Gaussian Distribution, Normality Tests and Associated Constraints

(Gonzalo Bajo)

1. Gaussian distribution.

The Gaussian distribution, also known as the normal distribution, is one of the most basic statistical distributions used in probability theory and statistics. Applied to data analysis, it describes how information is distributed around an average or central value, so that it appears as a symmetrical bell-shaped distribution, and is particularly useful for modelling various natural and social phenomena. For example, population height or, in general, anthropometric data, and measurement errors have an approximately normal distribution.

A Gaussian distribution is characterised by its mean (μ) and standard deviation (σ). A normal distribution shows how most of the data analysed are concentrated around its mean value (μ), so that as we move away from this central value -dispersion or standard deviation (σ)- we observe less data clustered at the extremes. In other words, the frequency of the data is lower as we move away from the central value. It is identified by the shape of a symmetrical bell that peaks at the centre and extends smoothly and continuously on either side. The fact that it is defined by two parameters, mean and variance, makes it easy to interpret and control in statistical analysis, so that it is possible to understand how the data are concentrated around the mean value and what their dispersion is. Precisely, the shape of the bell is determined by the mean and the standard deviation.

An important and interesting property of the Gaussian distribution is to consider that approximately 68% of the data lies within the interval between the mean value and one standard deviation (+-). About 95% of the data lies within the interval between the mean and twice the value of the standard deviation.

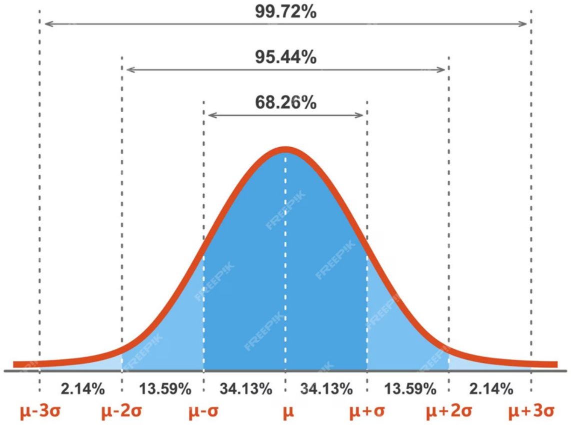

Figure 1. Gaussian or Normal distribution.

Gaussian distributions are important for statistical inference and hypothesis testing because many parametric statistical methods assume normality of the data distribution, so that confidence interval calculations, parameter estimations and statistical tests can be performed. Its mathematical and statistical properties make it a powerful tool for understanding and modelling the variability of observed data in various fields and performing normality tests. For example, applications can be found from scientific research to engineering and finance to quality control processes.

All stages of research in medicine, from data collection to evaluation of results, require the use of statistical methods. There are several ways to determine whether continuous data are normally distributed.

Generally, assumptions of normality can be assessed using both graphical and test methods. Graphical methods provide information on the shape of the distribution, but do not guarantee that the distribution is normal and are not able to test whether the difference between a normal distribution and a sampling distribution is significant. In addition, tests for normality can cause problems. Due to the small sample size, normality tests have little power to reject the null hypothesis that the data are from a normal distribution.

For this reason, a small sample size always passes the normality test. With larger sample sizes, small deviations from the normal distribution may be considered statistically significant even though they do not affect the results of the parametric test. Thus, the best way to determine if your data are normal is to evaluate the graph in conjunction with an appropriate normality test.

Normality tests are of importance in empirical and theoretical studies. The effectiveness of various parametric statistical inference methods depends on the underlying distributional assumptions. Some parametric statistical tests are able to prove that the distribution of data follows normality. However, under certain assumptions, these methods may not be able to determine a normal distribution of the data.

2. Normality tests

A normality test is a statistical tool used to assess whether a grouping of data follows a normal distribution. The normal distribution is a form of continuous probability distribution that has certain well-known characteristics, such as its bell-shaped form and its symmetry around the mean. There are different methods for testing whether a data set follows a normal distribution, and some common tests include:

- The Kolmogorov-Smirnov or K-S test, which compares the cumulative distribution of the data with the expected cumulative distribution for a normal distribution. It is used to determine whether the data in a sample are from a normal distribution. This test is applied in cases where the variables are quantitative and continuous, and when the sample size exceeds 50 cases.

This test compares the cumulative distribution of the data with the expected cumulative normal distribution and determines the P-value based on the most significant differences. For example, we have a sample of 500 employees, and we want to know whether the variable age follows a normal distribution. The null hypothesis (Ho) states that the sample comes from a normal distribution, while the alternative hypothesis (Ha) suggests that the data do not follow a normal probability model. Therefore, in order to accept Ho, the value of statistical significance (known as the p-value) must be greater than 0.05.

-Shapiro-Wilk or S-W test: This test is one of the best-known tests for diagnosing assumptions of normality and is based on the correlation between the given observations and the associated normal scores. It is considered suitable for small sample sizes of less than 50 cases. As it has a smaller or moderate sample size, it is more accurate compared to other tests. It is non-parametric, which means that it is not necessary to know the value of the parameters of the underlying normal distribution.

The S-W test has been recognised as the test of choice due to its remarkable power properties compared to a wide range of alternative tests.

- Anderson-Darling test: This test for normality is used to assess whether a sample comes from a population that has a specific distribution. It is considered an extension of the K-S test and sets additional weights to the tails of the distribution, which gives it greater sensitivity to deviations in those regions.

- Lilliefors test: This test is also a variant of the Kolmogorov-Smirnov test, which is applicable to small sample sizes. This test is not very useful in practice, since in most cases the value of the mean and standard deviation of the population is not known, so it is necessary to estimate these values for the theoretical comparison distribution. This makes the Kolmogorov-Smirnov test conservative, thus accepting the null hypothesis in most cases. In order to overcome this problem, the Lilliefors test tabulated the Kolmogorov-Smirnov statistic for the most common case where the population mean and variance are unknown and these values are estimated from the sample data.

These tests generate a p-value and a test value. The p-value indicates the probability of obtaining observed results if the data are normally distributed. If the p-value is greater than a pre-defined threshold (usually 0.05), the null hypothesis that the data follow a normal distribution is rejected.

It should be noted that the above tests may not be completely conclusive, especially when dealing with small sample sizes. Caution is needed when interpreting results, since a p-value of less than 0.05 does not necessarily guarantee that the data do not follow a normal distribution, and vice versa.

3. Limitations associated with normality testing

Although normality tests are useful tools, it is necessary to be aware that they can present a number of problems and limitations in specific cases. The following examples are associated with limitations of normality tests.

- Sensitivity to sample size: Different tests for normality may be sensitive to sample size. With large sample sizes, there is a greater chance of detecting small deviations from normality that may not be relevant in practice.

- Sensitivity to skewness: Normality tests may be affected by the presence of skewness in the data. Normality tests, in the presence of skewness, may reject the hypothesis of normality even though the distribution is approximately normal.

- Impact of outliers: Outliers can significantly alter the results of normality tests. Outliers can lead to erroneous rejection of the normality hypothesis.

- Context-related dependence: The interpretation of the results obtained from the normality test is subject to the purpose and context of the analysis. In certain cases, even if the data do not strictly correspond to a normal distribution, parametric methods can be robustly applied.

- P-value as unique criterion: The exclusive use of the p-value may lead to erroneous conclusions. A p-value close to the significance threshold should not be interpreted rigidly. It is important to consider the magnitude of the observed deviations.

- Use in large data sets: When using large data sets, normality tests can be very sensitive for the detection of small deviations from normality, which lack practical relevance.

In **conclusion**, normality tests are very useful as an exploratory tool because of their ease of application. However, some caution should be exercised in the interpretation of the data and consideration should be given to different aspects of the data set and statistical analysis. In different cases, the choice of reliable statistical techniques and consideration of the context may be more important in certain situations than strict compliance with the normality hypothesis.

Bibliography:

- Das, K., & Imon, A.H. (2016). A Brief Review of Tests for Normality. *American Journal of Theoretical and Applied Statistics, 5*, 5. DOI: 10.11648/j.ajtas.20160501.12
- Dogde, Y. (2008). Anderson–Darling Test. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_11
- Öztuna, D.; Elhan, A. H.; and Tüccar, Ersöz (2006) Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions, *Turkish Journal of Medical Sciences*: Vol. 36: No. 3, Article 7. https://journals.tubitak.gov.tr/medical/vol36/ iss3/7
- Porras Cerron, J. C. (2016). Comparación de Pruebas de Normalidad Multivariada. *Anales Científicos*, *77*(2), Pág. 141-146. https://doi.org/10.21704/ac.v77i2.483
- Romero Saldaña, M. (2016). Pruebas de bondad de ajuste a una distribución normal, Revista Enfermería del Trabajo**,** vol 6 nº 3, pág.114. Dialnet-TestsOfFitnessToANormalDistribution-5633043.pdf

Master Degree on Pharmacological Research
Gonzalo Bajo López