Normal Probability Plots and Tests for Normality

Thomas A. Ryan, Jr. and Brian L. Joiner, Statistics Department, The Pennsylvania State University 1976

Acknowledgments: Helpful assistance from Dr. Barbara Ryan and discussions with Dr. James J. Filliben and Dr. Samuel S. Shapiro are gratefully acknowledged. Please see Dr. Thomas Ryan's Note on a Test for Normality at the end of this document.

Introduction

Normal probability plots are often used as an informal means of assessing the non-normality of a set of data. One problem confronting persons inexperienced with probability plots is that considerable practice is necessary before one can learn to judge them with any degree of confidence. Some objective measure of the straightness of a probability plot would he helpful, especially for students just beginning their statistical education.

One rather obvious way to judge the near linearity of any plot is to compute its "correlation coefficient." When this is done for normal probability plots, a formal test can be obtained that is essentially equivalent to the powerful Shapiro-Wilk test W and its approximation W. This note is basically an exposition of the utility of this simple yet powerful procedure.

An Example

Figure 1 shows a normal probability plot of 70 IQ scores that were obtained as a covariate in a study concerning the relative effectiveness of color versus black and white visual materials. This particular plot provides an example of the need for a simple objective way to assess the straightness of probability plots. This plot may seem curved enough at the ends to cast serious doubt upon the hypothesis on normality. However, the "correlation coefficient" of this plot was 0.990, which is not significant at a = 0.10 (critical value = 0.9856). In fact, plots as curved as this occur fairly often with normal data (see, e.g., [6]). Of course, there can still be practically significant departures from normality, even though the hypothesis of normality is not rejected.

Figure 1: Normal Probability Plot of IQ Scores of 70 Students

```
140.0+
     -
     -                                                  *
     -                                             2
     -                                       2 3
130.0+                                     2*
     -                                   3
     -                                 4
     -                               23
     -                             *3
120.0+                            3*
     -                          22
     -                         22
     -                       4*
     -                      3
110.0+                   * 5
     -                  22
     -                 2
     -               2*
     -           * 3
100.0+         2
     -      *
     -  *
     -
     -
 90.0+
       +---------+---------+--------+---------+---------+
     -2.50               -0.50               1.50
               -1.50               0.50                2.50
```

More Details

A normal probability plot (see, e.g., [6], [8], or [19]) is basically a plot of the **ordered observations** from a sample against the corresponding percentage points from the standard normal distribution. If we denote the ordered observations in a sample of size n by {Yi}, then a normal probability plot can be produced by plotting the Yi on normal "probability" graph paper against some simple function like $p_i = \left(i - \frac{3}{8}\right) / \left(n + \frac{1}{4}\right)$ or $p_i = \frac{i}{n+1}$.

Using the special graph paper is equivalent to plotting the {Yi} on standard arithmetic graph paper against {bi} where bi is the pith percentage point of the standard normal distribution. That is, $b_i = \Phi^{-1}(p_i)$.

If the data come from a normal distribution, they will fall on an approximately straight line, whereas if they come from some alternative distribution, the plot will exhibit some degree of curvature. If the data fall nearly on a straight line, the "correlation coefficient" will be near unity, whereas if the plot is curved, the "correlation coefficient" will be smaller. If it falls below an appropriate critical value, doubt will be cast on the null hypothesis of normality. Thus the "probability plot correlation coefficient" version of the Shapiro-Wilk test is given by the familiar formula for a correlation coefficient, namely

$$Rp = \frac{\sum (Y_i - \overline{Y})(b_i - \overline{b})}{\sqrt{\sum (Y_i - \overline{Y})^2 \sum (b_i - \overline{b})^2}}$$

Since $\overline{b}$ = 0, Rp can be simplified to

$$Rp = \frac{\sum Y_i b_i}{\sqrt{\sum (Y_i - \overline{Y})^2 \sum b_i^2}}, \text{ or } \frac{\sum Y_i b_i}{\sqrt{s^2 (n-1) \sum b_i^2}}$$

where s2 denotes the sample variance.

Filliben [9, 10] suggested plotted the {Yi} against {Ci} where Ci is the **median** of the ith order statistic in samples from the standard normal distribution. The Ci may also be viewed as F-1(pi) where pi is the median of the ith order statistic in samples from the uniform distribution. For simplicity of computation, we suggest the use of pi or pi rather than pi since it does not appear to make any practical difference. Either test has the highly desirable feature of linking together a graphical display of the data with a simple, objective test statistic.

Some may object to the use of the term "correlation coefficient" since the {bi} are not random variables. However, another view is that, given any set of points in the plane, one can use the "correlation coefficient" associated with those points as a descriptive measure of how close they are to a straight line. In this sense, Rp can be thought of as a correlation coefficient. However, since Rp does not arise from sampling a bivariate distribution, it is not the same as the usual correlation coefficient. In fact, since both {bi} and {Yi} are ordered, Rp 0, and, in most practical cases, Rp is very large, even if the Yis come from a non-normal population.

A very useful approximation for making probability plots and/or computing Rp is [17, 12]

$$b_i \approx 4.91\left[p_i^{.14} - \left(1 - p_i\right)^{.14}\right]$$

A slightly more accurate approximation is [11]

$$b_i \approx \frac{g_0 + g_1 u + g_2 u^2}{1 + g_3 u + g_4 u^2 + g_5 u^3} - u,$$

where u = [-2 log e (p i)], and (g 0 , g 1 , ..., g 5) = (2.515517; 0.802853; 0.010328; 1.432788; 0.189269; 0.001308).

Either of these approximations is adequate. Use of these simple formulas in computer programs obviates the need to store the large tables of coefficients required for W and W.

Relationship to Shapiro-Wilk and Shapiro-Francia Tests

There is a very close relationship between R and the Shapiro-Wilk [16] test W and the Shapiro-Francia [15] approximation W. In fact, $\sqrt{W'}$ can be viewed as the "correlation coefficient" of a probability plot in which the expected values of the standardized normal order statistics m_i are used as plotting positions rather than the normal percentage points b_i. Similarly, $\sqrt{W}$ is proportional to the "correlation coefficient" associated with a probability plot in which the plotting positions are the coefficients a_i of the "best linear unbiased estimate" (BLUE) of the standard deviation σ . Since the expected values of the normal order statistics, the normal percentage points and the (scaled) BLUE coefficients are all quite similar (see, e.g., Table 1), similar properties should be expected among the three statistics W, W, and R_p. This indeed turns out to be the case as shown in the section entitled Power. Note in particular in Table 1 that the coefficients for W and R_p are in especially close agreement. The closeness of these three test statistics can also be anticipated from the theory of BLUEs and their approximations (see, e.g., [7]).

Table 1: Coefficients for the Three Tests W, W and R_p for n = 20.

Test:	R_p	W	W		
	Coefficients			**Ratios**	
i	**b_i**	**m_i**	**a_i 4.4122**	**m_i b_i**	**(a_i 4.4122) b_i**
11.	0.0619	0.0620	0.0618	1.0011	0.9979
12.	0.1867	0.1870	0.1862	1.0013	0.9971
13.	0.3146	0.3149	0.3137	1.0011	0.9972
14.	0.4478	0.4483	0.4470	1.0012	0.9983
15.	0.5895	0.5903	0.5886	1.0014	0.9986
16.	0.7441	0.7454	0.7439	1.0017	0.9997
17.	0.9191	0.9210	0.9199	1.0020	1.0008
18.	1.1281	1.1310	1.1317	1.0025	1.0032
19.	1.4034	1.4076	1.4168	1.0030	1.0095
20.	1.8683	1.8675	2.0887	0.9996	1.1180

Correlation between (all 20) b_i and m_i values = 1.0000.

Correlation between (all 20) b_i and a_i values = 0.9986.

Thus, R_p is basically a new way of viewing very good established procedures. R_p is easy to explain to students in an elementary statistics course and to researchers from other fields, since it is linked to a graphical technique (probability plots) and is based on a technique taught early in most courses (correlation coefficient).

In addition, R_p is very easy to calculate, especially on a computer, since no special tables are required for its computation. And the critical values needed to complete the test can be easily calculated using formula (1) in the section entitled Critical Values. For example, in Minitab [14] there is a command called NSCORES that computes the b_i (normal scores) for any sample size. The following brief program reads in a batch of data, computes the normal scores, does a probability plot, and computes R_p. Note that no new commands need to be added to the system to compute R_p.

```
SET THE FOLLOWING IQ SCORES INTO COLUMN C1
(data come here)
NSCORES FOR DATA IN COL C1, PUT IN COL C2
PLOT COL C1 VS COL C2 (PROBABILITY PLOT)
CORRELATION BETWEEN C1 AND C2 (R-SUB-P)
STOP
```

This program produced the plot in Figure 1.

The critical importance of linking together graphical displays with objective test statistics cannot be overemphasized. This advantage is theoretically available with W and W but their use in this connection and their relationship with the "correlation coefficient" of normal probability plots has not previously been noted. In fact, when W was introduced [16] as a test based on the ratio of two estimates of variance, the authors stated that "Heuristic considerations augmented by some fairly extensive empirical sampling results suggest that the mean values of W for non-null distributions tends to shift to the left of that for the null case." Had the authors noted the connection with the "correlation coefficient" on a normal probability plot, it would have been apparent why low values of W, and thus low correlations, would have been indicative of non-normality.

Critical Values

Approximate critical values of R_p were obtained from Monte Carlo simulations using Chen's [3] algorithm to obtain the random normal samples. Five hundred independent random samples were generated for each value of n between 11 and 77, and 3500 samples were generated for each value of n between 3 and 10. The empirical critical values were computed for $\alpha = 0.10$, 0.05, and 0.01. The results were then smoothed for each value of α using a function of the form

$$CV(n) = b_0 + \frac{b_1}{\sqrt{n}} + \frac{b_2}{n} + \frac{b_3}{n^2} \qquad (1)$$

There was no detectable lack of fit using these functions, so it would appear that the simple approximations listed in Table 2 give critical values yielding α accurate to within 0.007 for = 0.10; to within 0.005 for = 0.05; and to within 0.002 for = 0.01. These uncertainties represent estimated limits to the error in α and are computed from upper bounds to twice the standard error of the fitted values using weighted least squares fits of (1).

Table 2: Approximate critical values for Rp.

n	α = 0.10	α = 0.05	α = 0.01
4	.8951	.8734	.8318
5	.9033	.8804	.8320
10	.9347	.9180	.8804
15	.9506	.9383	.9110
20	.9600	.9503	.9290
25	.9662	.9582	.9408
30	.9707	.9639	.9490
40	.9767	.9715	.9597
50	.9807	.9764	.9664
60	.9835	.9799	.9710
75	.9865	.9835	.9757

Approximate critical values for intermediate values of n are given by the following equations:

$$cv(n) \approx 1.0071 - \frac{0.1371}{\sqrt{n}} - \frac{0.3682}{n} + \frac{0.7780}{n^2}$$, for α = 0.10;

$$cv(n) \approx 1.0063 - \frac{0.1288}{\sqrt{n}} - \frac{0.6118}{n} + \frac{1.3505}{n^2}$$, for α = 0.05;

$$cv(n) \approx 0.9963 - \frac{0.0211}{\sqrt{n}} - \frac{1.4106}{n} + \frac{3.1791}{n^2}$$, for α = 0.01.

Limiting "Correlations"

The most important question is usually not "Is the population normal?" because we already know that no real population is *exactly* normal. Rather, the important questions are "How non-normal is the population?" and "How much is the non-normality going to hurt?". The statistic R_p can be used to provide an indication of the answer to the first of these questions. This will be particularly true with larger samples. Thus it will be useful to have a means of interpreting R_p asymptotically.

If increasingly larger samples are drawn from some alternative distribution F, the test statistic R_p converges in probability to the "**limiting correlation**"

$$\rho(F, \Phi) = \frac{\int_0^1 F^{-1}(p)\,dp}{\sigma_F}$$

where σ_F is the standard deviation of F. Roughly, $\rho(F, \Phi)$ may be viewed as the "correlation" between the two distributions F and Φ associated with the plot of $F^{-1}(x)$ versus $\Phi^{-1}(x)$. It is also instructive to think of R_p as a sample estimate of $\rho(F, \Phi)$. In this way, R_p can be used as an estimate of how far the population is from normality. Use of this procedure in conjunction with the normal probability plots of alternative distribution functions, as given by Chambers and Fowlkes [2], might be particularly informative. In addition, $\rho(F, \Phi)$ is useful as an indicator of what sort of power one can expect of R_p, as will be seen in the following section.

To make the concept of a distance from normality more precise, we can define a metric based on ρ. If F and G are any two distributions, then we define

$$\rho(F, G) = \frac{\int_0^1 (F^{-1}(p) - \mu_F)(G^{-1}(p) - \mu_G)\,dp}{\sigma_F \sigma_G}$$

where μ_F, μ_G, σ_F^2, and σ_G^2 are the means and variances of F and G. Then $d(F,G) = \sqrt{1 - \rho^2(F,G)}$ is a distance function between classes of distribution functions where two distributions are in the same class if they differ only in location and scale parameters. It is routine to verify that d(F, G) is non-negative, is zero if and only if F and G are in the same class, and is symmetric. The triangle inequality is readily established using the fact that for any random variables X_1, X_2, X_3,

$$\begin{aligned} 1 - \rho_{13}^2 &= \min_b E\left[(X_3 - bX_1)^2\right] \\ &\le E\left[(X_3 - \rho_{12}X_2) + \rho_{23}(X_2 - \rho_{12}X_1)\right]^2 \\ &\le (1 - \rho_{12}^2) + (1 - \rho_{13}^2) + 2\sqrt{(1 - \rho_{12}^2)(1 - \rho_{13}^2)} \ . \end{aligned}$$

Hence, $d(F, \Phi)$ may be used as a measure of distance from normality and $\sqrt{1 - R_p^2}$ may be used as an estimate of this distance. Limiting "correlations" for some alternative distributions are given in Table 3.

Table 3: Limiting correlations between the normal distribution and selected alternative distributions.

Distribution	ρ ***(F, Φ)***
Uniform	0.9770
Right Triangle	0.9730
Exponential	0.9025
Weibull (c=2)	0.9857
t with 3 d.f.	0.9082
t with 5 d.f.	0.9841

Power

To obtain estimates of the power of R_p, and more particularly, of the difference in power between R_p and W, additional computer simulations were performed. Using the uniform random number generator of Lewis et. al. [13], 500 sets of data of size n = 10 and n = 20 were simulated for each alternative distribution listed in Table 4. The notation for the contaminated normals may be explained in term of the entry (0.10, 5), which means that with probability 0.10 an observation was drawn from a normal distribution with $\sigma = 5$ and with probability 0.90 from a normal distribution with $\sigma = 1$, always with $\mu = 0$. The empirical power of R_p and W were then computed for these alternatives. The alternative distributions considered are essentially a subset of those investigated by Shapiro, Wilk, and Chen [17] and Chen [4].

Table 4: Empirical power of Rp and W for selected alternative distributions (α = 0.10)

	(power x 100)			
	n = 10		**n = 20**	
Distribution	**R_p**	**W**	**R_p**	**W**
Uniform	**13**	**18**	**20**	**37**
Right Triangle	**21**	**22**	**33**	**46**
Exponential	**53**	**54**	**89**	**90**
Weibull (c = 2)	**13**	**12**	**21**	**25**
Weibull (c = 0.5)	**93**	**94**	**100**	**100**
Lognormal (σ = 1)	**66**	**67**	**96**	**97**
Cauchy	**73**	**68**	**91**	**88**
Contaminated Normal (0.10, 5)	**42**	**41**	**62**	**59**
Contaminated Normal (0.05, 5)	**29**	**27**	**41**	**41**
Contaminated Normal (0.10, 3)	**23**	**20**	**38**	**33**
Contaminated Normal (0.05, 3)	**18**	**17**	**28**	**27**

The standard error of the power figures for R_p and W shown in Table 4 is always less than 0.025. The power values for W reported here agree well with those reported previously except for the Cauchy distribution with n = 10 where the value reported by Shapiro, Wilk, and Chen [17] appears to be too low. Since identical data sets were used in obtaining the empirical power functions of the two tests, the *comparisons* between R_p and W indicated in Table 4 are substantially more accurate. Table 4 shows that overall there is little difference between the powers of the two tests for most of the alternatives reported. The only appreciable difference is that for extremely short-tailed distributions like the uniform and triangular, W has more power than R_p, while for heavy-tailed distributions like the Cauchy and contaminated normals, R_p does slightly better. This difference in the (scaled) coefficients for the two statistics W and R_p is seen to be for the largest and smallest observations, where the a_i values are more extreme than the b_i values. Thus, the a_i values will tend to agree **better** with samples from long-tailed distributions, and therefore give W **less power** than R_p. Conversely, the a_i values will tend to disagree more with short-tailed samples and thus give W better power there. Since the m_i coefficients for W are nearly identical to the b_i values used for R_p, one would expect that the properties of W and R_p would agree even more closely than those for W and R_p. This appears to be the case, with the statistics $\sqrt{W'}$ and R_p agreeing to 3 decimal places in all samples observed.

A comparison between the limiting "correlations" in Table 3 and the corresponding power values in Table 4 suggests that ρ (F, Φ) does provide a useful indication of the non-normality of an alternative distribution function.

Concluding Remarks

The notion of using the familiar correlation coefficient as a means of judging the straightness of a normal probability plot is intuitively appealing. This test has the virtues of being simple, easily remembered, and powerful. It encourages the use and comparison of a visual test (the probability plot) with an objective measure (R_p). This test can also be used to provide an intuitive explanation of why the Shapiro-Wilk and the Shapiro-Francia tests work.

All four tests mentioned here (Filliben’s, W, W , and R_p) are intrinsically location and scale invariant and are thus readily usable against composite alternatives. The can also be used in the manner described by Wilk and Shapiro [20] to jointly assess the normality of several small data sets.

It would be each to extend R_p for use with censored samples. New tables of critical values would be required, but the basic procedure would be to simply omit the b_is corresponding to the censored observations, and compute the "correlation" between the observed Y_i and the corresponding b_i. The other tests could be extended in a similar fashion if viewed as "correlation coefficients."

The whole procedure can be extended to almost any other distribution, F. Simply compute $F^{-1}(p_i)$ and their correlation with the Y_i. A brief exploratory study of this extension to the exponential distribution has indicated that R_p may not be as powerful in that case as the exponential version of the W test. Thus the general efficacy of extending R_p to other situations remains in doubt.

References

[1.] Blom, Gunnar (1958). *Statistical Estimates and Transformed Beta-Variables*. New York: John Wiley & Sons.

[2.] Chambers, Mrs. E. and Fowlkes, E. B. (1966). "A Dictionary of Distributions: I. Comparisons With the Standard Normal." Paper presented at the Eastern Regional Meeting of the Institute of Mathematical Statistics, Upton, Long Island, New York.

[3.] Chen, E. H. (1971a). "A Random Number Generator for 32-bit-word Computers." *Journal of the American Statistical Association*, No. 66, pp. 400–403.

[4.] Chen, E. H. (1971b). "The Power of the Shapiro-Wilk W Test for Normality in Samples from Contaminated Normal Distributions." *Journal of the American Statistical Association*, No. 66, pp. 760–762.

[5.] Chernoff, H. and Lieberman, G. J. (1954). "Use of Normal Probability Paper." *Journal of the American Statistical Association*, No. 49, pp. 778–785.

[6.] Daniel, Cuthbert and Wood, Fred S. (1971). *Fitting Equations to Data*. New York: John Wiley & Sons.

[7.] David, H. A. (1970). *Order Statistics*. New York: John Wiley & Sons.

[8.] Dixon, Wilfrid J. and Massey, Frank J., Jr. (1969). *Introduction to Statistical Analysis*, 3rd Edition. McGraw-Hill, New York.

[9.] Filliben, James J. (1971). The Probability Plot Correlation Coefficient Test for Normality. (Unpublished manuscript.)

[10.] Filliben, James J. (1972). "Techniques for Tail Length Analysis." *Proceedings of the 18th Conference on the Design of Experiments in Army Research Development and Testing*. October 25–27, 1972.

[11.] Hastings, Cecil, Jr. (1955). *Approximations for Digital Computers*. Princeton University Press, Princeton, New Jersey.

[12.] Joiner, Brian L. and Rosenblatt, Joan R. (1971). "Some Properties of the Range in Samples from Tukey's Symmetric Lambda Distributions." *Journal of the American Statistical Association*, No. 66, pp. 394–399.

[13.] Lewis, P. A. W., Goodman, A. S. and Millo, J. M. (1969). "A Pseudo-Random Number Generator for the System/360." *IBM System Journal*, No. 8, pp. 136–146.

[14.] Ryan, T. A. and Joiner, B. L. "Minitab: A Statistical Computing System for Students and Researchers." *The American Statistician*, No. 27, pp. 222–225.

[15.] Shapiro, S. S. and Francia, R. S. (1972). "An Approximate Analysis of Variance Test for Normality." *Journal of the American Statistical Association*, No. 67, pp. 215–216.

[16.] Shapiro, S. S. and Wilk, M. B. (1965). "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika*, No. 52, pp. 591–611.

[17.] Shapiro, S. S., Wilk, M. B. and Chen, Mrs. H. J. (1968). "A Comparative Study of Various Tests for Normality." *Journal of the American Statistical Association*, No. 63, pp. 1343–1372.

[18.] Tukey, J. W. (1960). *The Practical Relationship Between the Common Transformations of Percentages or Fractions and of Amounts*. Technical report 36, Statistical Research Group, Princeton University.

[19.] Wilk, M. B. and Gnanadisikan, R. (1968). "Probability Plotting Methods for the Analysis of Data." *Biometrika*, No. 55, pp. 1–17.

[20.] Wilk, M. B. and Shapiro, S. S. (1968). "The Joint Assessment of Normality of Several Independent Samples." *Technometrics*, No. 10, pp. 825–839.

Note on a Test for Normality

Thomas A. Ryan, Jr.
October 4, 1990

This note updates the 1974 technical report that I wrote with Brian Joiner concerning using a correlation coefficient associated with a normal probability plot as a test for normality. As you would expect, there is a better, more recent reference for this test. *Goodness-of-Fit Techniques,* edited by Ralph B. D'Augostino and Michael A. Stevens (Dekker, 1986) describes this test in some detail, and provides a table that can be used to construct critical values for the test for n from 10 to 1000. See pages 195–205 (especially Section 5.7 and the table on page 203).

There is very good agreement between our critical values and those in D'Augostino and Stevens. Transforming their table gives the results below, along with our table for comparison. (I now believe, based on graphical displays, that the .01 critical value for n = 4 may be wrong.) (D'Augostino and Stevens also give results for other levels of alpha.)

	D'Augostino and Stevens			**Joiner and Ryan**		
alpha	*0.10*	*0.05*	*0.01*	*0.10*	*0.05*	*0.01*
N 4	–	–	–	0.8951	0.8734	0.8318
5	–	–	–	0.9033	0.8804	0.8320
10	0.9349	0.9176	0.8792	0.9347	0.9180	0.8804
15	–	–	–	0.9506	0.9383	0.9110
20	0.9602	0.9511	0.9270	0.9600	0.9503	0.9290
25	–	–	–	0.9662	0.9582	0.9408
30	–	–	–	0.9707	0.9639	0.9490
40	0.9769	0.9717	0.9579	0.9767	0.9715	0.9597
50	–	–	–	0.9807	0.9764	0.9664
60	0.9835	0.9799	0.9710	0.9835	0.9799	0.9710
75	–	–	–	0.9865	0.9835	0.9757
80	0.9871	0.9843	0.9776	–	–	–
100	0.9894	0.9871	0.9818	–	–	–
400	0.9969	0.9964	0.9950	–	–	–
600	0.9979	0.9975	0.9966	–	–	–
1000	0.9987	0.9984	0.9979	–	–	–

There is some other material in the technical report that has not, to my knowledge, been published anywhere. An example is the asymptotic values of the correlation coefficient for alternative distributions.