Chapter 42
Quantile-Quantile Plots, a Good Start for Looking at Your Medical Data (50 Cholesterol Measurements and 58 Patients)

General Purpose

A good place to start looking at your data before analysis is a data plot, e.g., a scatter plot or histogram. It can help you decide whether the data are normal (bell shape, Gaussian), and give you a notion of outlier data and skewness. Another approach is using a normality test like the chi-square goodness of fit, the Shapiro-Wilkens, or the Kolmogorov Smirnov tests (Cleophas, Zwinderman, Chap. 42, pp 469–478, Testing clinical trials for randomness, in: Statistics applied to clinical studies 5th edition, Springer Heidelberg Germany, 2012), but these tests often have little power, and, therefore, do not adequately identify departures from normality. This chapter is to assess the performance of another and probably better method, the Q-Q (quantile-quantile) plot.

Specific Scientific Question

Are Q-Q plots of medical records capable of identifying normality and departures from normality. Random samples of hdl cholesterol and ages are used for examples.

Q-Q Plots for Assessing Departures from Normality

hdl cholesterol values (mmol/l)
3,80
4,20
4,27
3,70

(continued)

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_42

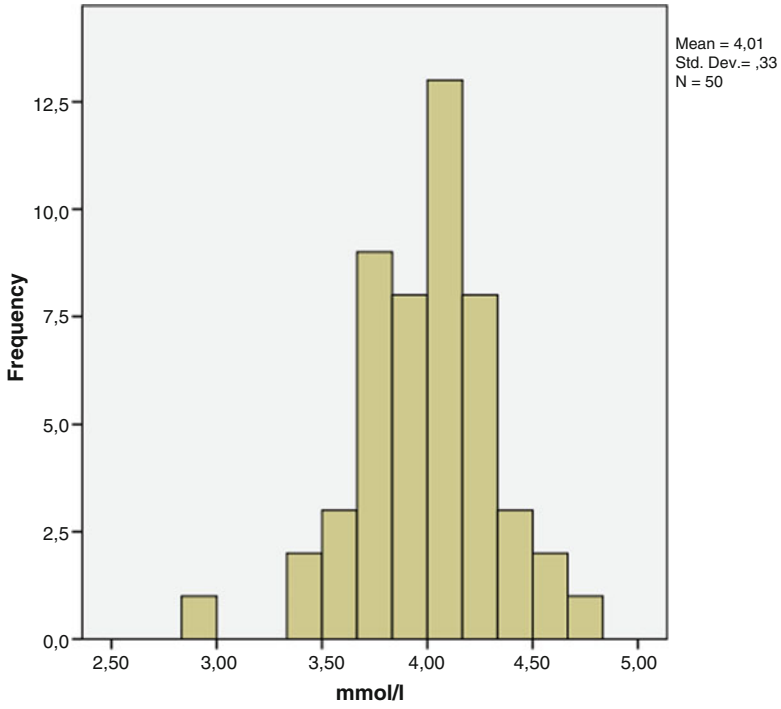
hdl cholesterol values (mmol/l)
3,76
4,11
4,24
4,20
4,24
3,63

The above table gives the first 10 values of a 50 value data file of hdl cholesterol measurements. The entire file is in the SPSS file entitled "q-q plot", and is available on the internet at extras.springer.com. SPSS statistical software is applied. Start by opening the data file in SPSS.

Command:

click Graphs....Legacy Dialogs....Histogram....Variable: enter hdlcholesterol....click OK.

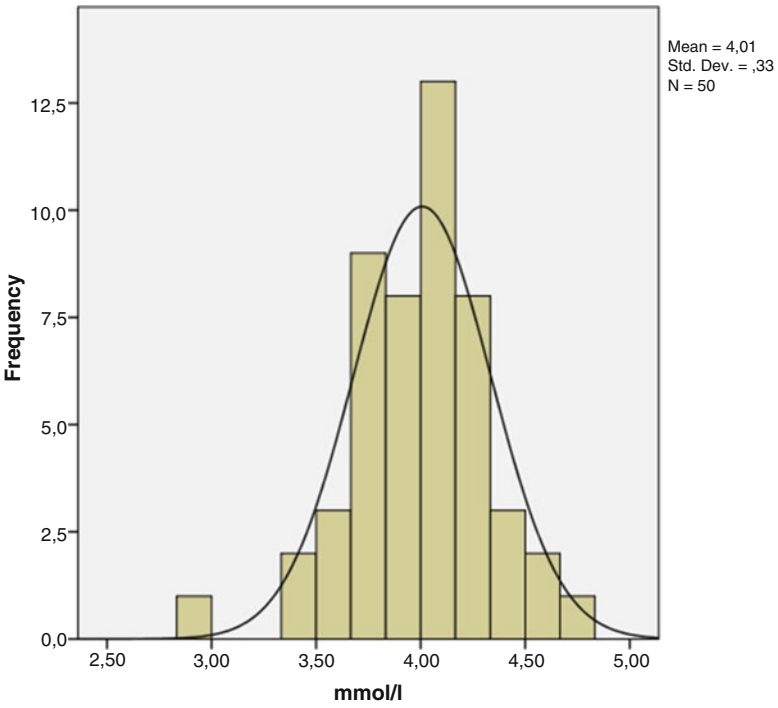
A histogram with individual hdl cholesterol values on the x-axis and "how often" on the y-axis is given in the output sheet: 50 hdl cholesterol values are classified in percentages (%) or quantiles (= frequencies = numbers of observations/50 here). E.g., one value is between 2,5 and 3,0, two values are between 3,0 and 3,5, etc. The pattern tends to be somewhat bell shape, but there is obvious outlier frequencies close to 3 mmol/l and close to 4 mmol/l. Also some skewness to the right is observed, and the values around 4 mmol/l look a little bit like Manhattan rather than Gaussian.



A Q-Q plot (quantile-quantile plot) can be helpful do decide what type of data we have here. First, the best fit normal curve is construed, e.g., based on the mean and standard deviation of the data. A graph of it is easy to produce in SPSS.

Command:

click Graphs....Legacy Dialogs....Histogram....Variable: enter hdlcholesterol .... mark: Display normal curve....click OK.

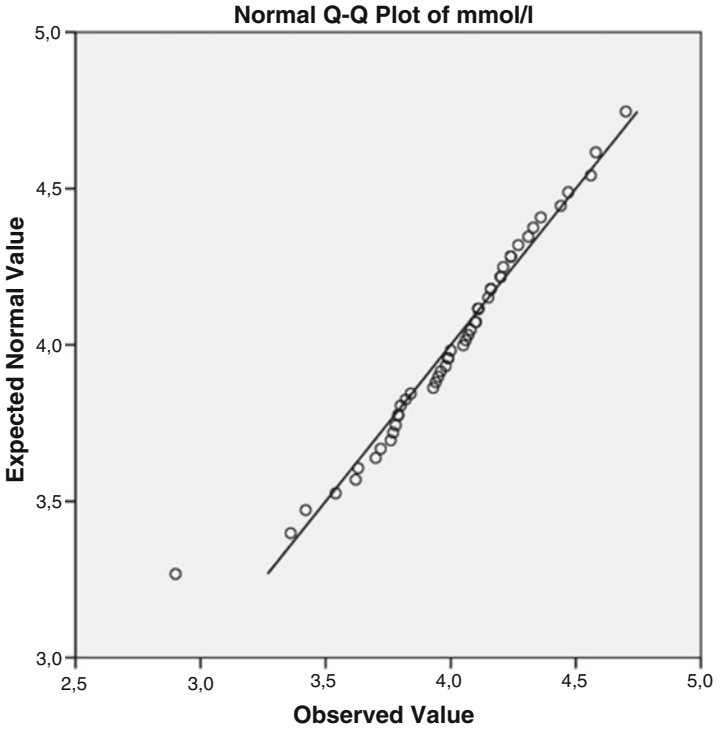


SPSS uses the curve to calculate the values for a Q-Q plot.

Command:

Analyze....Descriptive Statistics....Q-Q plots....Variables: enter hdlcholesterol.... click OK.

The underneath plot is construed of the observed x-values versus the same x-values taken from the above best fit normal curve. If our data perfectly matched the best fit Gaussian curve, then all of the x-values would be on the 45° diagonal line. However, we have outliers. The x-value close to 3 mmol/l is considerably left from the diagonal, and thus smaller than expected. The value close to 4 mmol/l is obviously on the right side of the diagonal, and thus larger than expected. Nonetheless, The remainder of the observed values vary well fit the diagonal, and it seems adequate to conclude that normal statistical test for analysis of these data will be appropriate.



Q-Q Plots as Diagnostics for Fitting Data to Normal (and Other Theoretical) Distributions

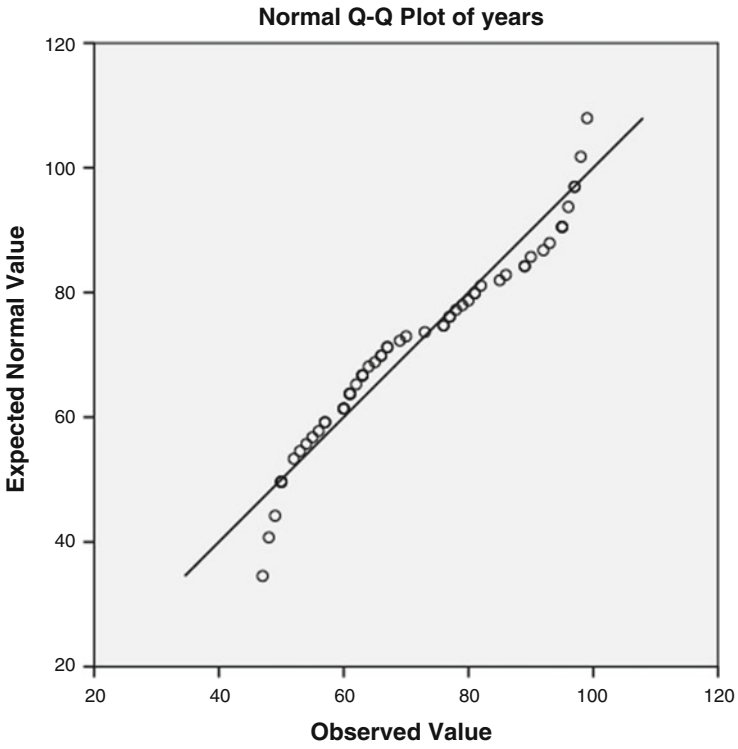
Age (years)
85,00
89,00
50,00
63,00
76,00
57,00
86,00
56,00
76,00
66,00

The above table gives the first 10 values of a 58 value data file of patients with different ages. The entire file is in the SPSS file entitled “q-q plot”, and is available

on the internet at extras.springer.com. SPSS statistical software is applied. Start by opening the data file in SPSS.

Command:

Analyze....Descriptive Statistics....Q-Q plots....Variables: enter age....click OK.



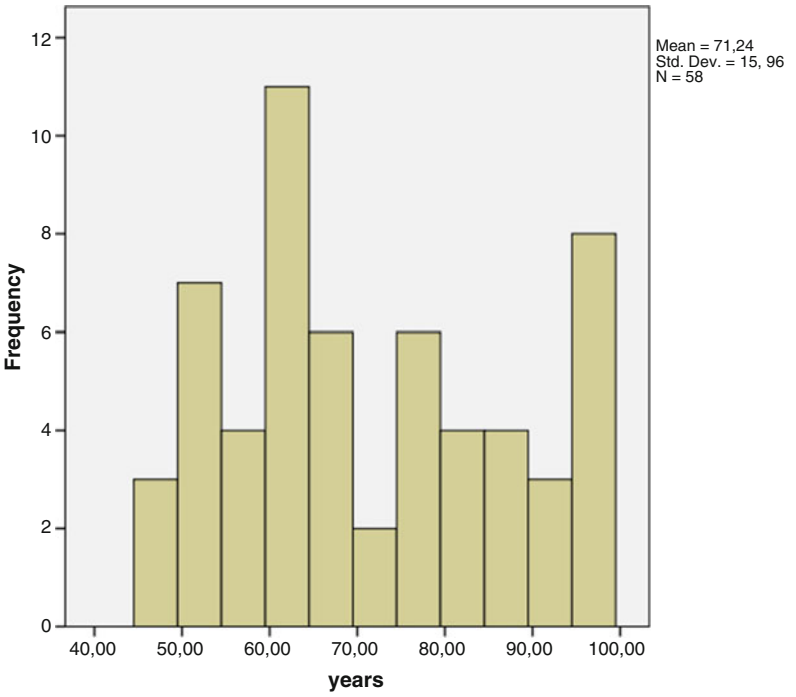
In the output sheets is the above graph. It shows a pattern with the left end below the diagonal line and the right end above it. Also the overall pattern seems to be somewhat undulating with the initially an increasing slope, and then a decreasing slope. The possible interpretations of these patterns are the following.

1. Left end below and right end above the diagonal may indicate a bells shape with long tails (overdispersion).
2. In contrast, left end above and right end below indicates short tails.
3. An increasing slope from left to right may indicate skewness to the right.
4. In contrast, a decreasing slope suggests skewness to the left.
5. The few cases with largest departures from the diagonal may of course also be interpreted as outliers.

The above Q-Q plot can hardly be assumed to indicate Gaussian data. The histogram confirms this.

Command:

click Graphs....Legacy Dialogs....Histograms....Variables: enter age....click OK.



The histogram given in the output sheets seems to confirm that this is so. The Q-Q plot method is somewhat subjective, but an excellent alternative to underpowered goodness of fit tests, and provides better information regarding normality than simple data plots or histograms do, because each datum assessed against its best fit normal distribution counterpart. We should add that SPSS and other software also offer the construction of Q-Q plots using other than normal distributions.

Conclusion

Q-Q plots are adequate assess whether your data have a Gaussian-like pattern. Non-Gaussian patterns and outliers are visualized, and often an interpretation can be given of them. The Q-Q plot method is similar to the less popular P-P (probability-probability) plot method, which has cumulative probabilities (= areas under curve left from the x-value), instead of the x-values on the x-axis and their expected counterparts on the y-axis. They are a little bit harder to understand.

Note

More background, theoretical and mathematical information of frequency distributions and goodness of fit testing is in the Chap. 42, pp 469–478, Testing clinical trials for randomness, in: Statistics applied to clinical studies 5th edition, Springer Heidelberg Germany, 2012, from the same authors as the current work.