

A Tutorial on Quantile-Quantile Plots

Abstract – This is a tutorial on quantile-quantile plots, a technique for determining if different data sets originate from populations with a common distribution. The technique can be used to determine if a data set is normally distributed, and to optimize the transformation parameter of variance-stabilizing transformation models. An Excel link is provided.

Keywords: tutorial, quantiles, qq plots, normal distribution, data transformations, box-cox

Published: 10-02-2012; Updated: 07-08-2018

© E. Garcia, PhD; admin@minerazzi.com

Introduction

Although no data set is *exactly* normally distributed, most statistical analyses require that the data be *approximately* normally distributed for their findings to be valid.

One way of testing for normality is through a quantile-quantile (qq) plot, a tool that helps to determine if data sets originate from populations with a common distribution. These plots can be constructed to address a wide range of problems. Quantiles are points taken at regular intervals from the cumulative distribution function of a random variable (Wikipedia, 2012a).

In this tutorial you will determine if a data set is normally distributed by comparing its quantiles against those of a theoretical normal distribution. You will also learn how to make a data set nearly normally distributed. The data to be analyzed are the microwave radiation measurements reproduced by NIST (2012) from Johnson and Wichern's book (1998).

Background

A data set can be converted into z scores, sorted in ascending order, and then ranked to form a new set $S = \{1, 2, 3, \dots, j \dots k\}$, where j/k are the quantiles of a theoretical normal distribution.

The last quantile of S , k/k , corresponds to the 100th percentile. Because this is the maximum value of said distribution which is often infinite (Wikipedia, 2012b), the quantiles must be shifted over using the plotting position $(j - 0.5)/k$.

A qq plot of theoretical vs. experimental z scores can then be constructed by taking quantiles for probabilities and mapping these to theoretical z scores. To test for normality, the paired data can be fitted to a regression curve and a Pearson correlation coefficient, r , computed. The further r varies from 1, the greater the indication of deviation from normality.

This was the procedure developed when we first published this tutorial. Back then, we processed the data with *Microsoft Excel* version 2007. We chose *Microsoft Excel* because this spreadsheet software comes with several built-in functions and chart layouts that greatly simplify the analysis. Its chart wizard works better if the x-axis values are to the left of the y-axis values. You can reproduce this tutorial procedure with the instructions given in the next section.

Instructions

1. Open a new spreadsheet workbook and label cells A1 through F1 as follows:
rank, probability, z score *theo*, **data, unsorted score**, and **z score** *exp*
2. Enter data in cells D2 - DK where $K=k+1$ and k is the sample size (in this example $K=43$).
3. In cell E2, enter =STANDARDIZE(D2,AVERAGE(D\$2:D\$K),STDEV(D\$2:D\$K)) and paste it into the remaining column cells. This generates unsorted experimental z scores.
4. In cell F2, enter =SMALL(\$E\$2:\$E\$K,ROW()-ROW(\$E\$2)+1) and paste it into the remaining column cells. This sorts the experimental z scores in ascending order.
5. In cell A2, enter =RANK(F2,\$F\$2:\$F\$K,1)+(COUNT(\$F\$2:\$F\$K) + 1 - RANK(\$F2,\$F\$2:\$F\$K,0) - RANK(\$F2,\$F\$2:\$F\$K,1))/2 and paste it into the remaining column cells. This function considers rank ties, returning an average rank. To properly discern the averages, format cells to a proper number of decimal places. If you do not need to consider rank ties, enter instead =RANK(F2,\$F\$2:\$F\$K,1) in the column cells.
6. In cell B2, enter =(A2-0.5)/COUNT(A\$2:A\$K) and paste it into the remaining column cells. This converts ranks into probabilities of a normal distribution curve.
7. In cell C2, enter =NORMSINV(B2) and paste it into the remaining column cells. This function accepts a probability corresponding to a normal distribution and returns a z score that corresponds to an area under the curve of that distribution.
8. To generate the qq plot, construct a scatterplot of **z score** *theo* (x-axis) versus **z score** *exp* (y-axis) by choosing the **Scatter** menu from the **Insert** tab of *Excel* and selecting the **Scatter with only Markers** chart type.
9. Change the overall layout of the graph by selecting **Layout 9** from the **Chart Layouts** of the **Design** tab. *Excel* fits the data points to a straight line, returns a regression equation, and a coefficient of determination as R^2 .
10. Pearson correlation coefficient can be reported to two decimal places as $r = (R^2)^{1/2} = R$.

Results

You should be able to reproduce the results given in Tables 1 and 2 and Figures 1, 2, and 3. A link to an Excel .xlsx file that reproduces Table 1 is given at the end of the tutorial.

Table 1. Excel workbook data.

A	B	C	D	E	F
rank	probability	<i>z score</i> <small>theo</small>	data	unsorted score	<i>z score</i> <small>exp</small>
1.50	0.02	-1.98	0.15	0.22	-1.18
1.50	0.02	-1.98	0.05	-0.78	-1.18
4.00	0.08	-1.38	0.10	-0.28	-1.08
4.00	0.08	-1.38	0.05	-0.78	-1.08
4.00	0.08	-1.38	0.08	-0.48	-1.08
6.00	0.13	-1.12	0.20	0.71	-0.98
9.00	0.20	-0.83	0.09	-0.38	-0.78
9.00	0.20	-0.83	0.08	-0.48	-0.78
9.00	0.20	-0.83	0.10	-0.28	-0.78
9.00	0.20	-0.83	0.03	-0.98	-0.78
9.00	0.20	-0.83	0.18	0.52	-0.78
12.00	0.27	-0.60	0.20	0.71	-0.58
14.00	0.32	-0.46	0.18	0.52	-0.48
14.00	0.32	-0.46	0.10	-0.28	-0.48
14.00	0.32	-0.46	0.02	-1.08	-0.48
16.50	0.38	-0.30	0.05	-0.78	-0.38
16.50	0.38	-0.30	0.10	-0.28	-0.38
22.00	0.51	0.03	0.30	1.71	-0.28
22.00	0.51	0.03	0.10	-0.28	-0.28
22.00	0.51	0.03	0.07	-0.58	-0.28
22.00	0.51	0.03	0.10	-0.28	-0.28
22.00	0.51	0.03	0.15	0.22	-0.28
22.00	0.51	0.03	0.20	0.71	-0.28
22.00	0.51	0.03	0.30	1.71	-0.28
22.00	0.51	0.03	0.05	-0.78	-0.28
22.00	0.51	0.03	0.02	-1.08	-0.28
27.00	0.63	0.33	0.01	-1.18	-0.18
28.00	0.65	0.40	0.10	-0.28	-0.08
30.00	0.70	0.53	0.11	-0.18	0.22
30.00	0.70	0.53	0.40	2.71	0.22
30.00	0.70	0.53	0.12	-0.08	0.22
32.50	0.76	0.71	0.01	-1.18	0.52
32.50	0.76	0.71	0.40	2.71	0.52
35.00	0.82	0.92	0.15	0.22	0.71
35.00	0.82	0.92	0.30	1.71	0.71
35.00	0.82	0.92	0.30	1.71	0.71
38.50	0.90	1.31	0.08	-0.48	1.71
38.50	0.90	1.31	0.10	-0.28	1.71
38.50	0.90	1.31	0.10	-0.28	1.71
38.50	0.90	1.31	0.09	-0.38	1.71
41.50	0.98	1.98	0.02	-1.08	2.71
41.50	0.98	1.98	0.05	-0.78	2.71

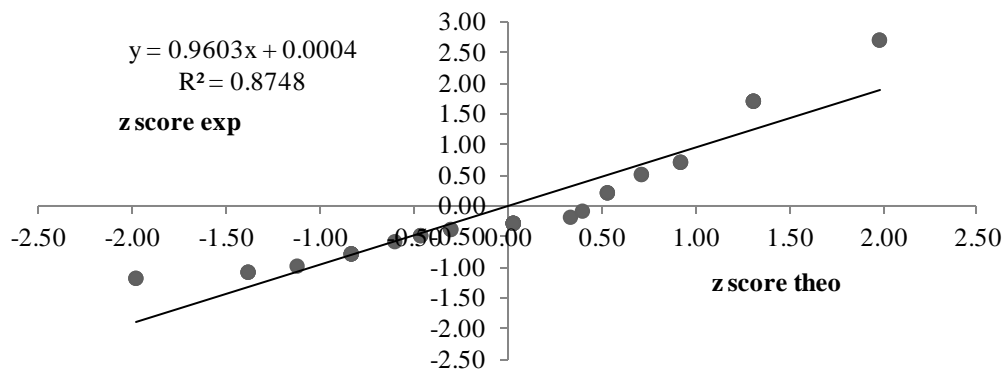


Figure 1. Quantile-quantile plot generated from Table 1.

Applications

Data that is normally distributed produce a qq plot with a Pearson correlation coefficient, r , close to 1. The further r is from 1, the greater the deviation from normality.

The data analyzed in this tutorial returned $r = (0.8748)^{1/2} = 0.9353 \approx 0.94$. While a relatively high value, this result shows that the data is not free from deviations. Said deviations can be minimized by applying variance-stabilizing transformations to the original data set before constructing the qq plot.

In general, a transformation that stabilizes the variance makes a distribution normal and vice versa. Table 2 shows two of the best-known data transformation models: Tukey and Box-Cox (Tukey, 1957; Box & Cox, 1964). Sakia (1992) and Hossain (2011) reviewed these models.

In Table 2, y is an experimental value, y^* is the transformed value, p is a power transformation parameter, and c is a positive constant that offsets any negative or zero value(s) of the data set. For $p = 1$ and $c = 0$, the Tukey model does not change the data ($y^* = y$), but the Box-Cox model subtracts 1 ($y^* = y - 1$). This should not alter the analysis, though.

For $p = 0$, both models return the same logarithmic transformation, but by different means. In the Tukey model, we evaluate the derivative dy^*/dp at $p = 0$. By contrast, in the Box-Cox model, we apply *l'Hôpital's Rule*.

Among others, the following transformations are commonly obtained from both models by setting p : quadratic ($p = 2$), square root ($p = 0.5$), and reciprocal ($p = -1$). A Box-Cox Power Transformations tool is available from Minerazzi (Garcia, 2016).

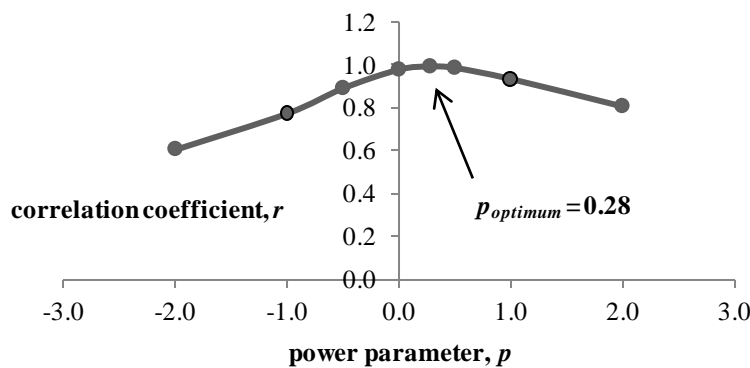
Table 2. Common data transformation models.

Model	Condition	$p \neq 0$	$p = 0$
Tukey	$y + c > 0$	$y^* = (y + c)^p$	$y^* = \ln(y + c)$
	$y > 0, c = 0$	$y^* = y^p$	$y^* = \ln(y)$
Box-Cox	$y + c > 0$	$y^* = [(y + c)^p - 1]/p$	$y^* = \ln(y + c)$
	$y > 0, c = 0$	$y^* = (y^p - 1)/p$	$y^* = \ln(y)$

With so many p values, which ones should be used? In most cases, the optimum value is determined recursively. Essentially, we select an initial p value within the $[-2, +2]$ interval, transform the data, construct a qq plot from it, and compute a correlation coefficient.

Said process is repeated for different p values, using the same original data. The p value that maximizes r is taken for the optimum value. Some authors maximize the coefficient of determination, R^2 , instead of r , but this is a matter of preferences, and does not change the results.

Figure 2 shows the results of maximizing r by applying the Box-Cox model to the data given in Table 1, where $c = 0$ and $r_{maximum} = 0.9932 \approx 0.99$ corresponds to $p_{optimum} = 0.28$. This result agrees with the likelihood method used at NIST (2012). The optimum transformation is one that is slightly closer to square roots than to logarithms.



p	r
2.00	0.8079
1.00	0.9353
0.50	0.9866
0.28	0.9932
0.00	0.9787
-0.50	0.8921
-1.00	0.7753
-2.00	0.6077

Figure 2. QQ plot-based optimization of Box-Cox power transformation parameter.

An advantage of this approach is that it is not limited to Box-Cox's model. Since a qq plot is constructed for different p values, we can estimate deviations from normality at each step of the recursive process. In addition, the qq plot allows one visualize outliers and how these are affected by the transformations.

Figure 3 shows the qq plot after transforming the data from Table 1, column D using $p_{optimum} = 0.28$. Compare with Figure 1.

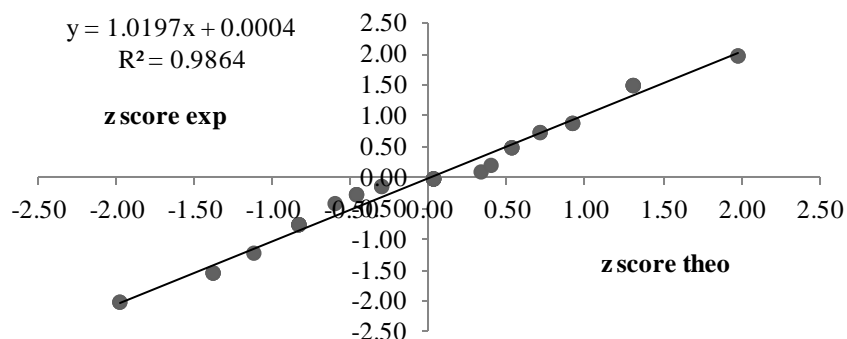


Figure 3. QQ plot after transforming Table 1, column D data using $p = 0.28$

Conclusion

In this tutorial, we showed how to construct qq plots using *Excel*, without the need for additional software. Customized spreadsheet functions were used. We also showed how the qq plots can be used to optimize the transformation parameters of variance-stabilizing transformation models.

Integrating a data transformation model to a qq plot tool using *Excel* reduces to programming the corresponding model into a different worksheet from the same workbook and referencing the transformed data into column D. In this second worksheet, you may want to call the p and c parameters from separate cells of the worksheet, so these can be easily modified.

In the second worksheet, setting $p = 1$ and $c = 0$ as default values lets you use the first worksheet as a standalone qq plot tool. Optionally, both worksheets can be combined into a single one.

As of 10-18-2016, readers interested in reproducing Table 1 can do so by downloading the following Excel file: <http://www.minerazzi.com/tutorials/quantile-raw-data-example.xlsx>.

References

Garcia, E. (2016). Box-Cox Power Transformations Tool. Retrieved from

<http://www.minerazzi.com/tools/power-transformations/box-cox.php>

Hossain, M. Z. (2011). The use of Box-Cox transformation technique in economic and statistical analyses. *Journal of Emerging Trends in Economics and Management Sciences* 2(1):32-39.

Retrieved from

<http://jetems.scholarlinkresearch.com/articles/The%20Use%20of%20Box-Cox%20Transformation%20Technique%20in%20Economic%20and%20Statistical%20Analyses.pdf>

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, Fourth Ed., Prentice Hall, Upper Saddle River, NJ. NIST (2012). What to do when data are non-normal.

Retrieved from

<http://www.itl.nist.gov/div898/handbook/pmc/section5/pmc52.htm>

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The Statistician* 41:169-178. Retrieved from

<http://staff.ustc.edu.cn/~zwp/teach/Reg/Boxcox.pdf>

Wikipedia (2012a). Quantile. Retrieved from

<http://en.wikipedia.org/wiki/Quantile>

Wikipedia (2012b). QQ Plot. Retrieved from

http://en.wikipedia.org/wiki/Q-Q_plot