Comparing the test power for the normal distribution of the direct multiple measurements

N M Borbats and T V Shkolina¹

Department of Quality Management, Standardization and Metrology, Bryansk State Technical University, Bryansk, 241035, Russia

¹E-mail: shkolina.tv@yandex.ru

Abstract. An experimental study of the test power (Shapiro-Wilk W-test, Geary test and composite test) was carried out. The test power was used to check the normal distribution of a relatively small number of the multiple measurements. The study results confirm the high power of the W-test for detecting the deviation of the probability distribution from the normal law for a wide range of alternative distributions. A slightly better sensitivity of the Geary test for alternative distributions close to normal was also established. The results obtained can be used to justify the inclusion of the Shapiro-Wilk test in the next edition of the corresponding national standard.

1. Introduction
The procedure for processing the results of direct multiple measurements includes several operations, one of which is finding the confidence limits of the random error of the measurement result. At the same time, the corresponding national standard GOST R 8.736-2011 State system for ensuring the uniformity of measurements. Multiple direct measurements. Methods of measurement results processing. Basic principles) establishes a methodology for calculating confidence limits only for the measurement results belonging to a normal distribution.

According to the standard provisions, with the number of measurement results $15 < n \leq 50$, it is recommended to check whether they belong to the normal distribution using a composite test. With the number of measurement results $n \leq 15$, their belonging to the normal distribution is not checked, and using the method for determining the confidence limits of the random error established by the standard is possible only if it is known in advance that distributing measurement results is described by the normal law.

The composite test recommended by the standard, as the name suggests, consists of two separate tests, the first of which is the Geary test [1–3], and the second is based on the properties of the normal distribution (hereinafter N-test). The hypothesis about the normal distribution of a group of measurement results is rejected based on applying a composite test if at least one of its constituent criteria indicates its deviation. At the same time, it has long been known that one of the most powerful criteria designed to test the deviation of the probability distribution from the normal one with an amount of data $8 \leq n \leq 50$ is the Shapiro-Wilk W-test [4]. However, as part of a comparative study of the W-test with other well-known and often used in practice goodness-of-fit tests, the power of the Shapiro-Wilk test and Geary test, and even more the composite test, have not been compared [5]. In this regard,
it is of interest to compare the power of these tests to introduce possible changes in the provisions of the corresponding national standard.

2. Describing the compared tests
Let \( x_1, x_2, \ldots, x_n \) – be the results of individual observations in a group of multiple measurements (sample data) with the number \( n \) (sample size), and \( \bar{x} \) – be their arithmetic mean:

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

Then, to test the hypothesis \( H_0 \) about the normal distribution \( x_1, x_2, \ldots, x_n \) a certain function (statistic) is calculated from them and the fulfillment of the rejection criterion \( H_0 \) is checked. The type of statistics and the test for rejecting the null hypothesis depend on the type of the test used. This paper compares the power of the following tests.

2.1. Shapiro-Wilk W-test
The test is based on statistics of the form:

\[
W = \frac{U^2}{(n-1)S^2}
\]

where \( S^2 \) – is sampling variance:

\[
S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

\( U \) – is variability estimate calculated as:

\[
U = \sum_{i=1}^{k} a_{n,i} [x_{(n+1-i)} - x_{(i)}]
\]

where \( x_{(i)} \) – is \( i \)-th ordinal statistics; \( a_{n,i} \) – are coefficients determined from the look-up tables depending on the sample size \( n \) and index \( i, i = 1, 2, \ldots, k \), where

\[
k = \begin{cases} 
\frac{n}{2} & \text{if } n \text{ even} \\
\frac{(n-1)}{2} & \text{if } n \text{ odd}
\end{cases}
\]

The hypothesis \( H_0 \) about the normal distribution law is rejected at the level of significance \( \alpha \) if the condition is fulfilled

\[
w_n \leq w_{\alpha,n}
\]

where \( w_n \) – is the statistic value calculated by the formula (1); \( w_{\alpha,n} \) – is the percentage point selected from the look-up tables for a given significance level \( \alpha \) and the sample size \( n \).

2.2. Geary d-test
The test is based on statistics of the form:

\[
d = \frac{1}{ns} \sum_{i=1}^{n} |x_i - \bar{x}|
\]

where \( s \) – is biased estimate of the standard deviation, calculated by the formula:
The normal distribution hypothesis is not rejected if the condition is fulfilled:

\[ d_{\alpha/2,n} \leq d_n \leq d_{1-\alpha/2,n} \]  \hspace{1cm} (5)

where \( d_n \) is the statistic value calculated by the formula (4); \( d_{\alpha/2,n} \) and \( d_{1-\alpha/2,n} \) – are lower and upper percentage points, respectively, selected depending on the level of significance \( \alpha \) and the sample size \( n \).

2.3. N-test

The test is based on counting the number of differences \( m_n \) \( |x_i - \bar{x}| \), \( i = 1, 2, \ldots, n \), for which the condition is fulfilled:

\[ |x_i - \bar{x}| > z_{p/2} S \]  \hspace{1cm} (6)

where \( z_p \) is \( p \)-level quantile of the standard normal distribution; \( S \) is sample standard deviation, defined as \( S = \sqrt{S^2} \), where \( S^2 \) is sample variance calculated by the formula (2).

The hypothesis \( H_0 \) about the normal distribution law is rejected at the level of significance \( \alpha \) if the condition is fulfilled

\[ m_n > m \]  \hspace{1cm} (7)

where \( m \) is critical number of differences. The values \( m \) and \( P \) (in formula (6)) are found according to the reference tables depending on the accepted level of significance \( \alpha \) and the sample size \( n \).

2.4. Composite test (C-test)

The test is based on the combined use of the \( d \) and \( N \) test described in paragraphs 2.2 and 2.3, respectively. In this case, the hypothesis \( H_0 \) about the normal distribution law is rejected if it is rejected by at least one of these tests.

3. Research methodology

In fact, the study consisted of two parts. In the first part, the power of the Shapiro-Wilk W-test and the Geary d-test was compared in accordance with the methodology used in [5]. First, the Monte Carlo method obtained empirical zero distributions of statistics for each of the two tests for the sample sizes \( n = 11, 15, 20, 30, 40 \) and \( 50 \). Since both tests can be used to check complex hypotheses (that is, their statistics are invariant with respect to scaling and the beginning of the report), for each of the volumes \( n \), \( 10^6 \) random samples were generated from the standard normal distribution, the values of the test statistics were calculated using formulas (1) and (4), and their empirical cumulative distribution functions (cdf) were constructed. On the basis of the obtained empirical cdf, the values of the distribution quantiles for \( p \)-levels from 0.0025 to 0.9975 were determined. This sequence of actions was repeated five times, forming five replicas of the experiment. The arithmetic mean of five replicas was taken as the quantile final value of the \( p \)-level zero distribution of the corresponding statistics. The standard error of the quantile estimates obtained in this way did not exceed \( 3.10 \cdot 10^{-4} \).
Table 1. Comparing the quantiles known values of the W-test statistics zero distribution with their estimates.

| Sample size n | P-level | 0.01 | 0.05 | 0.10 | 0.50 | 0.90 | 0.95 | 0.99 |
|---------------|---------|------|------|------|------|------|------|------|
|               | w_p    | w_p  | w_p  | w_p  | w_p  | w_p  | w_p  | w_p  |
| 15            | 0.835  | 0.836 | 0.881 | 0.882 | 0.901 | 0.901 | 0.950 | 0.949 | 0.975 | 0.975 | 0.980 | 0.980 | 0.987 | 0.987 |
| 30            | 0.900  | 0.903 | 0.927 | 0.929 | 0.939 | 0.939 | 0.967 | 0.967 | 0.983 | 0.985 | 0.985 | 0.990 | 0.990 |
| 40            | 0.919  | 0.921 | 0.940 | 0.941 | 0.949 | 0.949 | 0.972 | 0.972 | 0.985 | 0.985 | 0.987 | 0.987 | 0.991 | 0.991 |
| 50            | 0.930  | 0.933 | 0.947 | 0.948 | 0.955 | 0.955 | 0.974 | 0.974 | 0.985 | 0.985 | 0.988 | 0.988 | 0.991 | 0.992 |

Table 1 shows a comparison of the known values $w_p$ of the distribution quantiles of the W-test statistics, found on the basis of approximating the Johnson curve $S_B$ [6] and the estimates $w_p^*$ obtained by the procedure described above for different $p$-levels and sample sizes $n$. The analysis of the table indicates a satisfactory accuracy of the values obtained.

The power of the tests was determined by their ability to detect deviations of data from the normal distribution if the alternative hypothesis was true. The types of the used alternative distributions with their parameter values are given in Table 2. Also, the table for each alternative distribution shows the values of the standardized third and fourth moments $\sqrt{\beta_3}$ and $\beta_4$, which are the characteristics of the distribution asymmetry and kurtosis, respectively [7].

Table 2. Types of distributing alternative hypotheses.

| Distribution (parameters) | Code designation | Parameter values | $\sqrt{\beta_3}$ | $\beta_4$ |
|---------------------------|------------------|------------------|-------------------|----------|
| Chi-square ($\nu$)        | $\chi^2(\nu)$    | $\nu = 2$        | 2.00              | 9.00     |
|                           |                  | $\nu = 10$       | 0.89              | 4.20     |
| Lognormal ($\mu, \sigma$)| LN                | $\mu = 0, \sigma = 1$ | 6.18              | 113.94   |
| Student’s ($\nu$)         | t($\nu$)         | $\nu = 1$        | 0                 | –        |
|                           |                  | $\nu = 4$        | 0                 | –        |
|                           |                  | $\nu = 10$       | 0                 | 4.00     |
| Beta ($a, b$)             | BE ($a, b$)      | $a = 1, b = 1$   | 0                 | 1.80     |
|                           |                  | $a = 2, b = 2$   | 0                 | 2.14     |
|                           |                  | $a = 2, b = 1$   | -0.57             | 2.40     |
|                           |                  | $a = 3, b = 2$   | -0.29             | 2.36     |
| Weibull ($a, b$)          | WE               | $a = 1, b = 0.5$ | 6.62              | 87.72    |
| Logistic ($\mu, \sigma$) | L                | $\mu = 0, \sigma = 0.5$ | 0                 | 4.20     |

To obtain the power estimates of the compared tests for each distribution indicated in table 2, $10^6$ random samples of the above volumes $n$ were generated using the MATLAB random number generator. For each sample, the values of statistics were calculated, after which their alternative cumulative distribution functions were constructed and the values of these functions were determined for the corresponding values of the zero distribution quantiles. That is, if $x_p$ – is the quantile of the corresponding zero distribution of the $p$-level statistics, and $F_n^+(x)$ – is the alternative cdf statistics, then a value $F_n^+(x_p)$ was found for each alternative distribution.

Since the W-test uses the left-sided critical region, the power of the test at the significance $p$-level for a given alternative distribution was estimated as:

$$1 - \beta = F_n^+(x_p)$$
In turn, since the Geary test uses a two-sided critical region, the power estimate at the significance level for a given alternative was estimated as:

$$1 - \beta = F_{n}^{*} \left(x_{p/2}\right) + 1 - F_{n}^{*} \left(x_{1-p/2}\right)$$

The described procedure was repeated five times and the arithmetic mean of five replicas was taken as the power estimate. The standard error of the estimates did not exceed $4.22 \times 10^{-4}$.

In the second part of the study, the power of all four tests for alternative distributions from table 2 was compared at a significance level of 0.05 (which is often used in practice by default). For this purpose, $10^6$ random samples of size $n = 11, 15, 20, 30, 40, \text{ and } 49$ were generated from each alternative distribution. For each sample, the values of the test statistics were calculated and fulfilling the conditions set by equations (3), (5) and (7) was carried out. The power of each test was determined as the ratio of the number of hypothesis deviations $H_0$ according to this criterion to the total number of the tests performed. The arithmetic mean of five replicas was taken as the power estimate. The maximum standard error of the estimates did not exceed $1.68 \times 10^{-4}$.

4. Results of comparing the test power

This section provides a brief description of the power comparison results for the tests under consideration. First, the results of comparing the power of the Shapiro-Wilk W-test and the Geary test for different levels of significance are presented, then the results of comparing the power of all four tests at a fixed level of significance are given.

4.1. Comparing the power of the W-test and the Geary test

Some of the results obtained are shown in the form of the power dependence curves $1 - \beta$ on the level of significance $p$ for individual alternative distributions with the volume of observations $n = 11$ in figure 1 and with the volume of observations $n = 30$ in figure 2.

![Figure 1](attachment:image.png)

**Figure 1.** Power curves of the W-test (1) and the Geary test (2) for some alternative distributions for the volume of observations $n = 11$. 

It is convenient to analyze the obtained results of a comparative study of the considered test power by dividing all alternative distributions into five groups depending on the values of the characteristics $\sqrt{\beta_1}$ and $\beta_2$ [6]:

- Asymmetric distributions with large kurtosis ($\sqrt{\beta_1} > 0.3, \beta_2 > 3.0$).
- Asymmetric distributions with small kurtosis ($\sqrt{\beta_1} > 0.3, \beta_2 < 3.0$).
- Symmetric distributions with large kurtosis ($\sqrt{\beta_1} \leq 0.3, \beta_2 > 4.5$).
- Symmetric distributions with small kurtosis ($\sqrt{\beta_1} \leq 0.3, \beta_2 < 2.5$).
- Symmetric distributions close to normal ($\sqrt{\beta_1} \leq 0.3, 2.5 \leq \beta_2 \leq 4.5$).

4.1.1. **Asymmetric distributions with large kurtosis.** This group includes such distributions as $\chi^2(2), \chi^2(10)$, LN and WE, for which the $W$-test demonstrates better power compared to the Geary test. So, for a significance level of 10% and a data volume of $n = 11$, the power of the $W$-test compared to the Geary test is, respectively, 62% versus 20% for $\chi^2(2)$, 75% versus 31% for LN (figure 1), 96% versus 41% for WE and 21% versus 12% for $\chi^2(10)$. With an increase in the number of observations, the power of both tests rises, but the $W$-test still demonstrates overwhelming superiority.
4.1.2. Asymmetric distributions with small kurtosis. The distributions of this group include BE (2, 1) and BE (3, 2), and for them the W-test also demonstrates the best power. For example, 10% power for \( n = 30 \) is: 71% versus 25% for BE (2, 1) (figure 2) and 23% versus 17% for BE (3, 2).

4.1.3. Symmetric distributions with large kurtosis. The distributions of this group include \( t (1) \) and \( t (4) \). Both tests demonstrate approximately the same power for this group distributions, and with an increase in the sample size, the sensitivity of the Geary test becomes slightly higher than that of the W-test. So, 10% power with the data volume \( n = 11 \) for \( t (4) \) is 22% for the W-test and 19% for the Geary test (figure 1), while for the data volume \( n = 30 \) it is 42% for the W-test and 48% for the Geary test (figure 2).

4.1.4. Symmetric distributions with small kurtosis. This group includes distributions BE (1, 1) and BE (2, 2). The power of both tests is practically the same for small values of \( n \) and, at the same time, is not great. However, as the volume of the data grows, the power of both tests increases and the W-test becomes more sensitive than the Geary test. For example, with the alternative BE (1, 1) and 10% power with the data volume \( n = 11 \), the power of both tests is 19%, and with \( n = 30 \) the power of the W-test is already 63% versus 52% for the Geary test.

4.1.5. Symmetric distributions close to normal. This group includes the distributions \( t (10) \) and \( L \). For the distributions of this group, both tests demonstrate almost the same power, which is rather low for small amounts of data and only slightly rises with increasing \( n \). So for the distribution \( L \) with \( n = 30 \), the power of the W-test is 20% versus 21% for the Geary test (figure 2).

4.1.6. General conclusions based on the comparison results. The results obtained once again confirm the high sensitivity of the W-test in relation to alternative distributions of various types. At the same time, the Geary test demonstrates lower power against the alternatives with the pronounced asymmetry, but it turns out to be somewhat more sensitive to the symmetric distributions with kurtosis other than normal, which is consistent with the remarks made by E.S. Pearson [2].

4.2. Comparing the power of four tests at a fixed significance level

The results of evaluating the power of the Shapiro-Wilk W-test, Geary d-test, N-test and composite C-test for different alternative distributions at a significance level of \( \alpha = 0.05 \) and a different number of observations \( n \) are shown in table 3.

| Distribution | \( n = 15 \) | \( n = 30 \) | \( n = 40 \) | \( n = 49 \) |
|--------------|-------------|-------------|-------------|-------------|
| \( \chi^2 (2) \) | 67 24 0 24 | 97 35 1 35 | 100 40 6 43 | 100 45 18 53 |
| \( \chi^2 (10) \) | 18 13 0 13 | 35 16 0 16 | 47 17 1 17 | 56 18 4 20 |
| LN | 82 40 0 40 | 99 64 1 64 | 100 74 10 76 | 100 81 23 85 |
| \( t (1) \) | 76 78 2 78 | 95 97 3 97 | 98 99 12 99 | 99 100 21 100 |
| \( t (4) \) | 19 24 0 24 | 30 40 0 40 | 34 48 4 49 | 37 55 11 57 |
| \( t (10) \) | 8 13 0 13 | 10 17 0 17 | 11 19 1 19 | 11 21 5 22 |
| BE (1, 1) | 13 24 0 24 | 41 51 0 51 | 69 65 0 65 | 85 75 0 75 |
| BE (2, 1) | 13 24 0 24 | 41 51 0 51 | 69 65 0 65 | 85 75 0 75 |
| BE (3, 2) | 13 24 0 24 | 41 51 0 51 | 69 65 0 65 | 85 75 0 75 |
| WE | 99 54 1 54 | 100 81 3 82 | 100 90 14 91 | 100 94 28 96 |
| L | 10 14 0 14 | 13 20 0 20 | 13 24 2 24 | 13 27 6 29 |

From the analysis of table 2, it can be seen that at the significance level of \( \alpha = 0.05 \), the W-test turns out to be the most powerful for the majority of alternative distributions practically over the entire range of \( n \) data values. Exceptions are symmetric distributions close to normal (such as \( t (10) \) and \( L \)), for which the power of all tests is sufficiently small, but the Geary test is somewhat more sensitive. You can also
notice that the power of the N-test does not exceed 28% for the distributions with the pronounced asymmetry and large kurtosis. In this case, contributing the N-criterion to the composite test power turns out to be relatively small, and the sensitivity of the latter is almost completely determined by the Geary test.

5. Conclusions
The results of the study confirmed the high power of the Shapiro-Wilk W-test for a wide range of alternative distributions, even using a fairly small amount of data. However, the Geary test showed slightly better sensitivity to alternatives that might be of the greatest interest in the practice of multiple measurements, namely, to the distributions close to normal. At the same time, the N-test turned out to be sensitive only to distributions with strong asymmetry and with a moderately large amount of data, and its contribution to the total power of the composite test proved to be small.

Thus, it can be recommended in the next revision of the corresponding national standard, if there is no possibility to replace the composite test, then at least to include in it a recommendation on using the Shapiro-Wilk W-test, especially in the situations where there are suspicions about the asymmetric distribution of the multiple measurements.

References
[1] Geary R C 1935 The ratio of mean deviation to the standard deviation as a test of normality Biometrika 27 310-32
[2] Geary R C 1936 Moments of the ratio of the mean deviation to standard deviation for normal samples Biometrika 28 295-307
[3] Geary R C 1947 Testing for Normality Biometrika 34 209-42
[4] Shapiro S S and Wilk M B 1965 An analysis of variance test for normality (complete samples) Biometrika 52 591-611
[5] Shapiro S S, Wilk M B and Chen H J 1968 A comparative study of various tests for normality J. Amer. Stat. Assoc. 63 1343-72
[6] Shapiro S S and Wilk M B 1968 Approximations for the null distribution of the W statistic Technometrics 10 861-6
[7] Johnson N L, Kotz S and Balakrishnan N 1995 Continuous Univariate Distributions 2 (New York: Wiley)