Lilliefors/Van Soest's test of normality

Hervé Abdi[1]   &   Paul Molin

1   Overview

The normality assumption is at the core of a majority of standard statistical procedures, and it is important to be able to test this assumption. In addition, showing that a sample does not come from a normally distributed population is sometimes of importance *per se*. Among the many procedures used to test this assumption, one of the most well-known is a modification of the Kolomogorov-Smirnov test of goodness of fit, generally referred to as the *Lilliefors test for normality* (or Lilliefors test, for short). This test was developed independently by Lilliefors (1967) and by Van Soest (1967). The null hypothesis for this test is that the error is normally distributed (*i.e.,* there is no difference between the observed distribution of the error and a normal distribution). The alternative hypothesis is that the error is not normally distributed.

Like most statistical tests, this test of normality defines a criterion and gives its sampling distribution. When the probability associated with the criterion is smaller than a given α-level, the

alternative hypothesis is accepted (*i.e.,* we conclude that the sample does *not* come from a normal distribution). An interesting peculiarity of the Lilliefors' test is the technique used to derive the sampling distribution of the criterion. In general, mathematical statisticians derive the sampling distribution of the criterion using analytical techniques. However in this case, this approach fails and consequently, Lilliefors decided to calculate an approximation of the sampling distribution by using the Monte-Carlo technique. Essentially, the procedure consists of extracting a large number of samples from a Normal Population and computing the value of the criterion for each of these samples. The empirical distribution of the values of the criterion gives an approximation of the sampling distribution of the criterion under the null hypothesis.

Specifically, both Lilliefors and Van Soest used, for each sample size chosen, 1000 random samples derived from a standardized normal distribution to approximate the sampling distribution of a Kolmogorov-Smirnov criterion of goodness of fit. The critical values given by Lilliefors and Van Soest are quite similar, the relative error being of the order of 10^{-2}.

According to Lilliefors (1967) this test of normality is more powerful than others procedures for a wide range of nonnormal conditions. Dagnelie (1968) indicated, in addition, that the critical values reported by Lilliefors can be approximated by an analytical formula. Such a formula facilitates writing computer routines because it eliminates the risk of creating errors when keying in the values of the table. Recently, Molin and Abdi (1998), refined the approximation given by Dagnelie and computed new tables using a larger number of runs (*i.e., K* = 100,000) in their simulations.

2   Notation

The sample for the test is made of N scores, each of them denoted X_i. The sample mean is denoted M_X is computed as

$$M_X = \frac{1}{N}\sum_i^N X_i,\qquad(1)$$

2

the sample variance is denoted

$$S_X^2 = \frac{\sum\limits_{i}^{N} (X_i - M_X)^2}{N-1} \, , \tag{2}$$

and the standard deviation of the sample denoted S_X is equal to the square root of the sample variance.

The first step of the test is to transform each of the X_i scores into Z-scores as follows:

$$Z_i = \frac{X_i - M_X}{S_X} \, . \tag{3}$$

For each Z_i-score we compute the proportion of score smaller or equal to its value: This is called the *frequency associated* with this score and it is denoted $\mathscr{S}(Z_i)$. For each Z_i-score we also compute the probability associated with this score if is comes from a "standard" normal distribution with a mean of 0 and a standard deviation of 1. We denote this probability by $\mathscr{N}(Z_i)$, and it is equal to

$$\mathscr{N}(Z_i) = \int_{-\infty}^{Z_i} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}Z_i^2\right\} \, . \tag{4}$$

The criterion for the Lilliefors' test is denoted L. It is calculated from the Z-scores, and it is equal to

$$\mathsf{L} = \max_i \{|\mathscr{S}(Z_i) - \mathscr{N}(Z_i)|, |\mathscr{S}(Z_i) - \mathscr{N}(Z_{i-1})|\} \, . \tag{5}$$

So L is the absolute value of the biggest split between the probability associated to Z_i when Z_i is normally distributed, and the frequencies actually observed. The term $|\mathscr{S}(Z_i) - \mathscr{N}(Z_{i-1})|$ is needed to take into account that, because the empirical distribution is discrete, the maximum absolute difference can occur at either endpoints of the empirical distribution.

The critical values are given by Table 2. $\mathsf{L}_{\text{critical}}$ is the critical value. The Null hypothesis is rejected when the L criterion is greater than or equal to the critical value $\mathsf{L}_{\text{critical}}$.

3   Numerical example

As an illustration, we will look at an analysis of variance example for which we want to test the so-called "normality assumption" that states that the within group deviations (*i.e.,* the "residuals") are normally distributed. The data are from Abdi (1987, p. 93*ff.*) and correspond to memory scores obtained by 20 subjects who were assigned to one of 4 experimental groups (hence 5 subjects *per* group). The score of the *s*th subject in the *a*th group is denoted $Y_{a,s}$, and the mean of each group is denoted $M_{a.}$. The within-group mean square $MS_{S(A)}$ is equal to 2.35, it correspond to the best estimation of the population error variance.

	G. 1	G. 2	G. 3	G. 4
	3	5	2	5
	3	9	4	4
	2	8	5	3
	4	4	4	5
	3	9	1	4
$Y_{a.}$	15	35	16	21
$M_{a.}$	3	7	3.2	4.2

The Normality assumption states that the error is normally distributed. In the analysis of variance framework, the error corresponds to the *residuals* which are equal to the deviations of the scores to the mean of their group. So in order to test the normality assumption for the analysis of variance, the first step is to compute the residuals from the scores. We denote X_i the residual corresponding to the *i*th observation (with *i* going from 1 to 20). The residuals are given in the following table:

Y_{as}	3	3	2	4	3	5	9	8	4	9
X_i	0	0	−1	1	0	−2	2	1	−3	2

Y_{as}	2	4	5	4	1	5	4	3	5	4
X_i	−1.2	.8	1.8	.8	−2.2	.8	−.2	−1.2	.8	−.2

Next we transform the X_i values into Z_i values using the following formula:

$$Z_i = \frac{X_i}{\sqrt{MS_{S(A)}}} \qquad (6)$$

because $MS_{S(A)}$ is the best estimate of the population variance, and the mean of X_i is zero. Then, for each Z_i value, the frequency associated with $\mathscr{S}(Z_i)$ and the probability associated with Z_i under the Normality condition $\mathscr{N}(Z_i)$ are computed [we use a table of the Normal Distribution to obtain $\mathscr{N}(Z_i)$]. The results are presented in Table 1.

The value of the criterion is (see Table 1)

$$\mathsf{L} = \max_i \{|\mathscr{S}(Z_i) - \mathscr{N}(Z_i)|, |\mathscr{S}(Z_i) - \mathscr{N}(Z_{i-1})|\} = .250 . \qquad (7)$$

Taking an α level of $\alpha = .05$, with $N = 20$, we find (from Table 2) that the critical value is equal $\mathsf{L}_{\text{critical}} = .192$. Because L is larger than $\mathsf{L}_{\text{critical}}$, the null hypothesis is rejected and we conclude that the residuals in our experiment are not distributed normally.

4   Numerical approximation

The available tables for the Lilliefors' test of normality typically report the critical values for a small set of alpha values. For example, the present table reports the critical values for

$$\alpha = [.20, .15, .10, .05, .01].$$

These values correspond to the alpha values used for most tests involving only one null hypothesis, as this was the standard procedure in the late sixties. The current statistical practice, however, favors multiple tests (maybe as a consequence of the availability of statistical packages). Because using multiple tests increases the overall Type I error (*i.e.,* the Familywise Type I error or α_{PF}), it has become customary to recommend testing each hypothesis with a corrected α level (*i.e.,* the Type I error *per* comparison, or α_{PC}) such as the Bonferonni or Šidák corrections. For example, using a Bonferonni approach with a familywise value of $\alpha_{PF} = .05$, and

Table 1: How to compute the criterion for the Lilliefors' test of normality. N_i stands for the absolute frequency of a given value of X_i, F_i stands for the absolute frequency associated with a given value of X_i (*i.e.,* the number of scores smaller or equal to X_i), Z_i is the Z-score corresponding to X_i, $\mathscr{S}(Z_i)$ is the proportion of scores smaller than Z_i, $\mathscr{N}(Z_i)$ is the probability associated with Z_i for the standard normal distribution, $D_0 = |\, \mathscr{S}(Z_i) - \mathscr{N}(Z_i)\,|$, $D_{-1} = |\, \mathscr{S}(Z_i) - \mathscr{N}(Z_{i-1})\,|$, and max is the maximum of $\{D_0, D_{-1}\}$. The value of the criterion is L = .250.

X_i	N_i	F_i	Z_i	$\mathscr{S}(Z_i)$	$\mathscr{N}(Z_i)$	D_0	D_{-1}	max
−3.0	1	1	−1.96	.05	.025	.025	.050	.050
−2.2	1	2	−1.44	.10	.075	.025	.075	.075
−2.0	1	3	−1.30	.15	.097	.053	.074	.074
−1.2	2	5	−.78	.25	.218	.032	.154	.154
−1.0	1	6	−.65	.30	.258	.052	.083	.083
−.2	2	8	−.13	.40	.449	.049	.143	.143
.0	3	11	.00	.55	.500	.050	.102	.102
.8	4	15	.52	.75	.699	.051	.250	.250
1.0	2	17	.65	.85	.742	.108	.151	.151
1.8	1	18	1.17	.90	.879	.021	.157	.157
2.0	2	20	1.30	1.00	.903	.097	.120	.120

testing $J = 3$ hypotheses requires that each hypothesis is tested at the level of

$$\alpha_{PC} = \tfrac{1}{J}\alpha_{PF} = \tfrac{1}{3} \times .05 = .0167\,. \tag{8}$$

With a Šidák approach, each hypothesis will be tested at the level of

$$\alpha_{PC} = 1 - (1 - \alpha_{PF})^{\frac{1}{J}} = 1 - (1 - .05)^{\frac{1}{3}} = .0170\,. \tag{9}$$

As this example illustrates, both procedures are likely to require using different α levels than the ones given by the tables. In fact, it is rather unlikely that a table could be precise enough to provide the wide range of alpha values needed for multiple testing purposes. A

more practical solution is to generate the critical values for any alpha value, or, alternatively, to obtain the probability associated to any value of the Kolmogorov-Smirnov criterion. Such an approach can be implemented by approximating the sampling distribution "on the fly" for each specific problem and deriving the critical values for unusual values of α.

Another approach to finding critical values for unusual values of α, is to find a numerical approximation for the sampling distributions. Molin and Abdi (1998) proposed such an approximation and showed that it was accurate for at least the first two significant digits. Their procedure, somewhat complex, is better implemented with a computer and comprises two steps.

The first step is to compute a quantity called A obtained from the following formula:

$$
A = \frac{-(b_1 + N) + \sqrt{(b_1 + N)^2 - 4b_2\left(b_0 - \mathsf{L}^{-2}\right)}}{2b_2} , \tag{10}
$$

with

$$
\begin{aligned}
b_2 &= 0.08861783849346 \\
b_1 &= 1.30748185078790 \\
b_0 &= 0.37872256037043 .
\end{aligned} \tag{11}
$$

The second step implements a polynomial approximation and estimates the probability associated to a given value L as:

$$
\begin{aligned}
\Pr(\mathsf{L}) \approx\ &-.37782822932809 + 1.67819837908004\,A \\
&- 3.02959249450445\,A^2 + 2.80015798142101\,A^3 \\
&- 1.39874347510845\,A^4 + 0.40466213484419\,A^5 \\
&- 0.06353440854207\,A^6 + 0.00287462087623\,A^7 \\
&+ 0.00069650013110\,A^8 - 0.00011872227037\,A^9 \\
&+ 0.00000575586834\,A^{10} .
\end{aligned} \tag{12}
$$

For example, suppose that we have obtained a value of $\mathsf{L} = .1030$ from a sample of size $N = 50$. (Table 2 shows that $\Pr(\mathsf{L}) = .20$.)

To estimate Pr(L) we need first to compute A, and then use this value in Equation 12. From Equation 10, we compute the estimate of A as:

$$A = \frac{-(b_1 + N) + \sqrt{(b_1 + N)^2 - 4b_2 \left(b_0 - L^{-2}\right)}}{2b_2}$$

$$= \frac{-(b_1 + 50) + \sqrt{(b_1 + 50)^2 - 4b_2 \left(b_0 - .1030^{-2}\right)}}{2b_2}$$

$$= 1.82402308769590 \, . \tag{13}$$

Plugging in this value of A in Equation 12 gives

$$\text{Pr(L)} = .19840103775379 \approx .20 \, . \tag{14}$$

As illustrated by this example, the approximated value of Pr(L) is correct for the first two decimal values.

References

[1] Abdi, H. (1987). *Introduction au traitement statistique des données expérimentales*. Grenoble: Presses Universitaires de Grenoble.

[2] Dagnelie, P. (1968). A propos de l'emploi du test de Kolmogorov-Smirnov comme test de normalité, *Biométrie et Praximétrie* **9**, 3–13.

[3] Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *Journal of the American Statistical Association*, **62**, 399–402.

[4] Molin, P., Abdi H. (1998). *New Tables and numerical approximation for the Kolmogorov-Smirnov/Lillierfors/Van Soest test of normality*. Technical report, University of Bourgogne. Available from `www.utd.edu/~herve/MA_Lilliefors98.pdf`.

[5] Van Soest, J. (1967). Some experimental results concerning tests of normality. *Statistica. Neerlandica*, **21**, 91–97, 1967.

Table 2: Table of the critical values for the Kolmogorov-Smirnov/Lillefors test of normality obtained with $K = 100,000$ samples for each sample size. The intersection of a given row and column shows the critical value $L_{critical}$ for the sample size labelling the row and the alpha level labelling the column. For $N > 50$ the critical value can be found by using $f_N = \dfrac{.83 + N}{\sqrt{N}} - .01$.

N	$\alpha = .20$	$\alpha = .15$	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
4	.3027	.3216	.3456	.3754	.4129
5	.2893	.3027	.3188	.3427	.3959
6	.2694	.2816	.2982	.3245	.3728
7	.2521	.2641	.2802	.3041	.3504
8	.2387	.2502	.2649	.2875	.3331
9	.2273	.2382	.2522	.2744	.3162
10	.2171	.2273	.2410	.2616	.3037
11	.2080	.2179	.2306	.2506	.2905
12	.2004	.2101	.2228	.2426	.2812
13	.1932	.2025	.2147	.2337	.2714
14	.1869	.1959	.2077	.2257	.2627
15	.1811	.1899	.2016	.2196	.2545
16	.1758	.1843	.1956	.2128	.2477
17	.1711	.1794	.1902	.2071	.2408
18	.1666	.1747	.1852	.2018	.2345
19	.1624	.1700	.1803	.1965	.2285
20	.1589	.1666	.1764	.1920	.2226
21	.1553	.1629	.1726	.1881	.2190
22	.1517	.1592	.1690	.1840	.2141
23	.1484	.1555	.1650	.1798	.2090
24	.1458	.1527	.1619	.1766	.2053
25	.1429	.1498	.1589	.1726	.2010
26	.1406	.1472	.1562	.1699	.1985
27	.1381	.1448	.1533	.1665	.1941
28	.1358	.1423	.1509	.1641	.1911

Table continues on the following page …

Table 3: . . . *Continued.* Table of the critical values for the Kolmogo-rov-Smirnov/Lillefors test of normality obtained with $K = 100,000$ samples for each sample size. The intersection of a given row and column shows the critical value $L_{critical}$ for the sample size labelling the row and the alpha level labelling the column. For $N > 50$ the critical value can be found by using $f_N = \dfrac{.83 + N}{\sqrt{N}} - .01$.

N	$\alpha = .20$	$\alpha = .15$	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
29	.1334	.1398	.1483	.1614	.1886
30	.1315	.1378	.1460	.1590	.1848
31	.1291	.1353	.1432	.1559	.1820
32	.1274	.1336	.1415	.1542	.1798
33	.1254	.1314	.1392	.1518	.1770
34	.1236	.1295	.1373	.1497	.1747
35	.1220	.1278	.1356	.1478	.1720
36	.1203	.1260	.1336	.1454	.1695
37	.1188	.1245	.1320	.1436	.1677
38	.1174	.1230	.1303	.1421	.1653
39	.1159	.1214	.1288	.1402	.1634
40	.1147	.1204	.1275	.1386	.1616
41	.1131	.1186	.1258	.1373	.1599
42	.1119	.1172	.1244	.1353	.1573
43	.1106	.1159	.1228	.1339	.1556
44	.1095	.1148	.1216	.1322	.1542
45	.1083	.1134	.1204	.1309	.1525
46	.1071	.1123	.1189	.1293	.1512
47	.1062	.1113	.1180	.1282	.1499
48	.1047	.1098	.1165	.1269	.1476
49	.1040	.1089	.1153	.1256	.1463
50	.1030	.1079	.1142	.1246	.1457
> 50	$\dfrac{0.741}{f_N}$	$\dfrac{0.775}{f_N}$	$\dfrac{0.819}{f_N}$	$\dfrac{0.895}{f_N}$	$\dfrac{1.035}{f_N}$