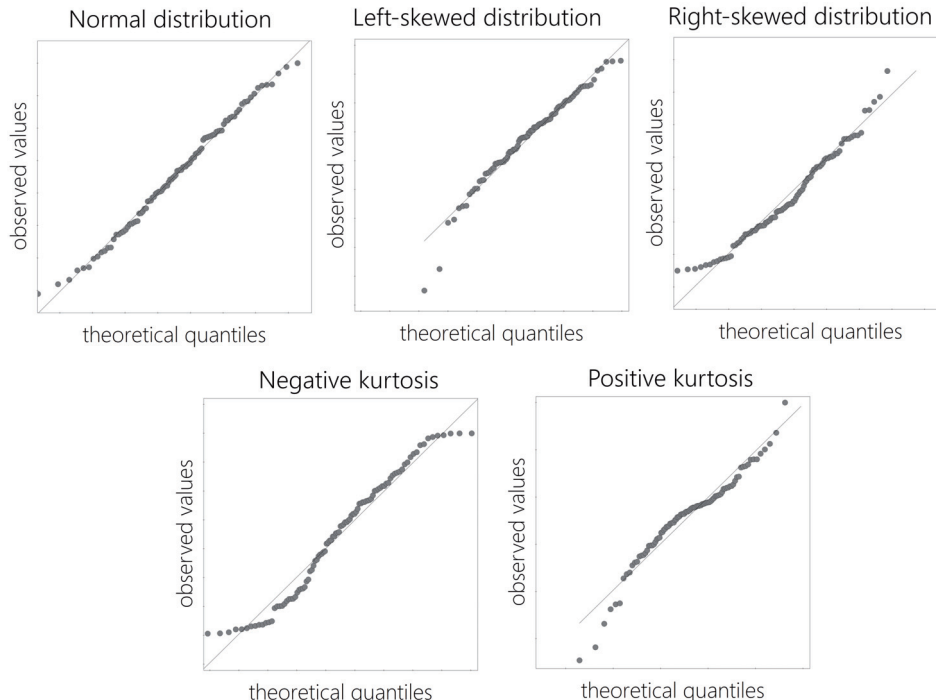
7. A Q-Q (quantile-quantile) plot is a graphical tool that examines whether a data set follows a particular theoretical distribution, such as a normal distribution. It compares the quantiles of the observed data with the quantiles of the expected theoretical distribution.

To construct a Q-Q plot, the values of the observed data set are first sorted in ascending order. Next, the corresponding quantiles of the expected distribution are calculated. These quantiles represent the values that would be expected if the observed data followed the specified distribution perfectly. The Q-Q plot then displays the observed quantiles on the x-axis and the expected quantiles on

the y-axis. Each data point represents a pair of observed and expected quantiles (see **Figure 8**).

If the observed data closely follow the expected distribution, the points on the plot will fall approximately on a straight identity line that follows the function  $x = y$ . This identity line assumes that the estimated points (on the y-axis) are the same as the observed points (on the x-axis). Departures from a straight line indicate deviations from the expected distribution.

These plots help to assess the normality assumption of a data set. It suggests that the data follows a normal distribution if the data points on the plot closely follow the identity line. However, if the points diverge from the line, this indicates deviations from normality, such as skewness or heavy tails. Q-Q plots can also be used to compare two sets of data. Plotting the quantiles of one data set against the quantiles of another makes it easy to see if the two data sets have similar distributions.



**Figure 8.** An example of a quantile-quantile (Q-Q) plot comparing quantiles representing the observations and their distribution with quantiles corresponding to the theoretical normal distribution. The points form a line along the identity line ( $y = x$ ) if both sets of quantiles come from the same distribution. Gaussian and other distributions such as uniform, exponential or Pareto can be compared using these plots. Q-Q plots are more diagnostic than comparing sample histograms, density plots, scatter plots, box-whisker plots or violin plots. With Q-Q plots, skewness and kurtosis are immediately visible. They are easy to examine. The multimodality of distributions can also be found. See an example in Figure 9.

Q-Q plots do not show measures of central tendency, but it is easier to see how the data deviate from normal distributions, whether skewed or have kurtosis.

## Standard numerical tests for normality testing

Many tests are used in medical research to analyse data distribution. The most common are:

1. *Kolmogorov-Smirnov test*. (1) This test determines whether a sample comes from a normal distribution. It compares the empirical data distribution with the cumulative distribution function of a theoretical normal distribution. Advantages: It is sensitive to differences in both location and shape between the sample and the normal distribution. Disadvantages: It is less powerful than other tests when the sample size is small.
2. *Shapiro-Wilk test*. (2) This test determines whether sample data come from a normal distribution based on the correlation between the observed data and the expected normal values. The test was originally proposed by Shapiro for small sample sizes. It is now used for data sets ranging from 3 to 5,000 samples. (3,4) Advantages: It is more potent than other tests when the sample size is small. Disadvantages: It is less powerful than other tests when the sample size is large.
3. *Shapiro-Francia test*. (5) This is similar but simpler than the Shapiro-Wilk test but has better power for small samples. It measures the deviation of the sample data from normality by comparing the sample distribution to a normal distribution with the same mean and variance. Advantages: It is more potent than other normality tests when the sample size is small and less sensitive to outliers than other normality tests. Apart from being less popular (not well known), there are no methodological disadvantages when used with small data sets.
4. *D'Agostino-Pearson (D'Agostino's K-squared) test*. (6,7) This test determines whether a sample comes from a normal distribution. It is based on the skewness and kurtosis of the sample as measures of deviation from normality. The D'Agostino-Pearson test provides a formal statistical test to support or challenge the visual assessment made, for example, by Q-Q plots. Both methods provide a more complete analysis of normality. Advantages: It is a powerful test of normality. Disadvantages: It may not be sensitive to forms of non-normality other than skewness and kurtosis, such as multi-modality or heavy tails.
5. *Anderson-Darling test*. (8) This parametric test uses the sample data to estimate the normal distribution parameters. The test statistic is based on the difference between the observed and expected cumulative distribution functions. Advantages: It is more powerful than other tests when the sample size is large. Disadvantages: It is less powerful than other tests when the sample size is small.
6. *Cramer-von Mises test*. (9) Similar to the Anderson-Darling test, but gives more weight to differences in the tails of the distributions. Advantages: It is a powerful test of normality for larger sample sizes. Disadvantages: It is sensitive to sample size.
7. *Jarque-Bera test*. (10) This test determines whether a sample comes from a normal distribution. It uses skewness and kurtosis as measures of deviation from normality. Advantages: It complements graphical methods such as Q-Q plots. Disadvantages: Its ability to identify certain types of non-normal distribution is limited as it primarily looks for deviations from the normal pattern based on skewness and kurtosis.
8. *Lilliefors test (Kolmogorov-Smirnov-Lilliefors test)*. (11,12) It is an extension of the Kolmogorov-Smirnov test but adjusted when the mean and variance of the data are also estimated. Advantages: It has better power than the original Kolmogorov-Smirnov test to detect deviations from normality. This is especially true for moderate sample sizes. Disadvantages: It can be overly conservative and not appropriate for small samples.
9. *Lobato-Velasco test*. (13) This test measures skewness and kurtosis and their correlation coefficients for observations. While assessing the normality of the data distribution, the Lobato-Velasco test provides consistent results for data that are correlated over time. Advantages: This test considers the specificity of dependent data and provides consistent results for data that are correlated over time.

Disadvantages: It is sensitive to deviations from the assumption of stationarity.

For smaller sample sizes < 50, it is advisable to employ the Shapiro-Wilk test or Shapiro-Franca test for their higher power in detecting deviations from normality in such cases. (9,14) If skewness or kurtosis are more important, the D'Agostino-Pearson or Jarque-Bera test and other tests focusing more on skewness and kurtosis work better. For sample sized > 50, other methods, particularly graphical like Q-Q plots, histograms, density plots, box-and-whiskers and other tests of normality can be used.

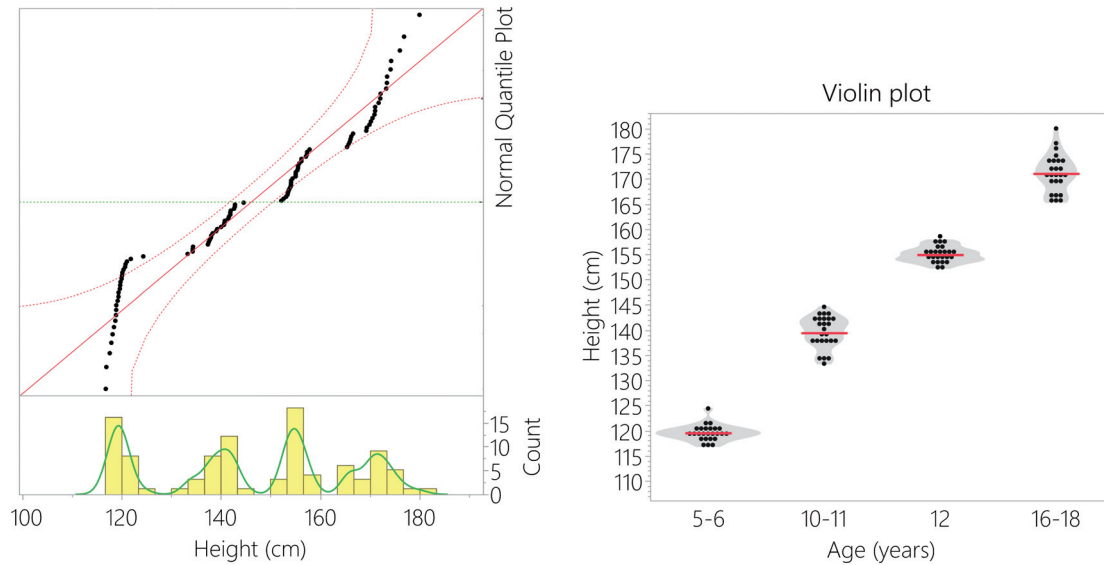
For very large samples, most normality tests are too sensitive and will detect even small deviations from normality. (15) It is advisable to use graphical tools to avoid prematurely labelling data as non-normal for small deviations that are unlikely to affect the interpretation of the data. These methods visualise the true distribution. They help to identify multimodality, asymmetry or excessive variance.

Multimodality is one of several reasons for the lack of normality in the distribution. In a multimodal distribution, clusters or subgroups of values are separated from each other. Statistical analysis and inference should take account of such clustered values and, where appropriate, apply specific tests for cluster analysis or multimodal modelling. These may facilitate under-

**Table 2.** Data distribution evaluation based on measures/coefficients, statistical tests and graphs.

Analysis type	Result of analysis	Result interpretation
Normality distribution tests	P-value < 0.05	Not a normal distribution
<b>Skewness assessment</b>		
Skewness coefficient	Positive (especially greater than 2*)	Right-skewed distribution
	Negative (especially less than -2*)	Left-skewed distribution
Tests for assessing skewness	P-value < 0.05	Skewed distribution
Histogram and density plots	Long right tail	Right-skewed distribution
	Long left tail	Left-skewed distribution
Box-whiskers, violin and dot plots	Extended top of the chart (an upper whisker)	Right-skewed distribution
	Extended lower part of the graph (a lower whisker)	Left-skewed distribution
Q-Q plot	Right and left tails significantly departing above the identity line	Right-skewed distribution
	Right and left tails significantly departing below the identity line	Left-skewed distribution
Mean versus Median	Mean distinctively above median	Right-skewed distribution
	Mean distinctively below median	Left-skewed distribution
<b>Kurtosis assessment</b>		
Kurtosis coefficient	Positive (especially greater than 4*)	Sample distribution is narrower than a normal distribution
	Negative (specifically less than -4*)	Sample distribution is flatter and wider than a normal distribution
Tests for assessing kurtosis	P-value < 0.05	Kurtosis atypical for normal distribution
Q-Q plot	Left tail above and right tail below the identity line	The distribution is more flattened than a normal distribution
	Left tail well below and right tail well above the fit line	Distribution is narrower than a normal distribution
Histograms and density plots	"Heavy" tails	The distribution is more flattened than a normal distribution
<b>Multimodality assessment</b>		
Histogram and density plot	Distinct clusters of bars (density) of similar height representing separate groups	Multimodality occurs
Q-Q plot	Multiple groups of points deviating in different directions from the fit line	
Violin and Point plot	Occurring in alternating wide and narrow shapes, separated clusters of multiple points represent different value groups with different centers	

\* Limits proposed for the district of significant deviation from normality (14).



**Figure 9.** An example of Q-Q plots, histograms and density plots (left panel) with the results of height measurements collected in a group of healthy children aged between 5 and 18 years. Four distinct peaks of lumped height values (local maxima) appear. An additional analysis (right panel) examining the height distribution against age explains that the four local maxima correspond to four different age groups of the children studied. Categorical, continuous and discrete data can all form multimodal distributions and can be analysed in this way.

standing of differences between two or more subgroups.

When the statistical mean is the primary measure describing the data, it is important to ensure that the data are normally distributed and not skewed. This is because the central limit theorem states that the sample mean of large samples approaches the true population mean. In other words, the means of such samples satisfy the normality of the distribution. However, this theorem does not define whether the data have a normal distribution. The theorem does not exempt the researcher from investigating how the data are distributed.

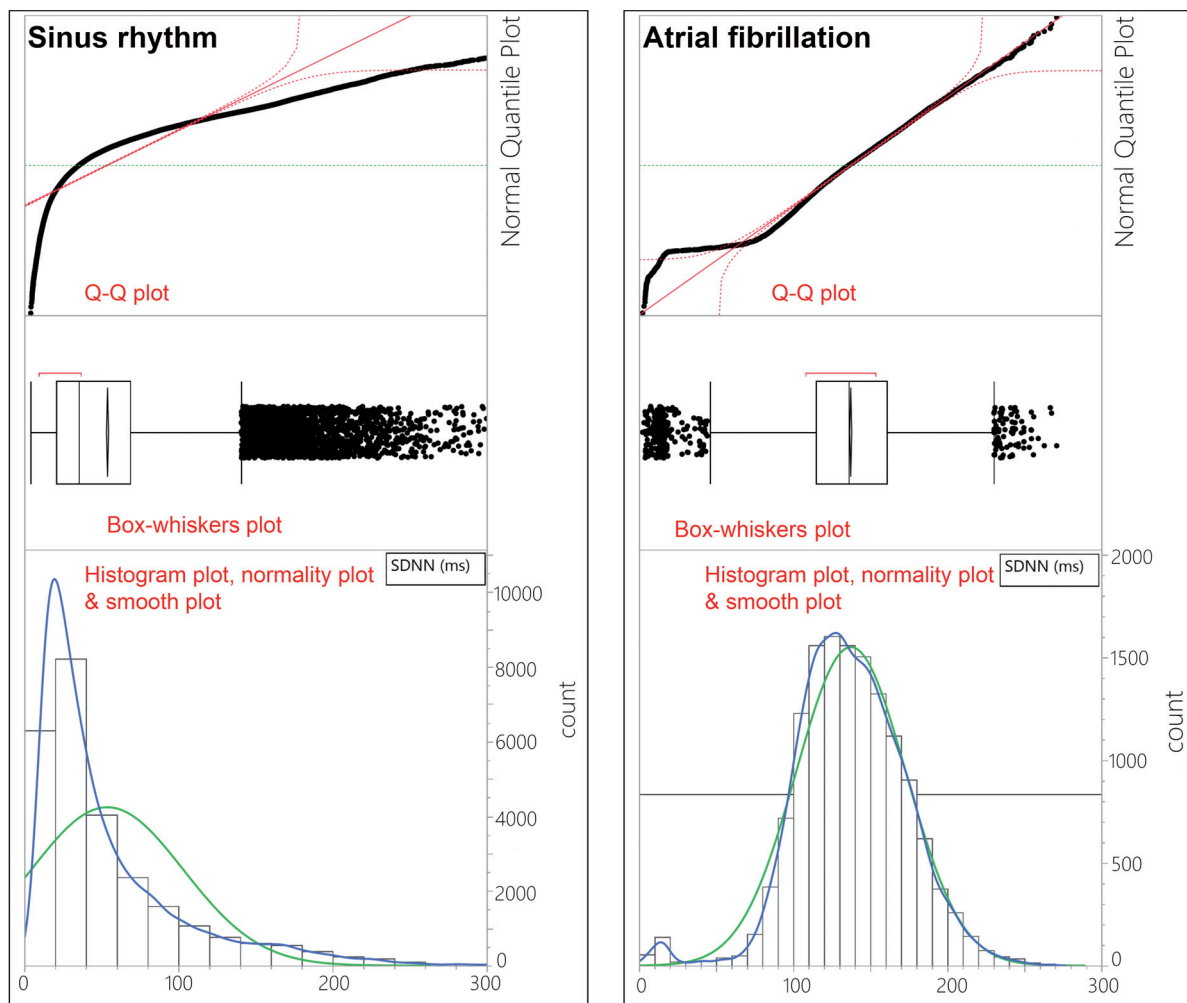
**Figure 10** shows over 20000 measurements of 1-minute total heart rate variability (SDNN) for sinus rhythm and for atrial fibrillation. In both cases, statistical tests detected significant ( $p$ -value < 0.0001) deviations from the normality of the distribution. However, only one graph shows data that are significantly out of normal distribution. For sinus rhythm, this is due to the high skewness of the data (mean 35.5 ms and median 53.8 ms). For atrial fibrillation, the distribution is less skewed (mean 137.5 ms and median 136.1 ms). Only on the left side is the proportion of observed values above the expected value slightly higher. The atrial fibrillation data can therefore be assumed to have a normal distribution.

With large data sets, the normality tests have too much power and may detect even minimal deviations from normality as significant. In such cases, graphical analysis is always essential and may be decisive.

However, only one graph shows data that are significantly out of normal distribution. For sinus rhythm, this is due to the high skewness of the data (mean 35.5 ms and median 53.8 ms). For atrial fibrillation, the distribution is less skewed (mean 137.5 ms and median 136.1 ms). Only on the left side is the proportion of observed values above the expected value slightly higher. The atrial fibrillation data can therefore be assumed to have a normal distribution.

## Discussion

Assessing the normality of a distribution is the first step in many statistical analyses. It should always start with a visual assessment, for example, using histograms or density plots. Unfortunately, due to the required time and uncertainty of interpreting such plots, statistical tests become the only tool for testing the normality of data distributions. Normality tests are central to statistical analysis. However, they should complement, not replace, graphical assessment of normality.



**Figure 10.** Q-Q plots, box and whisker plots, histograms and normal density plots showing analysis of SDNN calculated for 1 minute beat-to-beat values of RR interval duration. Left panel shows plots for normal sinus rhythm, right for AF. Each panel summarises the finding for more than 20,000 separate 1-minute files of RR intervals. For sinus rhythm, the data distribution is not normal, which can be seen in the Q-Q plot, box-whisker plot – greater distance between the median and Q3 and the right whisker, and a clear clustering of outliers outside this whisker. The histogram is also highly skewed. The distribution analysis of SDNN for AF appears to be Gaussian.

### Is normality testing necessary?

Normality tests aim to determine whether a data set is well-modelled by a normal distribution. A single normality test is usually sufficient. If the results are uncertain or borderline, other tests can be used to confirm or reject the normality of the distribution of the data being analysed. In such tests, the null hypothesis is that the distribution is normal, confirmed if the p-value exceeds 0.05. If  $p < 0.05$ , normality is rejected.

A normal distribution is symmetric, so data conforming to this distribution can be summarised with mean and SD and later analysed with parametric tests. True normality is considered a myth because real data, including medical data, usually deviate from the ideal normal dis-

tribution to some extent. For skewed non-normal data, mean and SD may be misleading and confusing because of potential over- or underestimation. The median, Q1 and Q3 are required for data with a non-Gaussian distribution. It is also convenient for readers to see both the mean and the median to decide whether the distribution is normal.

To date, statisticians have not reached a consensus on a single best test for assessing the normality of distribution for all possible data and situations. Normality tests with small group sizes often confirm a normal distribution, while tests with large groups tend to reject this assumption. Circumstances in which all tests agree in judging the normal distribution are straightforward. The

problem arises when different tests give different assessments of the distribution. What do you do when the statistical tests disagree with your assessment? This is another reason to return to graphical methods for assessing normal distribution. It is worth looking at the presence of outliers and whether errors are hidden among them, using additional graphical techniques such as Q-Q plots.

Assessing the distribution's normality should help select the best methods for further analysis. Choosing the right normality test can significantly impact the reliability and validity of the statistical analysis. Normality tests help determine whether parametric tests are appropriate for further statistical analysis. Parametric tests, such as t-test and ANOVA for comparisons or Pearson's correlation test and regression models based on least squares estimation, rely heavily on the normality assumption. Sample size estimation for design studies would not be possible without proper test selection.

Determining whether the data show a serious departure from normality is crucial. If there is any doubt about the normality of the data distribution, it is better to use non-parametric tests in further analyses. If the data are normally distributed in one subgroup but not in another, it is recommended that non-parametric tests be used for the subgroup that does not have normally distributed data.

Nonparametric tests do not assume that the data are normally distributed. Non-parametric methods should be used in further analyses for data that are not normally distributed. The simplest examples are the Mann-Whitney or Kruskal-Wallis tests for comparisons or the Spearman correlation test. They are more resistant to violations of this assumption. There are also robust statistical methods used in medical research to analyse data that may have outliers or other anomalies and to deal with such problems. (16) However, some statistical power is lost by using non-parametric tests rather than parametric tests. Alternatively, the data can be normalised by transforming them with some mathematical functions (e.g. logarithm, square root). Another solution is to treat the results as exploratory rather than conclusive.

Consistency in the presentation and interpretation of data is important, and the choice of

a particular approach should stand if the validity of the statistical method used has been established. Unwarranted changes from parametric to non-parametric tests or vice versa during the process may raise concerns about the reliability of the statistical analysis and affect the final result.

## Summary

Exploring clinical data is an integral part of medical research. One of the first steps is to distinguish whether the data is continuous, quasi-continuous or discrete. Since outliers of different origins can affect the final results, it is important to notice them and decide what to do about them. When analysing the data distribution, graphical and numerical methods should be used after adequately identifying whether the data have a normal distribution; non-parametric or parametric tests should be used in further analysis.

Reliable and correct statistical analysis is crucial in medical research for many reasons, including accurate data interpretation, findings validation, evidence-based decision-making, and generalisability of results. It underpins the credibility and impact of medical research, leading to advances in healthcare and improved patient outcomes.

## Acknowledgements

This paper has been prepared within the tasks of the Project "Development of the University Centre for Sports and Medical Studies in Poznan, Poland" (Number: NdS/544750/2021/2022) with Principal Investigator Prof. Przemysław Guzik. The Ministry of Education and Science, Warsaw, Poland, funded the Project within the "Science for Society" Programme.

## Conflict of interest statement

The authors declare no conflict of interest.

## Funding sources

There are no sources of funding to declare.

## References

1. Sulla ANK. Determinazione empirica di una legge didistribuzione. Giorn Dell'inst Ital Degli Att. 1933;4:89–91. doi: 10.12691/ajams-1-1-2.
2. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika*. 1965;52:591–611. doi: 10.2307/2333709.

3. Royston P. Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and computing*. 1992;2:117–119. doi: 10.1007/BF01891203.
4. Royston P. A Toolkit for Testing for Non-Normality in Complete and Censored Samples. *The Statistician*. 1993;42:37. doi: 10.2307/2348109.
5. Shapiro SS, Francia RS. An approximate analysis of variance test for normality. *Journal of the American statistical Association*. 1972;67:215–216. doi: 10.1080/01621459.1972.10481232.
6. D'Agostino R, Pearson ES. Tests for departure from normality. Empirical results for the distributions of  $b_2$  and  $\sqrt{b_1}$ . *Biometrika*. 1973;60:613–622. doi: 10.2307/2335012.
7. D'Agostino RB, Belanger A, D'Agostino Jr RB. A suggestion for using powerful and informative tests of normality. *The American Statistician*. 1990;44:316–321. doi: 10.2307/2684359.
8. Anderson TW, Darling DA. A test of goodness of fit. *Journal of the American statistical association*. 1954;49:765–769. doi: 10.2307/2281537.
9. Thode HC. *Testing For Normality*. New York: CRC Press; 2002.
10. Jarque CM, Bera AK. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*. 1980;6:255–259. doi: 10.1016/0165-1765(80)90024-5.
11. Lilliefors HW. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*. 1967;62:399–402. doi: 10.1080/01621459.1967.10482916.
12. Lilliefors HW. On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *Journal of the American Statistical Association*. 1969;64:387–389. doi: 10.1080/01621459.1969.10500983.
13. Lobato IN, Velasco C. A simple test of normality for time series. *Econ Theory* [Internet]. 2004 [cited 2023 Jun 8];20. doi: 10.1017/S0266466604204030.
14. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive Statistics and Normality Tests for Statistical Data. *Ann Card Anaesth*. 2019;22:67–72. doi: 10.4103/aca.ACA\_157\_18. Cited in : PMID: 30648682.
15. Demi R S. Comparison of Normality Tests in Terms of Sample Sizes under Different Skewness and Kurtosis Coefficients. *International Journal of Assessment Tools in Education*. 2022;9:397–409. doi: 10.21449/ijate.1101295.
16. Farcomeni A, Ventura L. An overview of robust methods in medical research. *Stat Methods Med Res*. 2012;21:111–133. doi: 10.1177/0962280210385865.