GRANT journal
ISSN 1805-062X, 1805-0638 (online), ETTN 072-11-00002-09-4

EUROPEAN GRANT PROJECTS | RESULTS | RESEARCH & DEVELOPMENT | SCIENCE

Verification of normality as a basic precondition for the use of quality management tools

Kateřina Bícová[1]
Josef Sklenička[2]

[1] Západočeská univerzita v Plzni, Fakulta strojní; Univerzitní 22, 301 00 Plzeň; kbicova@kto.zcu.cz;
[2] Západočeská univerzita v Plzni, Fakulta strojní; Univerzitní 22, 301 00 Plzeň; sklenick@kto.zcu.cz

Grant: SGS-2019-008
Název grantu: Research and Development for Innovation in the Field of Manufacturing Technology - Machining Technology III.
Oborové zaměření: JS - Řízení spolehlivosti a kvality, zkušebnictví

Abstract The aim of this paper is to present the possibilities of verifying the assumption of a normal distribution of data for further statistical processing, without the need to study an inexhaustible number of methods and hypotheses about statistical data processing. The main idea is that, for example, companies in the automotive industry, where standards such as IATF 16949 require 100% control and the use of statistical tools for process monitoring, have easy guidance on how to verify relevant input data for further statistical processing. The normal distribution of data is one of the most common distributions that data has. At the same time, it is the most suitable for statistical tools, because it is possible to predict that the evaluated process will behave the same under the same input conditions. Without this verification, further data processing would not have sufficient explanatory power about the monitored parameter.

Key words normality, Gauss, automotive, IATF 16949, quality, statistics

1. INTRODUCTION

The correct use of the vast majority of statistical quality management tools, as well as a number of statistical hypotheses, is based on the fact that the probability distribution of data is known in advance. This means that the input data with which the analysis will be performed correspond to the given distribution, in our case normal. The normal distribution of data is one of the most common distributions that data has. At the same time, it is the most suitable for statistical tools, because it is possible to predict that the evaluated process will behave the same under the same input conditions.
The aim of this paper is to present the possibilities of verifying the assumption of a normal distribution of data for further statistical processing, without the need to study an inexhaustible number of methods and hypotheses about statistical data processing. MS Excel, which is one of the most widespread and well-known software supports, is mainly used for verification. The main idea is that, for example, companies in the automotive industry, where standards such as IATF 16949 require 100% control and the use of statistical tools for process monitoring, have easy guidance on how to verify relevant input data for further statistical processing. For example, for data processing using control diagrams and subsequent evaluation of process capability.

2. THE NORMAL DISTRIBUTION

The normal distribution or Gaussian distribution (according to Carl Friedrich Gauss) is one of the most important probability distributions of a continuous random variable. Random events occurring in nature or society can be well modelled by normal distribution. The normal distribution includes the often mentioned random errors, such as measurement errors, caused by a large number of unknown and mutually independent causes. Therefore, normal distribution is also referred to as the law of error. According to this law, the distribution of some physical and technical quantities is also theoretically governed. [1] [2]
The normal distribution is fully characterized by two constants: the mean value μ and the variance σ^2. The Gaussian curve is symmetric, the mean value of μ lies just below its peak. The shape of the curve with the extreme at the location of the mean value actually means that when repeating a random experiment following a normal distribution, the values around the mean value will most often come out. The symmetry of the curve then says that results deviated above and below the mean will be published about the same time. The parameter σ^2 determines how closely the curve fits the mean value; the lower this parameter, the "sharper" the graph. In practice, the so-called three sigma rule is often used, sometimes even two or one sigma. It holds that the result of a random experiment with the distribution $N(\mu, \sigma^2)$ lies in the interval [3]:

- $(\mu - \sigma, \mu + \sigma)$ with a probability of 68.27%,
- $(\mu - 2\sigma, \mu + 2\sigma)$ with a probability of 95.45%,
- $(\mu - 3\sigma, \mu + 3\sigma)$ with a probability of 99.73%.

Results near the mean value of μ are therefore more likely than outliers, see Fig. 1.

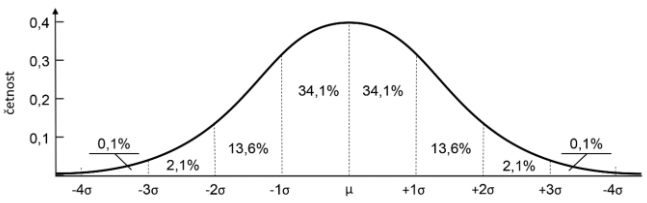

Fig. 1: The normal distribution (Gaussian curve) [3]

3. SELECTED METHODS FOR VERIFICATION OF NORMALITY

There are many methods to verify that the data corresponds to a normal probability distribution. These are numerical and graphical methods. Tests of the hypothesis that the random selection $x_1, x_2, \ldots, x_n$ comes from the assumed normal distribution are called goodness-of-fit tests.
Probably the best known graphic method is the histogram, which is a simple and fast tool. In addition, other simple graphical tools can be used, such as the *Q-Q* graph (quantile-quantile), which is slightly more accurate than the histogram and is more suitable for testing normality at distribution edges, or the *P-P* graph (probability-probability), which emphasizes deviations from normal distribution near the mean value.
As for numerical methods, there are a number of tests that vary in strength and complexity. These include, for example, Shapir-Wilk, Anderson-Darling, Kolmogorov-Smirnov, Lilliefors and others. The test is usually not performed manually, but due to the high complexity, the calculations are performed on a computer.
One graphical method and two numerical methods will be discussed in more detail for this paper. A histogram is chosen as a representative of graphic methods. Numerical methods χ^2 - goodness-of-fit test or Kolmogorov-Smirnov goodness-of-fit test with normal distribution are also selected. [4] [5] [9]

3.1 The histogram

The histogram is one of the basic tools of quality management. It is a graphical representation of the data using a bar graph with columns of the same width, expressing the width of the intervals, while the height of the columns expresses the frequency of the monitored quantity in the given interval. The histogram will help us assess the set of values in terms of data normality, symmetry, multimodality or the occurrence of outliers. Histograms are also a great way to view the results of running data. [6] [7] [8]
The following figures (nr.2 and 3) show the differences in display depending on the selection range. All these histograms represent random selections from the normal distribution with a mean value of $\mu = 30$ and a standard deviation of $\sigma = 3$. However, it can be seen that the larger the sampling range n, the better the selection distribution shown by the histogram corresponds to the distribution in the base set shown probability density. With the commonly used range $n = 100$, the visual assessment may not be objective and the shape of the histogram may be additionally influenced by the choice of interval limits. [5]

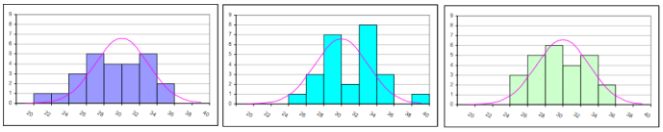

Fig. 2: Data with normal distribution, selection in the range n = 25, μ = 30 and σ = 3 [5]

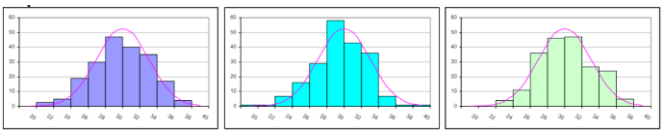

Fig. 3: Data with normal distribution, selection in the range n = 200, μ = 30 and σ = 3 [5]

3.2 The Pearson χ^2 - goodness-of-fit test

It is actually testing a statistical hypothesis, where the last step is to formulate the conclusion of testing, which can be done in two ways [2]:

a) by comparing the calculated test criterion with the critical value, which is determined depending on the selected level of significance α. If the value of the calculated test statistic exceeds the critical value, it means that there is evidence to reject the null hypothesis (ie "that the difference is confirmed"). Conversely, if the calculated test statistic finds itself within the domain of acceptance of the null hypothesis H_0, the null hypothesis does not have to be rejected and is therefore assumed to be valid. The agreement between the empirical and the theoretical distribution is assessed using the test criterion:

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}, \qquad (1)$$

where n_j are the empirical (real) frequencies in the interval j ($j = 1, 2, \ldots, k$) and n_{pj} are the theoretical frequencies (determined on the basis of probability) in the interval j. The formula of the test criterion can be easily adjusted to an equivalent form:

$$\chi^2 = \sum_{j=1}^{k} \frac{n_j^2}{np_j} - n \qquad (2)$$

During the validity of H_0, the statistics have asymptotically χ^2 - distribution of *k-c-1* degrees of freedom (c is the number of parameters that are not specified by H_0, so for a normal distribution 2)
The critical field for the H_0 test therefore has the form:

$$K = \{\chi^2 > \chi^2_{\alpha(k-c-1)}\} \qquad (3)$$

where $\chi^2_{\alpha(k-c-1)}$ is the critical value of χ^2 - distribution.
If $\chi^2 > \chi^2_{\alpha}$, the null hypothesis is rejected, the alternative hypothesis holds, which states that the random selection is not from a basic set with a given probability distribution. The reliability of the χ^2 - goodness - of - fit test increases with increasing range of selection n.

b) by converting the test statistic to a probability scale and calculating the probability p, which quantifies the probability of realizing the value of the test statistic, if the null hypothesis holds. So the rule for formulating a conclusion is as follows:
* If the p-value is less than the significance level α (error α), the null hypothesis H_0 is rejected. Symbolically, the conclusion can be used:
$p < 0.05$ "statistically significant difference" or
$p < 0.01$ "statistically highly significant difference".
* If the p-value is greater than the significance level α (error α), the null hypothesis H_0 cannot be rejected and it is therefore assumed that it holds. Symbolically it is possible to write: $p > 0.05$ ("statistically insignificant difference").

3.3 The Kolmogorov-Smirnov goodness-of-fit test with normal distribution

If the theoretical distribution is fully known, ie. its type and relevant parameters, is a very advantageous and simple test of conformity Kolmogorov-Smirnov test, which is applicable even in cases where χ^2 - goodness-of-fit test is not applicable (eg in case of small scale selection, large proportion of theoretical frequencies less than 5).
Its advantage is that it is based on the original individual observed values and not on data sorted into classes (groups). This prevents the information contained in the selection from being lost.
The test is used to verify the hypothesis that the selection obtained comes from a distribution with a continuous distribution function $F(x)$, which, however, must be fully specified, including all parameters. [5] [10]
The test is performed using the test criterion:

$$D = \frac{1}{n} max|N_j - H_j|, \qquad (4)$$

where Nj are the empirical cumulative frequencies, H_j the theoretical cumulative frequencies, n the frequency of the observed set and

$max|N_j - H_j|$ is the largest difference between cumulative empirical and theoretical frequencies. If the value of the test criterion D exceeds the critical value D_α found in the table for a given range of sample n and the chosen level of significance α, we reject the null hypothesis of agreement between the empirical and theoretical distribution. [10]

4. THE VERIFICATION OF NORMALITY

4.1 The histogram

The histogram was subsequently used for the numerical method, namely the Kolmogorov-Smirnov goodness-of-fit test with the normal distribution. The histogram is compiled so that the data are first divided into individual classes (intervals) of a specified width. The graph then shows the frequencies of values in individual classes. The following table lists the default values for histogram assembly.

Tab. 1: The occurrence of specific measurement values

Intervals	*Frequencies of values*
28,005	0
28,01	0
28,015	1
28,02	1
28,025	3
28,03	10
28,035	18
28,04	32
28,045	39
28,05	34
28,055	26
28,06	23
28,065	21
28,07	8
28,075	4
28,08	3
28,085	0
28,09	0

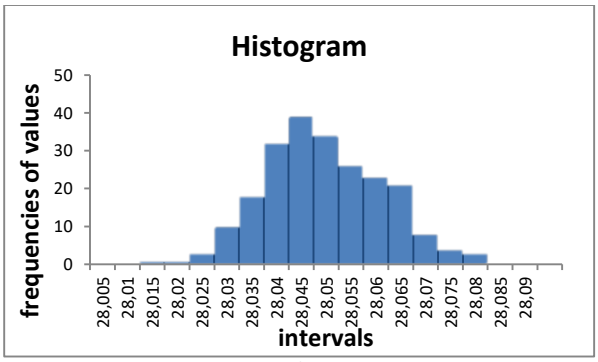

Fig. 4: Histogram of measurement values

From the previous figure we can conclude that this is really a concordance with the normal distribution, but it cannot be said unequivocally. Therefore, to better illustrate compliance, it is appropriate to use other tools or methods to confirm this.

4.2 The Pearson χ^2 - goodness-of-fit test

In this test, the calculated test criterion is compared with a critical value, which is determined depending on the selected level of significance α. If the value of the calculated test statistic exceeds the critical value, it means that there is evidence to reject the null hypothesis (ie "that the difference is confirmed"). Conversely, if the calculated test statistic finds itself within the scope of H_0 acceptance, the null hypothesis does not have to be rejected and is therefore assumed to be valid.

For the analysed data, for a 5% level of significance, the critical value is for χ^2crit = 6.244766 (from the tables for χ^2) [11]

Test criteria:

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j} = 0{,}155182574,$$

Since 0.155182574 < 6.244766, it follows that the null hypothesis holds = **it is an agreement with the normal distribution.**

It is a more extensive calculation, so it was performed in MS Excel [12]. In addition, the correctness of the calculation was verified in the Matlab program (see the following figure 5), where again the null hypothesis H_0 assumes that the sample has a distribution of a certain type, in this case normal.

```
>> [h,p,st] = chi2gof(x,'NBins',15,'Alpha',0.05)
h =
     0
p =
    0.1797
st =
    chi2stat: 10.1589
          df: 7
       edges: [28.0100 28.0260 28.0313 28.0367 28.0420 28.0473 28.0527 28.0580 28.0633 28.0687 28.0900]
           O: [7 16 27 44 32 28 29 19 12 10]
           E: [10.9775 13.6780 22.9458 32.1275 37.5450 36.6212 29.8139 20.2583 11.4890 8.5439]
```

Fig. 5: Verifying the null hypothesis in Matlab

H = 0 in the Matlab program means that the null hypothesis for the 5% level of significance is not rejected, ie, the hypothesis holds = **it is a coincidence with the normal distribution.**

4.3 The Kolmogorov-Smirnov goodness-of-fit test with normal distribution and histogram

Another possibility to verify normality is actually a combination of graphical and numerical methods. The graphical method in this case is a previously constructed histogram, which is more for the initial estimation of the shape of the data. Subsequently, the Kolmogorov-Smirnov test is used, in which the histogram is interpolated by a Gaussian curve. The value of criterion D is compared with the critical value D_α for the significance level $\alpha = 5\%$, ie *0.05.*

Tab. 2: Calculated values for the Kolmogorov-Smirnov test

Intervals	*Frequencies of values*
28.005	0
28.01	0
28.015	1
28.02	1
28.025	3
28.03	10
28.035	18
28.04	32
28.045	39
28.05	34
28.055	26
28.06	23
28.065	21
28.07	8
28.075	4
28.08	3
28.085	0
28.09	0

Calculated values	
Nr.of value =	224
x bar tot =	28.0466
s tot =	0.012449
Max =	28.09
Min =	28.01
Span =	0.08
Number int. =	15
Widht intervals =	0.005333
α =	0.05
D =	0.063711
D; crit. value. =	**0.090869**

Conclusion:
D < D; crit.value.
Not reject the normality

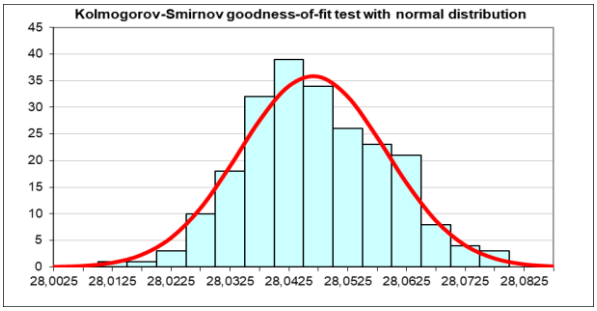

Fig. 6: Kolmogorov-Smirnov test - histogram with interpolated curve of normal probability density

From the previous figure nr.6 it is possible to compare the plotted Gaussian curve with the constructed histogram. The agreement with the normal distribution can therefore be stated not only from the point of view of graphical rendering, but also from the calculated values.
Numerically, the Kolmogorov-Smirnov test is expressed similarly to Pearson's χ^2 - goodness-of-fit test, by comparing the test criterion and the critical value. In this particular case, for a 5% significance level, the critical value is = 0.090869 (table - source [11]) and the calculated value = 0. 063711.
Since 0.063711 < 0.090869, it follows that the hypothesis holds = **it is a coincidence with the normal distribution.**
For an even better graphical representation of the match, it is possible to construct a distribution function. See the plot in Matlab (fig.7 and 8), where the empirical and theoretical distribution functions are compared.

```
>> cdfplot(x)
>> hold on
>> xx = [28 : 0.0005 : 28.1];
>> plot(xx, normcdf(xx,28.047,0.0124))
>> legend('Empiricka distr. fce',' Teoreticka distr. fce', 2)
>> title('Srovnani empiricke a teoreticke distribucni funkce')
>> ylabel('Fn(x),Φ(x)')
```

Fig. 7: Plotting a graph - comparison of empirical and theoretical distribution functions in Matlab

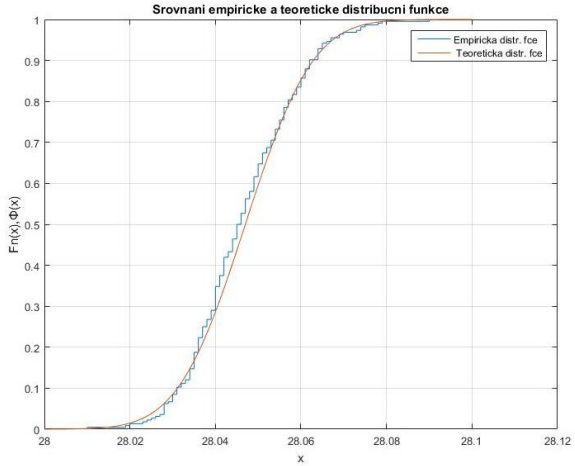

Fig. 8: Comparison of empirical and theoretical distribution functions in Matlab

When comparing the empirical and theoretical distribution functions, the agreement is evident.

5. THE CONCUSION

Each method and determination of indicators has its prerequisites for proper use. Therefore, the determination of indicators and further data processing is preceded by, for example, verification of normality, stability, etc.
The aim of this paper was to present the possibilities of verifying the assumption of a normal distribution of data for further statistical processing, without the need to study an inexhaustible number of methods and hypotheses about statistical data processing. The main idea was that, for example, companies in the automotive industry, where standards such as IATF 16949 require 100% control and the use of statistical tools for process monitoring, should have easy guidance on how to verify relevant input data for further statistical processing. For example, for processing using control diagrams and subsequent evaluation of process capability.
Normal distribution is a prerequisite for most other data processing tools. There are several methods for verification. These are numerical and graphical methods. This paper shows an example of using a simple graphical tool, namely a histogram. An important finding is that the analysed data, the distribution of which at first glance appears to be a normal distribution, is not always true. Therefore, numerical methods are used further.
Regarding numerical methods, two numerical methods were introduced here, namely χ^2 - goodness-of-fit test or Kolmogorov-Smirnov goodness-of-fit test with normal distribution. All presented methods were processed in MS Excel.
Overall, the shape of the curve characterizes the production or measurement process. So even on the basis of verifying the normality and evaluating the shape of the data, it is possible to draw conclusions about the properties of the data set or possible adverse effects on the process. In addition to the basic indicators, it is possible to determine other parameters such as accuracy, stability, bias and linearity for a more detailed evaluation of the data set, especially for the measurement system and for the production of skewness and sharpness.

Zdroje

1. Wikipedia, Normální rozdělení. [Online] [Cited: 28. 5 2021] https://cs.wikipedia.org/wiki/Norm%C3%A1ln%C3%AD_rozd%C4%9Blen%C3%AD
2. Bícová, K.: Příspěvek k hodnocení ukazatelů výrobního procesu v oblasti automobilového průmysl. Disertační práce, ZČU Plzeň 2016
3. WikiSkripta, Normální rozdělení. [Online] [Cited: 2. 6 2021] https://www.wikiskripta.eu/w/Norm%C3%A1ln%C3%AD_rozd%C4%9Blen%C3%AD
4. Papáková M.: Využití Chí kvadrát testů na příkladech experimentálních dat s využitím Geostatistical Analyst v softwaru ArcMap. Bakalářská práce. Olomouc. [Online] [Cited: 20. 5 2021] http://www.geoinformatics.upol.cz/dprace/bakalarske/papakova10/testy.html
5. Jarošová, E.; Král, J.: Ověřování předpokladu normality. Národní informační středisko pro podporu jakosti. 2006 [Online] [Cited: 21. 5 2021] http://www.csq.cz/fileadmin/user_upload/Spolkova_cinnost/Odborne_skupiny/Statisticke_metody/sborniky/2006/05_-_12_-_Testy_normality.pdf.
6. Statistické grafy. [Online] [Cited: 21. 5 2021] https://iastat.vse.cz/stat_grafy.html
7. Histograms. [Online] [Cited: 26. 5 2021] https://www.mathsisfun.com/data/histograms.html
8. Histogram. [Online] [Cited: 26. 5 2021] https://managementmania.com/cs/histogram
9. Testy normality. [Online] [Cited: 6. 6 2021] https://www.wikiskripta.eu/w/Testy_normality
10. Kolmogorovův-Smirnovův test. [Online] [Cited: 6. 6 2021] https://cs.wikipedia.org/wiki/Kolmogorov%C5%AFv%E2%80%93Smirnov%C5%AFv_test
11. Tabulky kritických hodnot a konstant. [Online] [Cited: 6. 6 2021] http://homel.vsb.cz/~dor028/Tabulky.pdf
12. Normality Test Using Microsoft Excel. [Online] [Cited: 6. 6 2021] https://www.inprolink.com/2019/02/20/normality-test-using-microsoft-excel