NORMALITY ASSUMPTION IN STATISTICAL DATA ANALYSIS

S.Ya. Shatskikh[1], L.E. Melkumova[2]

[1]Samara National Research University, Samara, Russia
[2]Mercury Development Russia, Samara, Russia

Abstract. The article is devoted to normality assumption in statistical data analysis. It gives a short historical review of the development of scientific views on the normal law and its applications. It also briefly covers normality tests and analyzes possible consequences of using the normality assumption incorrectly.

Keywords: normal law, normality assumption, normal distribution, Gaussian distribution, central limit theorem, normality tests.

Citation: Shatskikh SYa, Melkumova LE. Normality assumption in statistical data analysis. CEUR Workshop Proceedings, 2016; 1638: 763-768. DOI: 10.18287/1613-0073-2016-1638-763-768

“Mechanism" of the central limit theorem

Normal distribution can serve as a good approximation for processing observations if the random variable in question can be considered as a sum of a large number of independent random variables $X_1, \dots, X_n$, where each of the variables contributes to the common sum:

$$\operatorname*{Lim}_{n\to\infty} \mathbb{P}\left\{ \sum_{k=1}^{n} (X_k - m_k) \middle/ \left(\sum_{k=1}^{n} \sigma_k^2 \right)^{\frac{1}{2}} \le x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du,$$

$\mathbb{M}\{X_k\} = m_k,\ \mathbb{D}\{X_k\} = \sigma_k^2 < \infty.$

This theory was developed in the classical limit theorems, namely the De Moivre-Laplace theorem, the Lyapunov theorem and the Lindeberg-Feller theorem.

By the beginning of the 20th century this argument already had a rather wide support in practice. Many practical experiments confirmed that the observed distributions obey the normal law. The general acknowledgement of the normal law at that point of time was mainly based on observations and measurement results in physical sciences and in particular in astronomy and geodesy. In these studies measurement errors and atmospheric turbulence constituted the main source of randomness (Gauss, Laplace, Bessel).

In astronomy, where motion of celestial objects is dictated by the laws of classical mechanics it was possible to make measurements with high precision.

K. F. Gauss and P.S. Laplace were studying normal distribution in relation to their work on observation error theory (geodesy and astronomy). The woks of Gauss and Laplace had a great influence on the methods of observation data processing. During a long period of time it was believed that frequencies for a large number of observations have normal distribution, if enough number of accurate observations is removed from the sample.

In this respect a lot was known about measurements, errors and equations at that time.

Karl Pearson distributions

By the end of the 19th century normal distribution had lost its exclusive position. This was the result of many attempts to apply statistical methods to (mainly biological) research results. The distributions that came up in those studies were often asymmetrical or could have various other deviations from normality.

By that time Karl Pearson had suggested a system of 12 continuous distributions (in addition to the normal distribution), which could be used for smoothing empiric data. Today the discrete analogues of the Pearson-type distributions are also known.

Ronald Fisher to Egon Pearson polemic

However in the beginning of the 20th century the normal distribution has restored its value thanks to the thoughtful works of Ronald Fisher, who demonstrated that using the normality assumptions one can make conclusions of wide practical importance.

Nevertheless, after the R. Fisher’s book “Statistical methods for research workers” (1925) was published, Egon Pearson (Karl Pearson’s son) had made some critical remarks on if it’s justified to use the normality assumption in the statistical data analysis. According to E. Pearson, many of the tests in the Fisher’s book are based on the normality assumption for populations where the samples are taken from. But the question of the accuracy of the tests for the case when the population distributions depart from normal is never discussed. There is no clear statement, that the tests should be used with great caution in such situations.

Responding to the Pearson’s criticism, Fisher was stating his point of view, based on the statistical data that was obtained in experiments in the field of selection of agricultural plants.

Fisher believed that the biologists check their methods by using control experiments. So the normality assumption was tested by practice and not theory at that time.

By the time of this discussion some consequences of breaking the normality law were already known. Errors of this sort have slight effect on the conclusions about the mean values but can be dangerous for conclusions about the variance.

The last decades

By the end of the 20th century wide usage of statistical methods in biology, medicine, sociology and economics led the researchers to a conclusion that there is a wide variety of distributions that can be useful in these sciences. Aside from the normal distribution, distributions with “heavy” tails and asymmetric distributions took the stage.

This was caused by the fact that for many problems of these sciences the “mechanism” of the central limit theorem was problematic to establish. Also in contrast to physical sciences, one and the same experiment made with the same conditions can lead to different results.

For this reason the main cause of randomness (aside from the measurement errors) became the influence of various factors that were not taken into account and are interpreted as random.

This state of affairs led to a necessity to develop robust (to random deviations from given assumptions) methods of data analysis. Also there was a need for methods that don’t use the normality assumption, for instance methods of nonparametric statistics [7].

It’s worth noting, that in recent years these non-normal stable distributions became widely used in theoretical models of economics, financial mathematics and biology [8, 10].

It’s also worth noting that the stable non-normal Levy distribution was successfully used in the theory of laser cooling (Cohen-Tannoudji, Nobel prize in physics 1997). This theory uses the limit theorem of Levy - Gnedenko about convergence to stable non-Gaussian distributions [11].

Two quotations

J Tukey:

Today we can use the Gaussian shape of distribution in a variety of ways to our profit. We can:

a) use it freely as a reference standard, as a standard against which to assess the actual behavior of real data -- doing this by finding and looking at deviations.

b) use it, cautiously but frequently, as a crude approximation to the actual behavior, both of data itself and of quantities derived from data.

In using the Gaussian shape as such an approximation, we owe it to ourselves to keep in mind that real data will differ from the Gaussian shape in a variety of ways, so that treating the Gaussian case is only the beginning.[5]

Henri Poincaré:

There must be something mysterious in the normal law, since mathematicians think that this is the law of nature and physicists think this is a mathematical theorem.

Testing sample distributions for normality

Pearson's chi-squared normality test

Since the Gaussian distribution is continuous and it has two unknown parameters - mean and variance - when using the Pearson's test, the sample is usually divided into r classes and unknown values of the two parameters are replaced by their statistical estimates. As a result the limit distribution of the X^2 statistics will not be asymptotically equal to the chi-squared distribution with $r-3$ degrees of freedom. The distribution function of the X^2 statistics will lie lower, which means that the level of significance will be less than the nominal level. There are a few authors who claim that the chi-squared test is not a good choice for testing normality (see for example [9]).

Kolmogorov-Lilliefors test

Sometimes the Kolmogorov test, omega-squared test, chi-squared test can be used incorrectly to test normality of sample distribution. The Kolmogorov test is used to test the hypothesis that the sample is taken from a population with a known and completely defined continuous distribution function.

When testing normality of the distribution one can be unaware of the exact values of the mean and the variance. However it's well known that when the parameters of these distributions are replaced by their sample estimates, the normality assumption is accepted more often than it should be.

Besides in this case to test normality reliably one needs samples of large size (several hundred of observations). It's difficult to guarantee uniformity of observations for samples of this size [7].

Some recommendations for using the statistical tests can be found in [9]. Often it's appropriate to use the Lilliefors version of the Kolmogorov test.

Other normality tests

Starting the 30s many different distribution normality tests were developed. Some of the examples are: Cramer-von Mises test, Kolmogorov-Smirnov test, Anderson-Darling test, Shapiro-Wilk test, D'Agostino test and others (see [3]).

The Kolmogorov-Lilliefors and the Shapiro-Wilk normality tests are implemented in the ***Statistica*** and ***R*** software.

Consequences of breaking the normality assumption

The Student's **t**-statistics and the Fisher **F**-statistics relate to the case when the observed values have normal distribution and correlation between the observations is equal to zero. If (as it's usually the case) the distribution of the observations is not normal, then the distribution of the **t** and the **F** statistics differs from those, described above, especially for the **F**-statistics.

Comparing means of two samples

The most widely used test for comparing means of two samples with equal variance is based on Student's **t**-statistics. In this case the observations must be independent and with zero hypothesis must have equal normal distributions. When the distributions are not normal the level of significance of the **t**-test becomes almost accurate for the sample size more than 12.

Yet if variances of two samples are different, the Student's **t**-test will not give exact values for the levels of significance even for normal distributions (Behrens-Fisher problem, having no exact solutions today, only approximate).

Comparing variances of two samples

The test of equality of variances for two independent normal samples is based on the Fisher's **F**-statistics. The Fisher's test based on the **F**-statistics is very sensitive to deviations from normality.

Large samples

For large samples the law of large numbers and the central limit theorem "mechanism" both work. With corresponding norming applied the sample mean of the large number of observations will be close to the mean or will have a distribution close to normal, even if the observations themselves do not have normal distribution.

In this situation the means of a large number of the observation squares (Pearson's chi-squared statistics), as a rule, have almost chi-squared distributions.

We should keep in mind that proximity to the normal distribution and the chi-squared distribution depends on the sample size and the observation distributions.

Another example is related to the maximum likelihood estimates, which have many useful properties. However some of these properties hold only for very large samples. In real practice the samples are almost never very large.

Distribution of the Pearson's sample correlation coefficient r

Let ρ be the correlation coefficient of a couple of random variables X and Y:

$$\rho = \frac{\mathbb{M}\{(X - \mathbb{M}\{X\})(Y - \mathbb{M}\{Y\})\}}{\sqrt{\mathbb{D}\{X\}\mathbb{D}\{Y\}}},$$

and let r be the Pearson's correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

for a bivariate sample of observations of these variables.

For random variables X and Y with bivariate Gaussian distribution when $\rho \neq 0$ the distribution function and the density of the Pearson's correlation coefficient r can not be expressed via elementary functions but it can be represented using the hypergeometric function. For the case when $\rho = 0$, representations of the correlation coefficient density via elementary functions are known.

When $\rho = 0$ for large samples (the sample size $n \to \infty$) the Pearson's correlation coefficient r has asymptotically normal distribution.

However the convergence of the r coefficient to the normal distribution is too slow. It's not recommended to use normal approximation when $n < 500$. In this case the Fisher transformation for the r coefficient can be used. It leads to a new variable z, that has a distribution which is much more close to normal. Using this distribution it's possible to find confidence intervals for the r coefficient.

The research of the problem of sensitivity to deviations from the normal distribution of the r coefficient cannot be considered complete by this time. One of the reasons is that the distributions of r for non-normal samples are developed in detail for a relatively small number of certain cases. There are examples when the sensitivity of r to deviations from normality is high as well as examples when it's rather insignificant [2].

Acknowledgements

This work was partially supported by a grant of RFBR (project 16-01-00184 A)

References

1. Good P, Hardin J. Common errors in statistics, and how to avoid them. Wiley, 2003.
2. Johnson N, Kotz S, Balakrishnan N. Continuous univariate distribution. Vol. 2. 2ed, Wiley, 1995.
3. Kotz S, et al. Encyclopedia of statistical science, 16 volumes. 2ed, Wiley, 2005.
4. Lemann E. On the history and use of some standard statistical models. Probability and Statistics: Essays in Honor of D. Freedman. Vol. 2, 2008.
5. Tukey J. Exploratory data analysis. Addison Wesley, 1977.
6. Kobzar AI. Applied mathematical statistics. M.: FM, 2006. [In Russian]
7. Lagutin MB. Pictorial mathematical statistics. M.: BINOM, 2007. [In Russian]
8. Prokhorov YuV (ed.) Probability and Mathematical Statistics. Encyclopedia. M.: GRE, 1999. [In Russian]
9. Tyurin YuN, Makarov AA. Statistical data analysis on the computer. M.: INFRA–M, 1998. [In Russian]
10. Shiryaev AN. Essentials of stochastic finance: facts, models, theory, World Scientific, 1999.
11. Cohen-Tannoudji C. (et al.) Levy statistics and laser cooling. Cambridge university press, 2002.