

■ RICK HESSE, Feature Editor, Pepperdine University

Normal Probability Plots

By Rick Hesse, Pepperdine University

Since this column has dealt with spreadsheet applications of decision sciences, readers may be interested that during June 27-30, 1998, the Tuck School at Dartmouth College will host an intensive, three-day workshop on Teaching Management Science with Spreadsheets. The workshop is intended for teachers who are new to teaching management science with spreadsheets, as well as those who are experienced. It will bring together a faculty of 15 experts in teaching with spreadsheets (including me) to work closely with a participant group of no more than 60 on a wide range of topics, from how to utilize spreadsheets in the classroom effectively, to how to teach nonlinear programming using spreadsheets. The workshop is cosponsored by INFORMS, DSI, and IFORS. The cost is \$695 before May 15, 1998, and \$795 thereafter. Complete information and registration forms are available at:

www.dartmouth.edu/tuck/tmss

As I have started to teach graduate and undergraduate statistics again these last two years, I have been impressed with how many simple, visual tools are available to “look” at the data. A few months ago I shared a Box and Whiskers plot and now would like to share a normal probability plot. These plots are a quick and dirty visual graphing technique to “see” if a data set exhibits the properties of a normal distribution. The idea is to rank the data set and change the ranks into percentiles that would be converted to z-scores. If the data is indeed approximately normally distributed, then the converted data points should lie in a straight line. Since the human eye can distinguish simple lines easier than curved ones, it is a quick and dirty visual test of normality, rather than just the cumulative probability plot. Shown in Figure 1 is the cumulative probability plot for the grades for 17 students of mine in the summer statistics class. The data to build the step function is shown along with the plot.

Rick Hesse

is professor of quantitative methods at Pepperdine University in the Graziada Graduate School of Business. He received his B.S., M.S., and D.Sc. at Washington University School of Engineering in applied math and computer science. Dr. Hesse is the author of *Managerial Spreadsheet Modeling & Analysis and Applied Management Science: A Quick & Dirty Approach* (with Gene Woolsey), articles in numerous journals, and software for personal computers. Rick was the first professor to be awarded the Outstanding Civilian Service Medal by the Department of the Army at West Point in 1982, and was the winner of the Decision Sciences Institute’s Innovative Instructional Award in 1981.

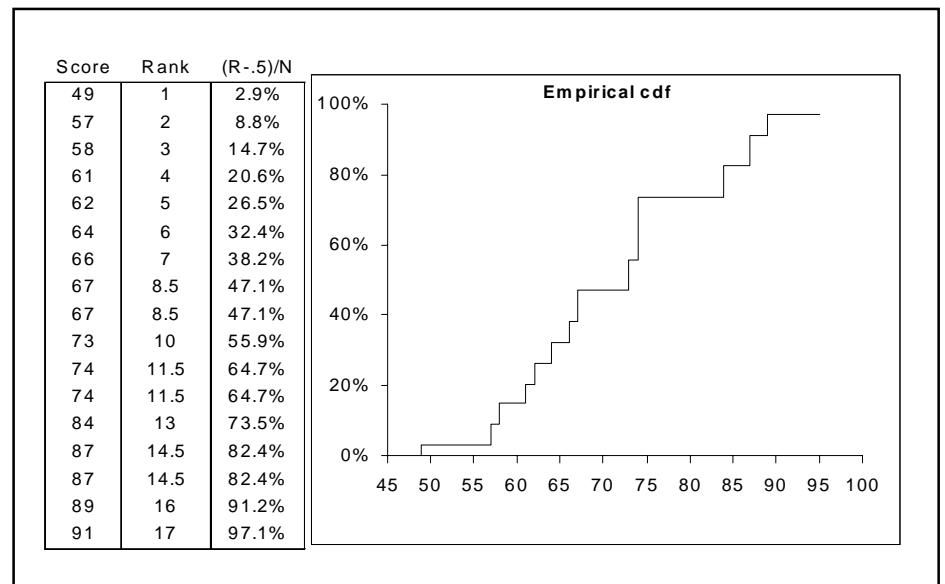


Figure 1: Cumulative plot.

If the data approximates a normal distribution, the curve should look like an "S," but this is a difficult visual to quantify for most people. How much does it have to look like an "S," for the data to approximate a normal curve? Human beings are better at identifying something that looks like a straight line. Therefore, a normal probability plot will be an easier way to check for normality. Excel 97, although it has many new features, does not have a Data Analysis program to convert a data set into a normal probability plot (or a cumulative step function for that matter), and so this article will give a simple template to accomplish this. A special feature of this template is that the data does not have to be sorted, but can be in any order, as long as it is in one column. Shown in Figure 2 is the data for the template, including the raw scores in column A (given the range name XS), the computed ranks (1 is lowest) in column B, the associated percentile with that rank in column C, and finally the z-scores for that percentile in column D (given the range name ZS). The important cell formulas are given below, with B4 being the most complicated, because Excel only gives the lowest rank for tied ranks.

B4: $(\text{COUNT}(\text{XS})+1+\text{RANK}(\text{A4},\text{XS},1) - \text{RANK}(\text{A4},\text{XS},0))/2$

C4: $(\text{B4}-0.5)/\text{COUNT}(\text{XS})$

D4: $\text{NORMINV}(\text{C4},0,1)$

Once the z-scores have been computed, columns A and D are graphed.

This is an XY scatter plot that uses points only (otherwise the graph would look like Etch-A-Sketch!). After the graph is created, the trendline and the equation are added on the chart by clicking on the data points, then using the right mouse button to bring up the trend line menu. By checking the option to show the equation and r^2 , we have all the information shown in Figure 3. This graph will allow us to visually check if the points are indeed close to a straight line or to see if there is a pattern to the points being above or below the line. If the data is somewhat close to being normally distributed, the points should lie approximately on the trend line, with the line crossing the x-axis at about the mean of the data, and the inverse of the slope should be close to the standard deviation of the data.

	A	B	C	D
1	Normal Probability Plot			
2	Score	Rank	(R-.5)/N	z-score
3	57	2	8.8%	-1.352
4	67	8.5	47.1%	-0.074
5	49	1	2.9%	-1.890
6	64	6	32.4%	-0.458
7	73	10	55.9%	0.148
8	91	17	97.1%	1.890
9	66	7	38.2%	-0.299
10	87	14.5	82.4%	0.929
11	74	11.5	64.7%	0.377
12	58	3	14.7%	-1.049
13	62	5	26.5%	-0.629
14	74	11.5	64.7%	0.377
15	84	13	73.5%	0.629
16	87	14.5	82.4%	0.929
17	67	8.5	47.1%	-0.074
18	61	4	20.6%	-0.821
19	89	16	91.2%	1.352

Figure 2: Data for normal probability plot.

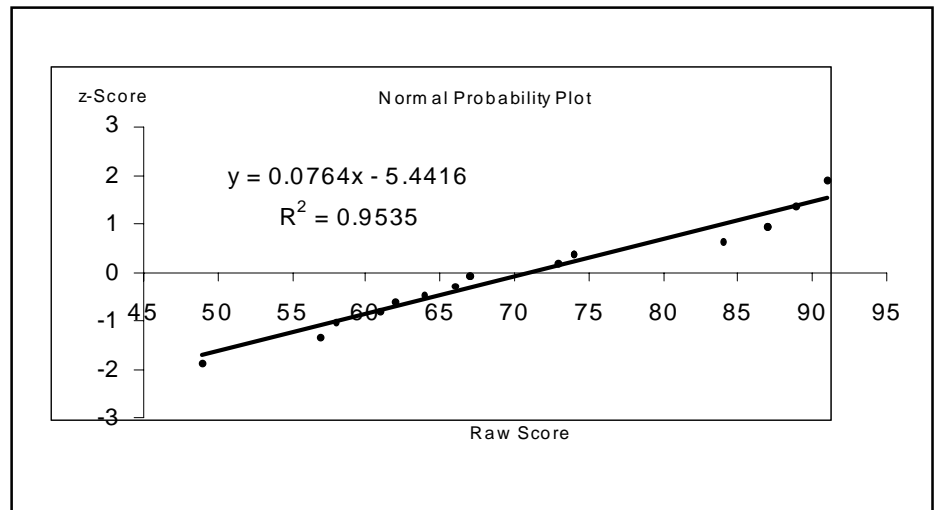


Figure 3: Regression line added to probability plot.

Excel gives the regression line coefficients on the chart, but they can also be computed using the formulas given for cells F18 and H18 (see Figure 5.)

F18: $= \text{TREND}(\text{ZS},\text{XS},1,1)-\text{H18}$

H18: $= \text{TREND}(\text{ZS},\text{XS},0,1)$

For normal probability plots, the mean is approximated by where the straight line crosses the x-axis (-intercept/slope) and the

approximate standard deviation is the reciprocal of the slope. These are computed in cells G3 and G4 and are also shown later in Figure 5.

G3: $= -\text{H18}/\text{F18}$

G4: $= \text{ABS}(1/\text{F18})$

Looking at Figure 3, it is pretty easy to see that these few data points seem to

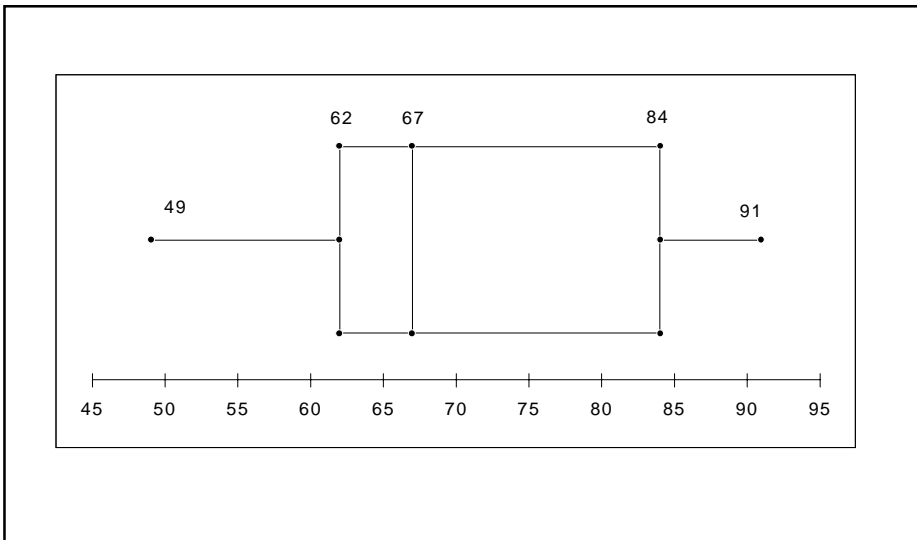


Figure 4: Box and whiskers plot.

	E	F	G	H
2		Calculated	Estimated	% difference
3	Average	71.18	71.19	0.02%
4	Std Dev	12.66	13.08	3.37%
5	Median	67.00		5.87%
6	P(< = avg)	52.94%		5.88%
18				
19	Slope	0.0764	Intercept	-5.4416

Figure 5: Calculated and estimated summary data.

fall on a straight line and that the data distribution seems to be normal. This visual check is a lot easier than using Figure 1 to see if the plot looks like an S-shaped curve. If a box and whiskers plot is used (see Figure 4), the plot is inconclusive, because a normal box plot should have two long whiskers and equal rectangles in the box. When there are a lot of data points, not all of the data points need be plotted—perhaps every other or every third. This is true both for the normal probability plot and box and whiskers. It is an easy matter to expand (or contract) the template in the middle (somewhere between rows 7-17). When all the appropriate formulas are copied, the graph is automatically redrawn and the trend line recomputed by Excel.

A final feature of the template is shown in Figure 5. The percent difference between the calculated and estimated mean and standard deviation are given in H3:H4. Also given is the median for purposes of comparing how close it is to the mean (which theoretically are equal for a normal distribution).

The median is calculated and compared to the calculated mean and then a simple logic check is used in K3:K19 (not shown) to compute the percentage of points below the mean in F6. The cell formulas are:

F3: = AVERAGE(XS)
 F4: = STDEV(XS)

F5: = MEDIAN(XS)
 F6: = AVERAGE(K3:K19)
 K3: = (A3<=\$F\$3)*1 copy to K3:K19
 H5: = ABS(F5/F3-1)
 H6: = ABS(1-F6/.5)
 H3: = ABS(G3/F3-1) copy to H4

This template, with both the visual plot and the actual versus estimated parameters, should be able to give the user a good idea if the data is “normal” without resorting to more difficult tests. ■

Dr. Rick Hesse, Graziadia Graduate School of Business, Pepperdine University, email: rickhesse@aol.com.