Colloquium Biometricum 45
2015, 67–78

IMPACT OF ALTERNATIVE DISTRIBUTIONS ON QUANTILE-QUANTILE NORMALITY PLOT

Zofia Hanusz, Joanna Tarasińska

Department of Applied Mathematics and Computer Science
University of Life Sciences in Lublin
Głęboka 28, 20-612 Lublin
e-mails: zofia.hanusz@up.lublin.pl, joanna.tarasinska@up.lublin.pl

Summary

Many statistical programs use normal quantile-quantile plots for geometrical visualization of experimental data sets. In the paper we observe how normal quantile-quantile points behave when data sets come from alternative distributions, characterized by different kind of departure from normality. Theoretical lines for these distributions are presented in order to observe relationship between type of departure from normality and location points in a normal Q-Q plot. Additionally, the Shapiro-Wilk's test is applied to check normality of the data sets.

Keywords and phrases: normality, quantile, visualization

Classification AMS 2010: 62-09, 62F03, 62E15

1. Introduction

The most common assumption in many statistical procedures is the normal distribution of data. Generally, to check this assumption one of known statistical tests can be used. However, the graphical methods should not be neglected as they may indicate not only that the assumption is not satisfied, but also suggest a proper distribution of data sets. In the literature of the subject, there are many papers and books concerning fitting data to any known distribution, for

example, Anscombe (1973), Chambers et al. (1983), Cleveland (1985), Wilk and Gnanadesikan (1968), Stephens (1974), Thode (2002). One of the most frequently used graphical method is a quantile-quantile (Q-Q) plot. The Q-Q plot shows points where one coordinate is empirical quantile of the sample, i.e. sample order statistics, and the second one is a quantile of theoretical distribution, i.e. theoretical quantile.

In one-dimensional case of checking normality of data sets we consider normal Q-Q plot, where theoretical quantiles are from standard normal distribution. If the data set comes from normal distribution then points on the graph should be arranged along a straight line. Systematic deviation of points from that line indicates non-normality.

The aim of the paper is to show how samples from different distributions affect the position of the points on normal Q-Q plots. Some results of this kind for generated samples are given in Thode (2002). Here, we give exact theoretical lines for different alternatives. Additionally, for these alternatives we generate random samples of size 20 and 100 to see the effect of sample size on the plot. Moreover, for each generated sample, the Shapiro-Wilk's test (Shapiro and Wilk, 1968) is used to assess normality.

All simulations, calculation and Q-Q plots were done using R language (R Core Team, 2014).

2. Normal Q-Q plot

Let us consider random sample $Y_1, Y_2, \cdots, Y_n$ of independent and identical distributed variables with cumulative distribution function (cdf) F. According to Blom (1958), in normal Q-Q plot n points $P_i\left(\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right), Y_{(i)}\right)$ $\left(i = 1,2,\cdots,n\right)$ are plotted, where Φ^{-1} is i-th quantile of a standard normal distribution and $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$ denote order statistics. The theoretical quantile-quantile line $y = F^{-1}\left(\Phi\left(z\right)\right)$ for distribution F is presented in Figure 1.

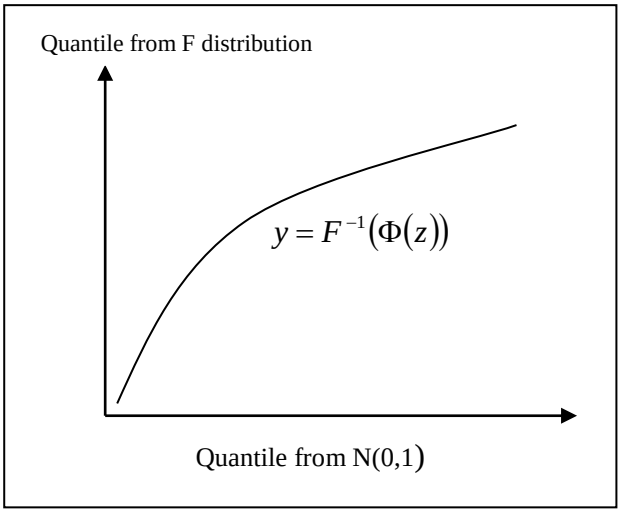

Fig. 1. Quantiles from distribution with cdf *F* against standard normal quantiles

If random samples come from normal distribution then the points P_i $(i = 1,2,\cdots,n)$ in normal Q-Q plot should fit the strait line $Y = \mu + \sigma Z$.

As an illustration, let us consider two samples of the size *n*=20 and *n*=100, generated from two normal distributions: standard normal and normal with mean equals 5 and variance equals 4. The normal Q-Q plot for sample from standard normal distribution and the theoretical line *y*=*z* are presented in Figure 2 (top), for $n = 20$ (left panel) and for $n = 100$ (right panel). Shapiro-Wilk's test applied for both samples gave $p = 0.7971$ and $p = 0.4749$, respectively. Therefore, for both samples normal distribution was not rejected at significance level 0.05.

For samples generated from $N(5,4)$ the points and the line *y*=5+2*z* on Q-Q plot are presented in Figure 2 (bottom). Applying Shapiro-Wilks's test we got $p = 0.5049$ for *n*=20 and $p = 0.3542$ for *n*=100. In both cases, normality was not rejected at significance level 0.05.

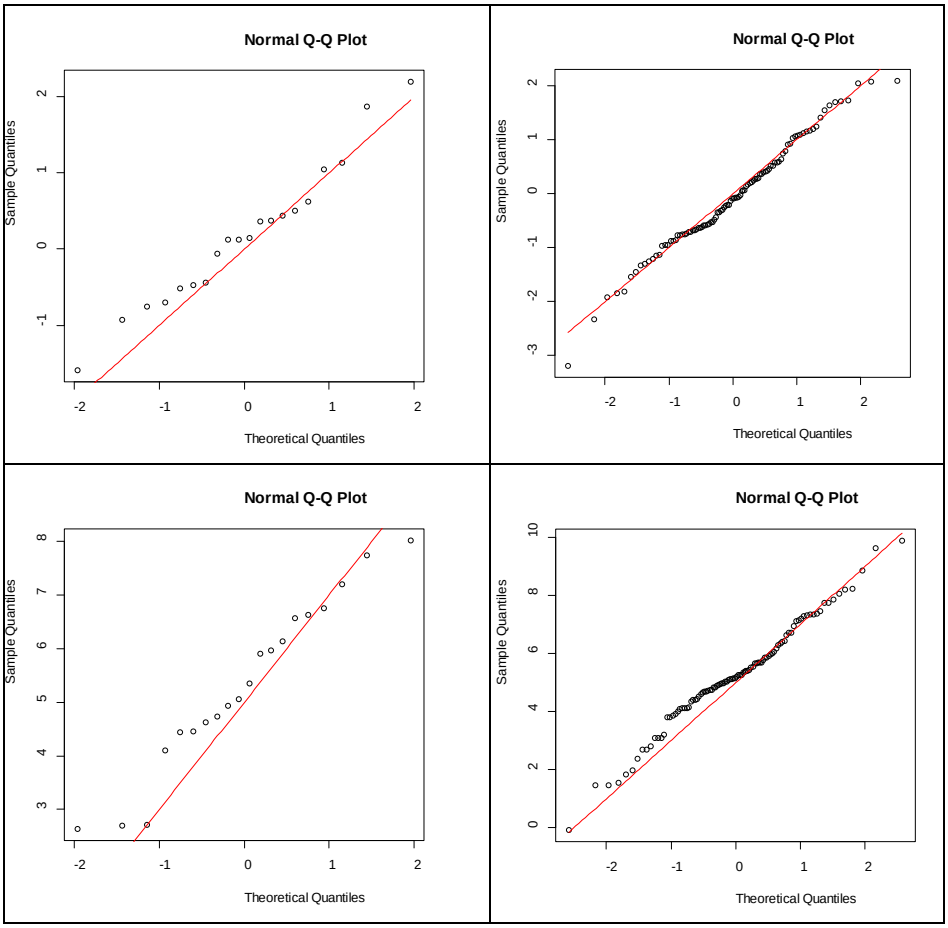

Fig. 2. Q-Q plots for samples generated from standard normal distribution (top) and from N(5,4) distribution (bottom) with size n=20 (left panel) and n=100 (right panel)

However, if random samples does not come from normal distribution, the points $P_i \ \left(i = 1,2,\cdots,n\right)$ in a normal Q-Q plot fit a line of some other shape than strait line. This shape can sometimes show what kind of departure from normality, for example skewness, kurtosis, multimodality we deal with.

In the next section we consider different alternative distributions and theoretical lines for them to observe relationship between type of departure from normality and location points in a normal Q-Q plot.

3. Impact of alternative distributions on normal Q-Q plot

In this section we consider different alternative distributions characterized by different type of departure from normality. We concentrate on impact of chosen distributions on the normal Q-Q-plot. The theoretical lines for alternative distributions $y = F^{-1}(\Phi(z))$ are given. Additionally, the sample of size 20 and 100 is generated and *p*-value of the Shapiro-Wilk's test for it is calculated in order to illustrate possible departure of points from the line.

3.1. Samples generated from symmetrical "light" tailed distributions

In this paper as a "light" tailed distribution we mean polykurtic distributions, i.e. with negative kurtosis. Let us remind that kurtosis for normal distribution is $\frac{\mu_4}{(\sigma^2)^2} - 3 = 0$. In this subsection we consider two types of distribution with "light" tails, namely, uniform distribution on the interval (0, 1) (kurtosis –1.2) and Beta distribution with shape parameters $\alpha = 2$ and $\beta = 2$ with the support on (0, 1) (kurtosis –6/7, Patel et al., 1976). The normal Q-Q plots for two samples from uniform distribution of the size n=20 (left panel) and n=100 (right panel) are presented in Figure 3 (top). We can also see the theoretical line of S-shape for this distribution. The Shapiro-Wilk's test for the generated sample of size n = 20 gave $p = 0.2096$ and normality is not rejected at significance level 0.05. For bigger sample n=100, the Shapiro-Wilk's test gave $p = 0.0197$ so normality is rejected at significance level 0.05.

The normal Q-Q plots for two samples from Beta (2,2) distribution of the size n=20 and n=100, respectively, together with theoretical line, are presented in Figure 3 (bottom). The *p*-value of the Shapiro-Wilk's test for generated sample of size n=20 is $p = 0.2265$ so normality is not rejected, while for n=100, $p = 0.0071$ so normality is rejected at significance level 0.05.

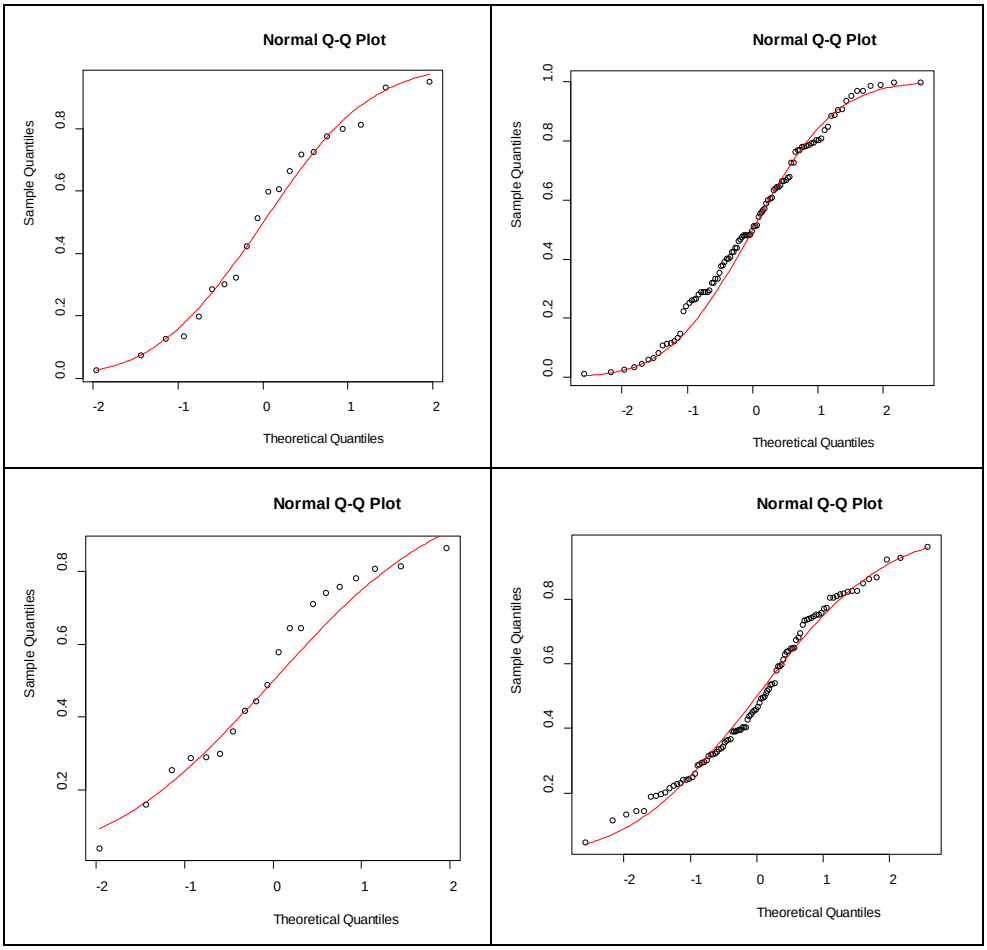

Fig. 3. Q-Q plots for samples generated from Uniform distribution on interval (0,1) (top) and Beta(2,2) distribution on support (0,1) (bottom) with size n=20 (left panel) and n=100 (right panel)

3.2. Samples generated from symmetrical "heavy" tailed distributions

As a "heavy" tailed distribution we mean distributions without finite variance and so called leptokurtic distribution with positive kurtosis. Let us consider Student t distribution with 1 (no finite variance) and 5 degrees of freedom (kurtosis equals 6). Theoretical lines together with quantiles points for samples generated from these distributions for 20 and 100 sizes are presented in Figure 4.

The Shapiro-Wilk's test for generated samples from Student t distributions with 1 degree of freedom with quantiles points in Figure 4 (top) gave

$p = 0.0010$ for n=20 and $p = 3.5\text{E}-15$ for n=100, so in both cases normality is rejected at significance level 0.05.

For samples generated from Student t distribution with 5 degrees of freedom with quantiles points in Figure 4 (bottom), the Shapiro-Wilk's test gave $p = 0.5541$ for n = 20 and did not reject normality at the significance level 0.05 while $p = 5.48\text{E}-05$ for n=100 caused the normality rejection.

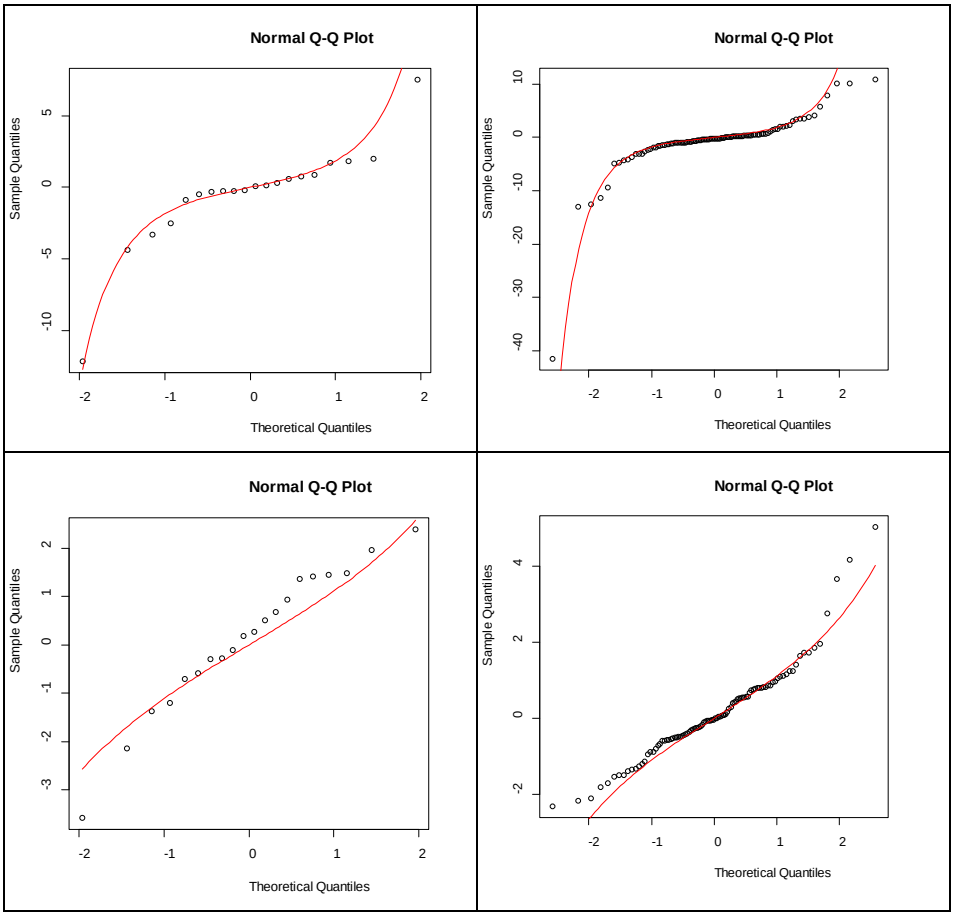

Fig. 4. Q-Q plots for samples generated from Student t distribution with 1 degree (top) and 5 degrees (bottom) of freedom, for size n=20 (left panel) and n=100 (right panel)

3.3. Samples generated from contaminated normal distributions

In this subsection we consider mixture of two normal distributions with proportion rate equals 0.5. In the first case we consider so called location contaminated normal, i.e. mixture of normal distributions with different means

and the same variance, namely, $N(0,1)$ and $N(5,1)$. The normal Q-Q plots for n = 20 (left panel) and n = 100 (right panel) are presented in Figure 5 (top). For n=20 the Shapiro-Wilk's test gave $p = 0{,}1305$ and normality was not rejected while for n=100, $p = 2.84\,\mathrm{E}-07$ and normality was rejected at the significance level 0.05.

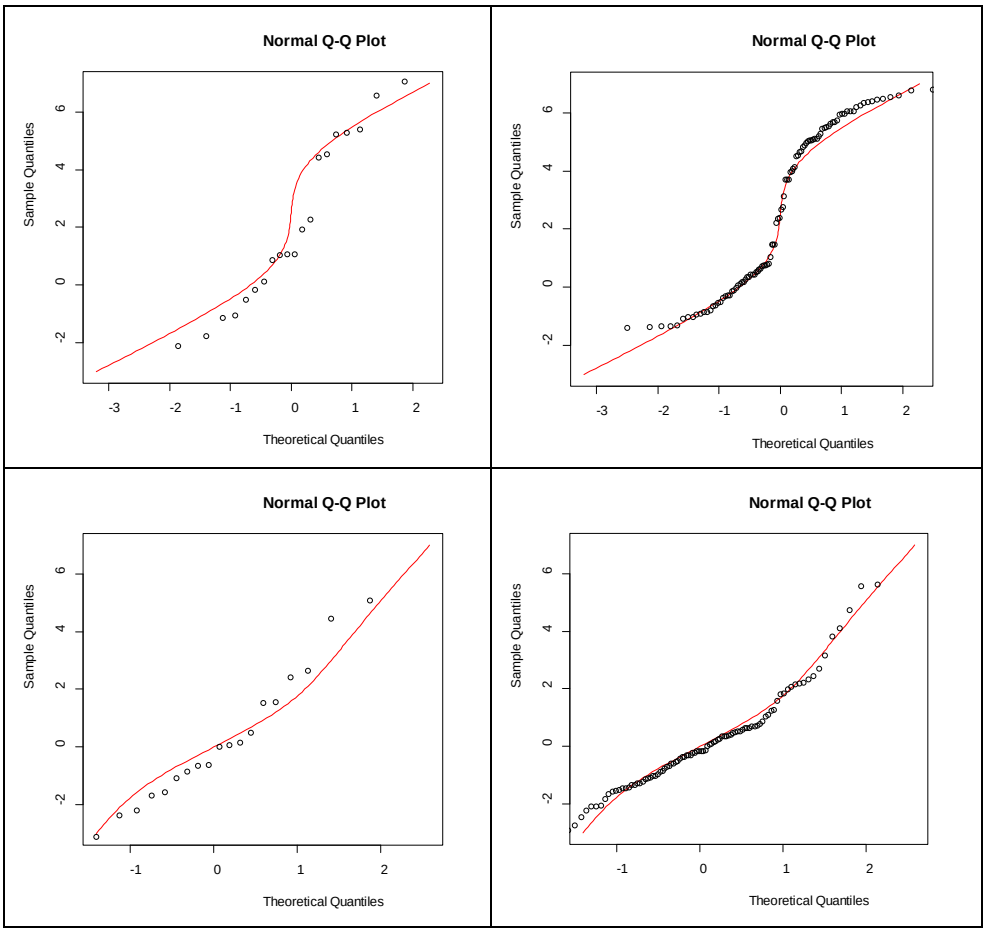

Fig. 5. Q-Q plots for samples generated from mixture of two normal distributions N(0, 1) and N(5, 1) (top) and mixture of N(0, 1) and N(0, 9) (bottom) for size n=20 (left panel) and n=100 (right panel)

The second case regards so called scale contaminated normal distribution (Tukey, 1960), i.e. mixture of two normal distributions with the same mean and different variances, namely, $N(0,1)$ and $N(0,9)$. The quantile points for n=20

(left panel) and n=100 (right panel) are presented in Figure 5 (bottom). For n=20 the Shapiro-Wilk's test gave $p = 0.3368$ and normality was not rejected but for n=100, $p = 8.45\text{E-}05$ and normality was rejected at the significance level 0.05.

3.4. Samples generated from asymmetric distributions

In this subsection we consider two kinds of asymmetry, depending on whether the distribution is skewed to the left (negative skewness) or to the right (positive skewness). Considering positive skewness we regard samples generated from chi-squared distributions with 1 and 3 degrees of freedom. The normal Q-Q plots for two samples from each distribution of the size n=20 (left panel) and n=100 (right panel), are presented in Figure 6.

In the case of samples generated from chi-squared distribution with 1 degree of freedom, p-value of Shapiro-Wilk's test was 8.52 E-05 for n=20 and 6.47 E-12 for n=100. In both cases normality was rejected at significance level 0.05.

In the case of samples generated from chi-squared distribution with 3 degrees of freedom (at the bottom of Fig. 6), p-value of the Shapiro-Wilk's test was 0.1121 for size n=20 so normality was accepted, for the sample of size n=100, $p = 5.11\text{E-}07$ so normality was rejected at significance level 0.05.

Samples with negative skewness were generated from noncentral Student t distribution with 4 degrees of freedom and negative noncentrality parameter –5 and Beta (5,2) distribution with shape parameters 5 and 2. The normal Q-Q plots for samples generated from noncentral t (top) and Beta (5,2) (bottom) of the size n=20 (left panels) and n=100 (right panels), are presented in Figure 7.

In the case of samples generated from noncentral t distribution with 4 degrees of freedom and noncentrality parameter –5, p-value of the Shapiro-Wilk's test was 0.0024 for size n=20 and 7.44 E-14 for n=100, so normality was rejected in both cases at significance level 0.05.

In the case of samples generated from Beta (5,2) distribution (at the bottom of Fig. 7), p-values of the Shapiro-Wilk's test were 0.2387 for size n=20 and 0.0072 for n=100. In the first case normality was not rejected and in the second one normality was rejected at significance level 0.05.

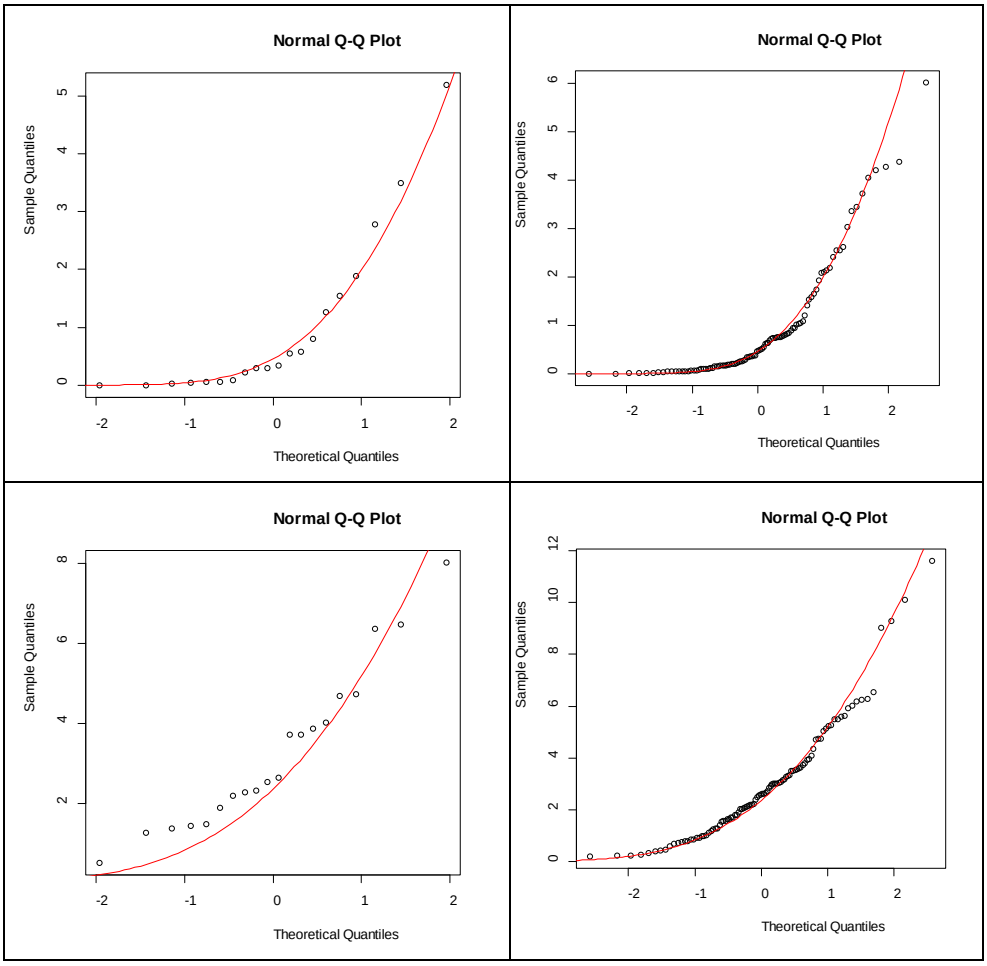

Fig. 6. Q-Q plots for samples generated from Chi-squared distribution with 1 degree of freedom (top) and 3 degrees of freedom (bottom) with size n=20 (left panel) and n=100 (right panel)

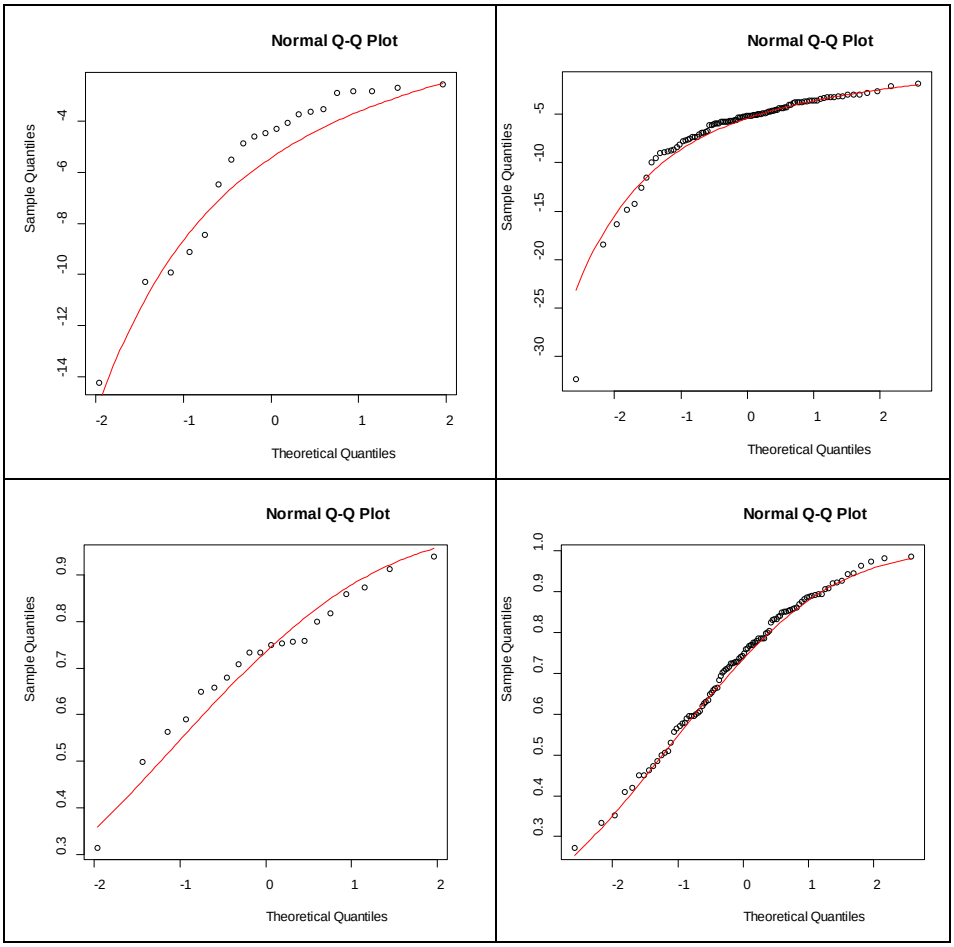

Fig. 7. Q-Q plots for samples generated from noncentral *t* distribution with 4 degrees of freedom and noncentrality parameter −5 (top) and Beta(5,2) (bottom) with size n=20 (left panels) and n=100 (right panels)

4. Conclusions

It is obvious that quantile-quantile points in the normal Q-Q plots form strait lines for samples generated from normal distributions.

For samples from different non-normal distributions considered in the paper we can observe that:

– quantile-quantile points in the normal Q-Q plot form lines of S-shape for distributions with “light” tails and conversely curved for “heavy” tails,

- for location contaminated normal (bimodal) the theoretical line has S-shape with clearly visible inflection point,
- right-skewed distributions form convex theoretical lines while left-skewed form concave lines.

However, it should be admitted that for small samples it is often difficult to discover non-normality on the basis of normal Q-Q plots.

References

Anscombe F. (1973). Graphs in Statistical Analysis. *The American Statistician*, 195-199.

Blom G. (1958). *Statistical estimates and transformed beta variables*. New York: John Wiley and Sons.

Chambers J.M., Cleveland W.S., Tukey P.A., Kleiner B. (1983). *Graphical Methods for Data Analysis*. Wadsworth & Brooks.

Cleveland W.S. (1985). *Elements of Graphing Data*. Wadsworth Publ. Co. Belmont, CA, USA.

Patel J.K., Kapadia C.H., Owen D.B. (1976). *Handbook of Statistical Distributions*. Marcel Dekker Inc. New York and Basel

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL http://www.R-project.org/.

Shapiro S.S., Wilk M.B. (1968). Approximations for the null distribution of the W statistic. *Technometrics* 10, 861-866.

Stephens M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association* 69, 730-737.

Thode H.C. Jr. (2002). *Testing for normality*. Marcel Dekker, Inc.

Tukey J.W. (1960). A survey of sampling from contaminated distributions. In: Olkin I., Ghurye S.G., Hoeffding W., Madow A.G., Mann H.B. eds. *Contributions to Probability and Statistics*. Stanford Univ. Press, CA.

Wilk M. B., Gnanadesikan R. (1968). Probability Plotting Methods for the Analysis of Data. *Biometrika* 5(5), 1-19.