

Probability plots as an instrument for comparing distributions

Holger Langkabel, Roche, Basel, Switzerland

ABSTRACT

In empirical studies it is often necessary to assess whether an assumed distribution reflects the given data appropriately. Maybe it is even necessary to develop a hypothesis about the data's distribution in the first place. Therefore, this paper wants to present how such tasks can be solved using probability plots (also referred to as quantile-quantile plots or Q-Q plots). By doing so, it is also presented how this method can be implemented in SAS. Additionally, certain analyses (e.g. in non-interventional studies) may require an assessment of the equality of distributions in treatment and control group. Also for this task, the use of probability plots is possible. However, there is no prefabricated SAS-procedure to create such a graphic (PROC UNIVARIATE is not able to apply this multivariate method). Therefore, this paper also presents a SAS-macro written by the author, which does the job.

INTRODUCTION

In empirical statistics researchers often encounter the problem that the true distribution of some continuous variable of interest is not known. On the other hand, an assumption on this distribution is also often necessary to perform inferential statistics. Therefore it is important to test for the validity of such a distributional assumption. One way to do so is to perform some formal statistical goodness-of-fit test which explicitly tests the assumption using objective inferential statistics. Another technique that might be used in conjunction or as a replacement are probability plots also referred to as Q-Q (quantile-quantile) plots. This is a graphical method which can be used to assess whether a distributional model fits the empirical data reasonably well.

Briefly summarized, this Q-Q plots use the fact that the empirically observed quantiles should be close to the theoretically derived quantiles of the assumed distribution. Therefore, plotting empirical against theoretical quantiles should result in predetermined patterns to be visible in the plot. Usually, probability plots are constructed in such a way that this pattern is a straight line. By the closeness to such a predetermined pattern, it can be assessed if the assumed distribution fits the empirical data or not.

A second application of Q-Q plots is the comparison of empirical quantiles in two distinct subgroups of the data (e.g. control vs. treatment group in a clinical trial), i.e. to compare empirical against empirical instead of theoretical quantiles. Doing so, it can be assessed whether the empirical distribution of some variable of interest is the same in both subgroups, i.e. whether the covariate is "balanced" between groups. In a usual randomized controlled trial, this will already be the case by virtue of the trial design itself. However, this might not be the case in non-interventional studies. Such studies are prone to intentional or even self-selection of patients into treatment and control group. As such selection might be based on covariates which are influencing outcome independent of treatment, it is important to check if these variables show the same empirical distribution across groups (e.g. that the distribution of BMI is the same across treatment and control group in a clinical trial on a new hypertension medication).

The remainder of this paper is organized as follows: The next section shortly treats some theoretical aspects of probability plots; thereby not going into too much mathematical detail. Section 3 presents how to create probability plots in SAS® using appropriate procedures and provides some examples of interpretation of probability plots. The concluding section presents the construction of Q-Q plots for the comparison of empirical distributions between groups. As this functionality is not provided by SAS, this section uses a macro developed by the author of this paper which is published in the appendix.

PROBABILITY PLOTS IN THEORY

To construct a probability plot, it is necessary to first rank the observations from smallest to largest according to their value. In mathematical notation, this means to re-arrange the sample $x_1, x_2, \dots, x_n$ to the ordered list of observations $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, where $x_{(1)}$ denotes the smallest value, $x_{(2)}$ the smallest but second value, and so forth with $x_{(n)}$ being the largest observed value in the sample. Now, consider a cumulative distribution function (cdf) $F(x)$ of some not further specified distribution. Then, the theoretical value of the cumulative distribution function at value $x_{(i)}$, $F(x_{(i)})$, (i.e. the theoretical probability for the occurrence of a value at most as large as $x_{(i)}$) would be expected to be close to its observed counterpart:

$$F(x_{(i)}) \approx \frac{i}{n}$$

(if n is sufficiently large). Here, i/n constitutes the estimated probability of observing a value at most as large as $x_{(i)}$ (because there are i observations with x equal or less than $x_{(i)}$ in the sample of n observations).

Re-arranging the above formula gives

$$x_{(i)} \approx F^{-1}\left(\frac{i}{n}\right)$$

where $F^{-1}(\cdot)$ is the inverse of the cumulative distribution function which delivers the quantiles of the distribution. Therefore, it can also be expected that the ordered values $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ are close to their corresponding theoretical quantiles $F^{-1}(1/n), F^{-1}(2/n), \dots, F^{-1}(n/n)$.

As a result of the second formula, it would be expected that a plot of the ordered values against their corresponding theoretical quantiles would scatter around a straight line passing through the origin with unity slope. Thus, a plotting instruction emerges which can be used for assessment of the suitability of a specific distribution model:

Plot instruction. Plot the ordered values of the sample against the hypothesized theoretical quantiles. If the hypothesized distribution is correct, the plotting points will scatter around a straight line that has slope 1 and passes through the origin.

By virtue of the re-arrangement used above, it can be seen that it would also be possible to plot the theoretical cdf against the empirical cdf. This is also a valid approach to assess the correctness of hypothesized distributions. This kind of plot is generally referred to as probability-probability plot (short: P-P plot).

PROBABILITY PLOTS IN PRACTICE

Let's have a look at some examples: Generating a dataset using the following code

```
data normal;
  call streaminit(1234);
  do i=1 to 100;
    normal1 = rand('normal',0,1);
    normal2 = rand('normal',10,4);
    exp = rand('weibull',1,0.5);
    output;
  end;
run;
```

the dataset WORK.NORMAL is created containing three variables with 100 observations each. The variable NORMAL1 is standard normally distributed, variable NORMAL2 has a normal distribution with mean 10 and standard deviation 4, and the variable EXP is exponentially distributed with $\lambda = 2$ (Please note that SAS does not offer an 'exponential'-option with the rand-function so that the exponential distribution has to be created as special case of the Weibull distribution.).

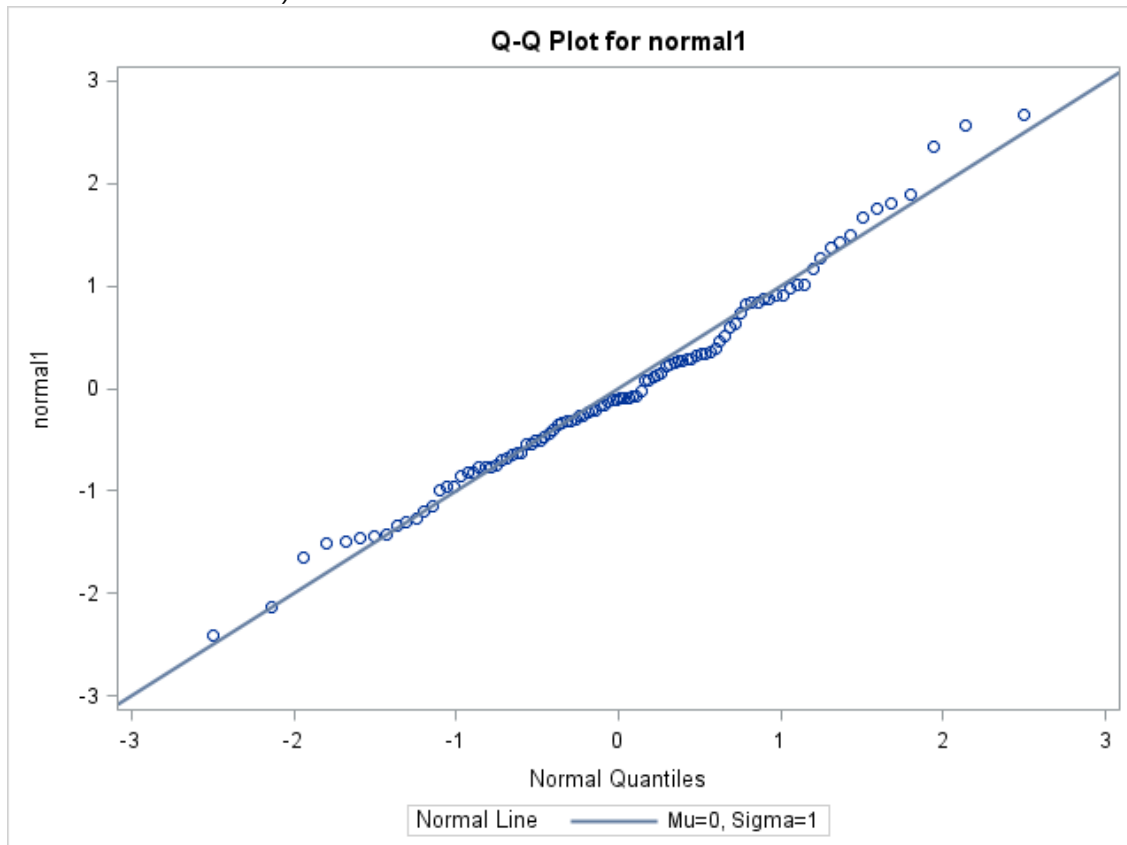


Figure 1: Q-Q plot for NORMAL1 assuming a standard normal distribution.

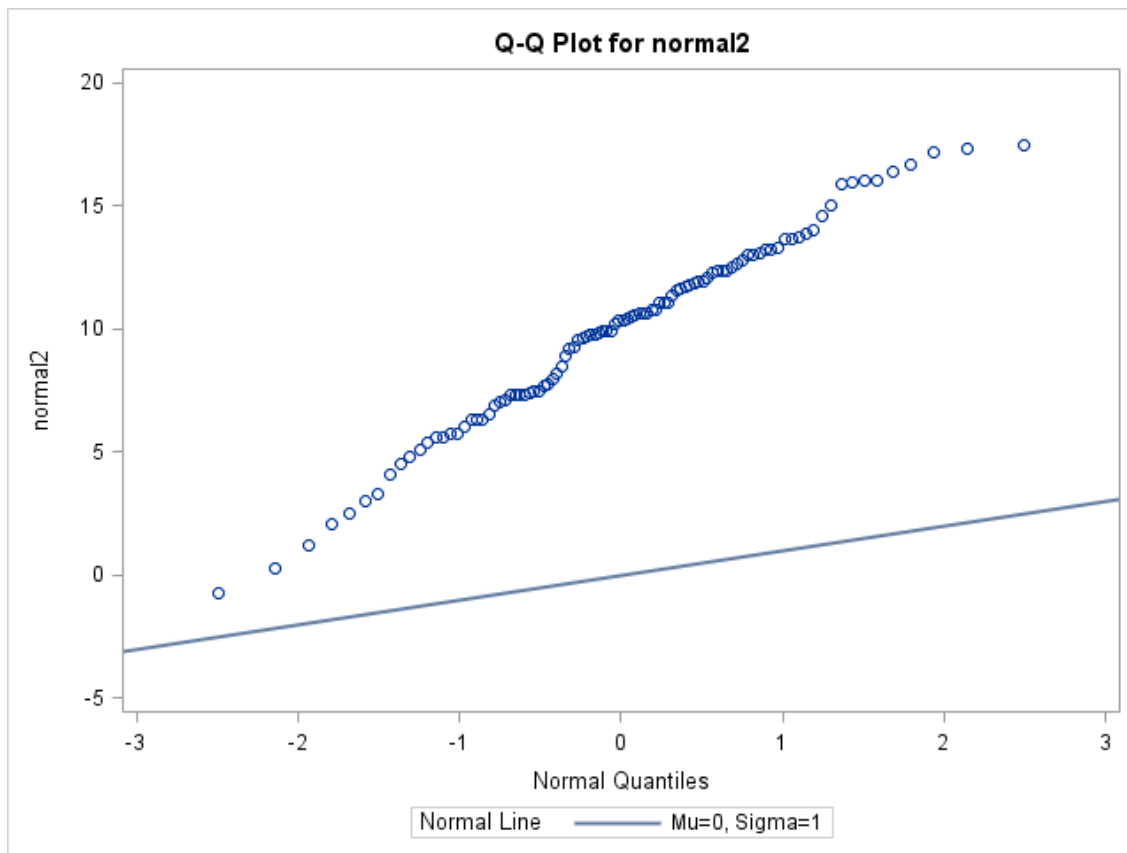


Figure 2: Q-Q plot for NORMAL2 assuming a standard normal distribution.

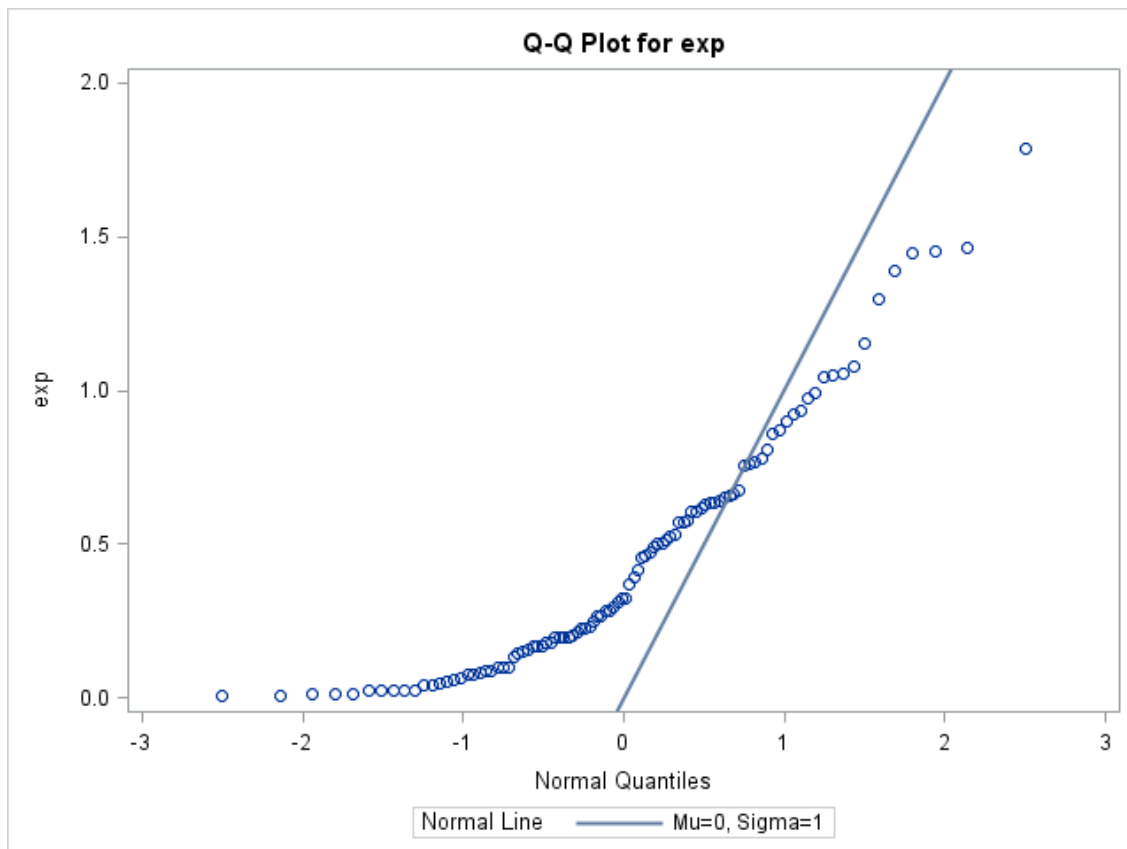


Figure 3: Q-Q plot for EXP assuming a standard normal distribution.

PhUSE 2016

SAS provides basic Q-Q plots in the UNIVARIATE-procedure. A simple call to the procedure using the defaults looks like this:

```
proc univariate data=normal noprint;
  qqplot normal1 normal2 exp / normal(mu=0 sigma=1);
run;
```

This results in Figure 1 through Figure 3. The `normal(mu=0 sigma=1)`-option is actually the default and would not have to be specified. But without specifying a comparison distribution, SAS does not plot reference lines. Figure 1 gives an example of a perfect result: the plotted points are not only representing a straight line, but they also scatter randomly around the reference line. This is of course due to the fact that the variable is truly standard normally distributed. Please note that it is not uncommon that the shape of the line gets a bit strewn at the tails of the distribution. Figure 2 shows a somewhat different picture: though the plotted points are still forming a rather straight line, they do not scatter around the reference line. Forming a straight line still indicates that the normal distribution is appropriate, while the difference to the reference line indicates that the parameters have been chosen incorrectly. As could be expected, the points in Figure 3 are not forming a straight line at all which is due to the fact that the variable is exponentially and not normally distributed.

Adjustments to Figure 2 could result in Figure 4 which compares the empirical quantiles of NORMAL2 against a normal distribution with mean 10 and standard deviation 4. Now, the points are scattering around the reference line. Analogously, variable EXP could be compared to an exponential distribution:

```
proc univariate data=normal noprint;
  qqplot exp / exponential(theta=0 sigma=0.5);
run;
```

The resulting Figure 5 shows a much better fit. Nonetheless, the correctness of the assumed distribution might be questioned by Figure 5 since the points for the higher quantiles are deviating quite much from the reference line. This fact points to the subjectivity of the method.

However, producing correct plots is rather easy if one already knows the true distribution. But once the suitability of a family of distributions has been assessed, it is possible to estimate the specific parameters out of the Q-Q plot; e.g. after assessing the suitability of the normal distribution in Figure 2 the parameters could have been estimated. This is illustrated in Figure 6, which shows a reference line for a normal distribution with the estimated parameters $\mu = 9.8668$ and $\sigma = 3.9224$ which are close to the true values of 10 and 4. This figure has been generated using the following code:

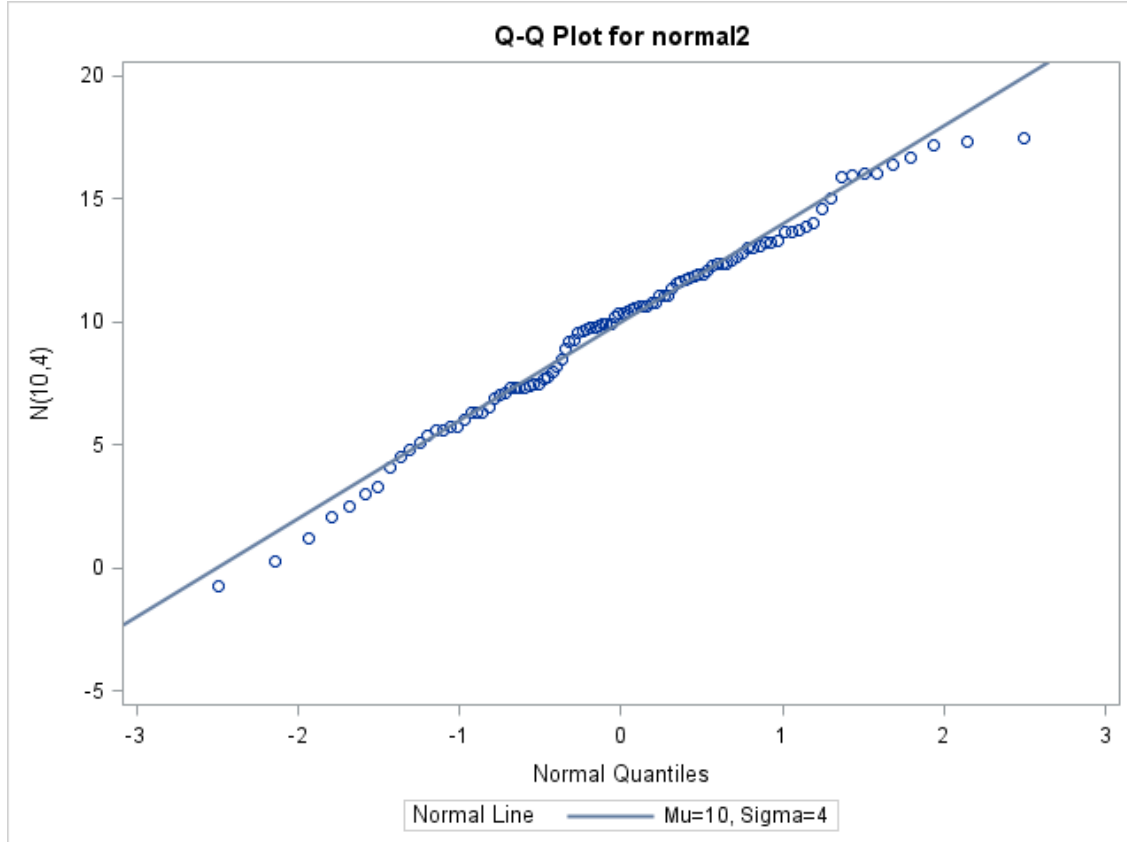


Figure 4: Q-Q plot for NORMAL2 assuming a $N(10, 4)$ -distribution.

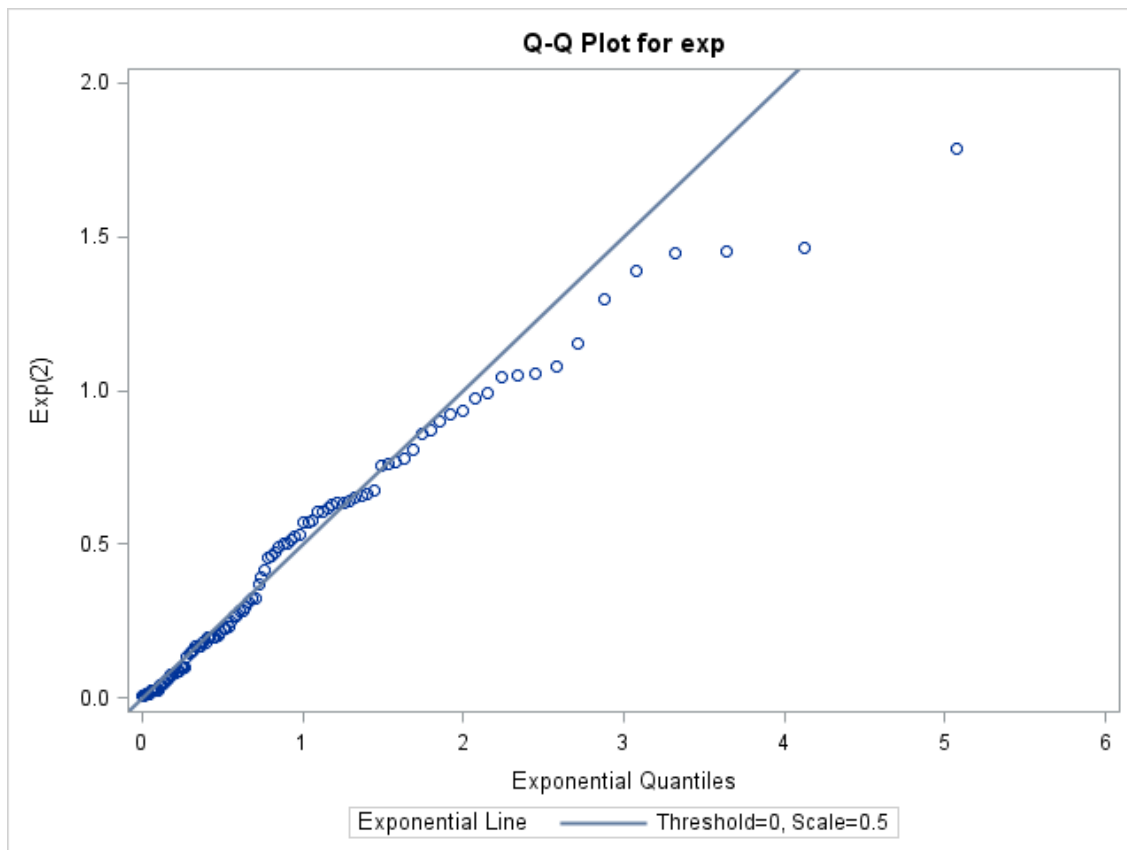


Figure 5: Q-Q plot for EXP assuming an exponential distribution ($\lambda = 2$).

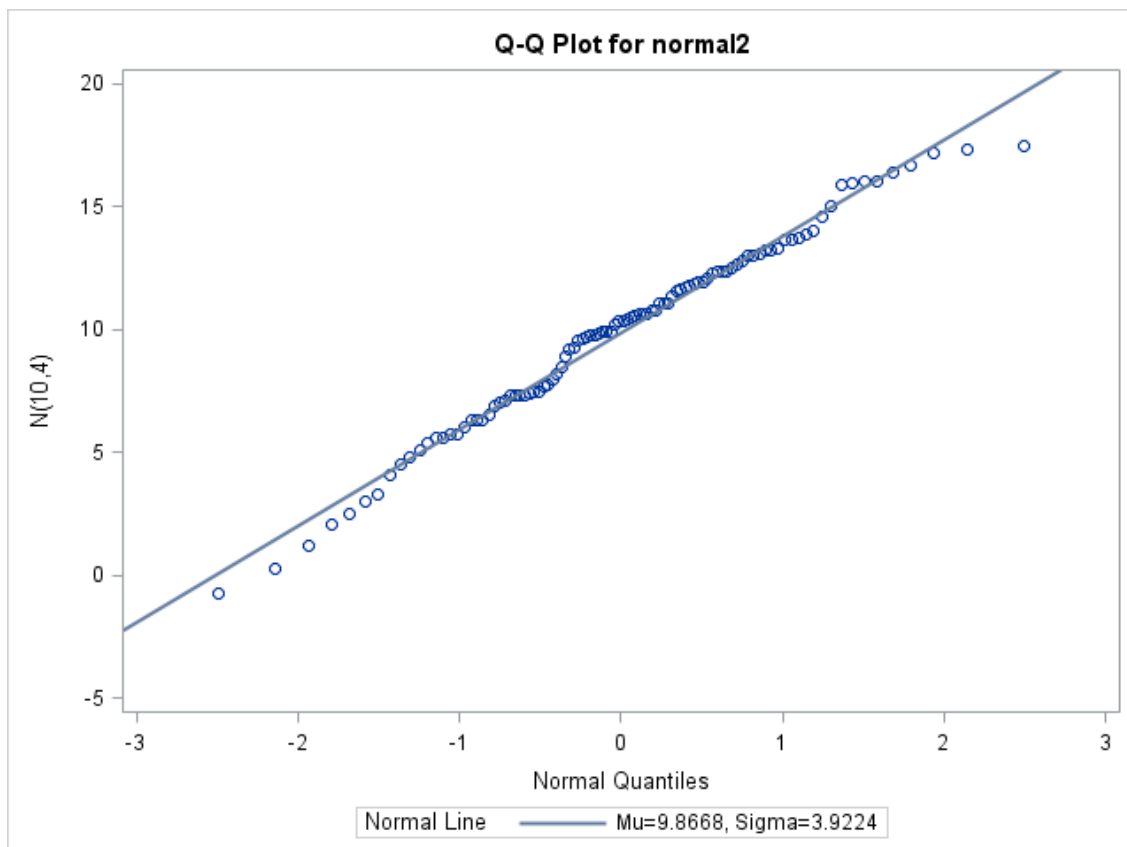


Figure 6: Q-Q plot for NORMAL2 assuming a normal distribution and estimating parameter values.

```
proc univariate data=normal noprint;
  qqplot normal2 / normal(mu=est sigma=est);
run;
```

For all of the distributions available in the UNIVARIATE-procedure (These are: beta, exponential, gamma, Gumbel, lognormal, normal, Pareto, power-function, Rayleigh, and two-parametrical as well as three-parametrical Weibull distributions.), it is possible to replace actual parameter values by the keyword “est”. If done so, SAS will estimate the parameter(s) from the data and print a reference line using the estimate(s).

Q-Q PLOT EMPIRICAL VS. EMPIRICAL

But what if the method is applied to empirical and not to simulated data? Let's try the SASHELP.BWEIGHT dataset of SAS 9.2. This data set provides 1997 birth weight data from the National Center for Health Statistics. Figure 7 shows a probability plot for the variable WEIGHT assuming a normal distribution, Figure 8 shows a probability plot assuming a Weibull distribution. Both plots show a rather good fit for the middle of the distribution, but perform very badly for the lower tail. Even the very flexible Weibull distribution doesn't seem to fit the data especially well.

But maybe you are not interested in the distribution of the birth weight variable itself, but you are rather interested in distributional differences across groups. E.g. you might want to evaluate the influence of a smoking mother on the infant birth weight. To do so you would either have to evaluate the distribution per group, i.e. to create two probability plots (one for smoking mothers and one for non-smoking mothers) – which might be very cumbersome since you do not know the underlying distribution in the first place – or you just compare the distributions in one probability plot because you do not have to compare an empirical distribution to a theoretical one but you could also put the quantiles of another empirical distribution on the x-axis – which also leads to a more straightforwardly interpretable and understandable graph.

Unfortunately, SAS does not provide a standard routine to create Q-Q plots for the comparison of two empirical distributions. For that reason, the appendix to this paper provides a SAS-macro which performs this task. It requires three parameters:

- `data`: the input dataset,
- `var`: the variable, which shall be compared across groups, and
- `class`: a binary character or numeric variable, which indicates group membership.

Internally, the macro is only calculating the quantiles of the variable of interest by group and then plotting these quantiles against each other using PROC SGLOT.

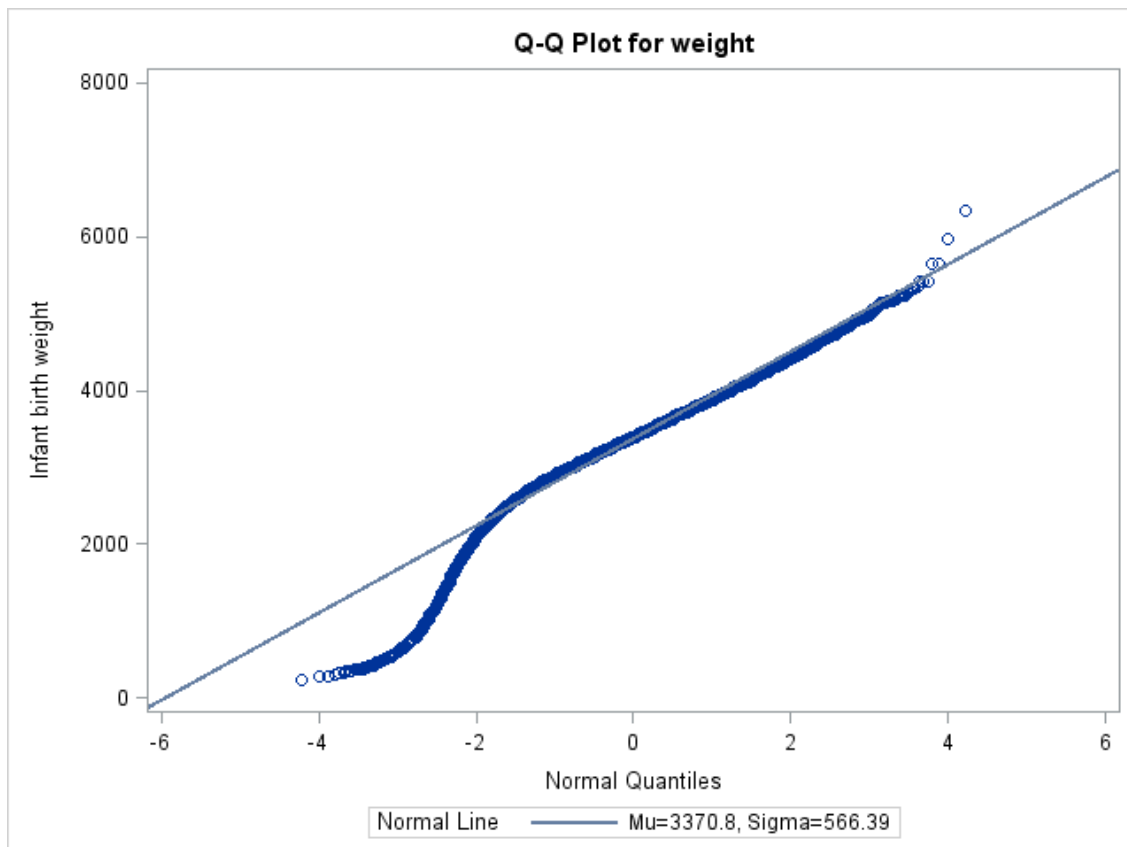


Figure 7: Q-Q plot for WEIGHT assuming a normal distribution and estimating parameter values.

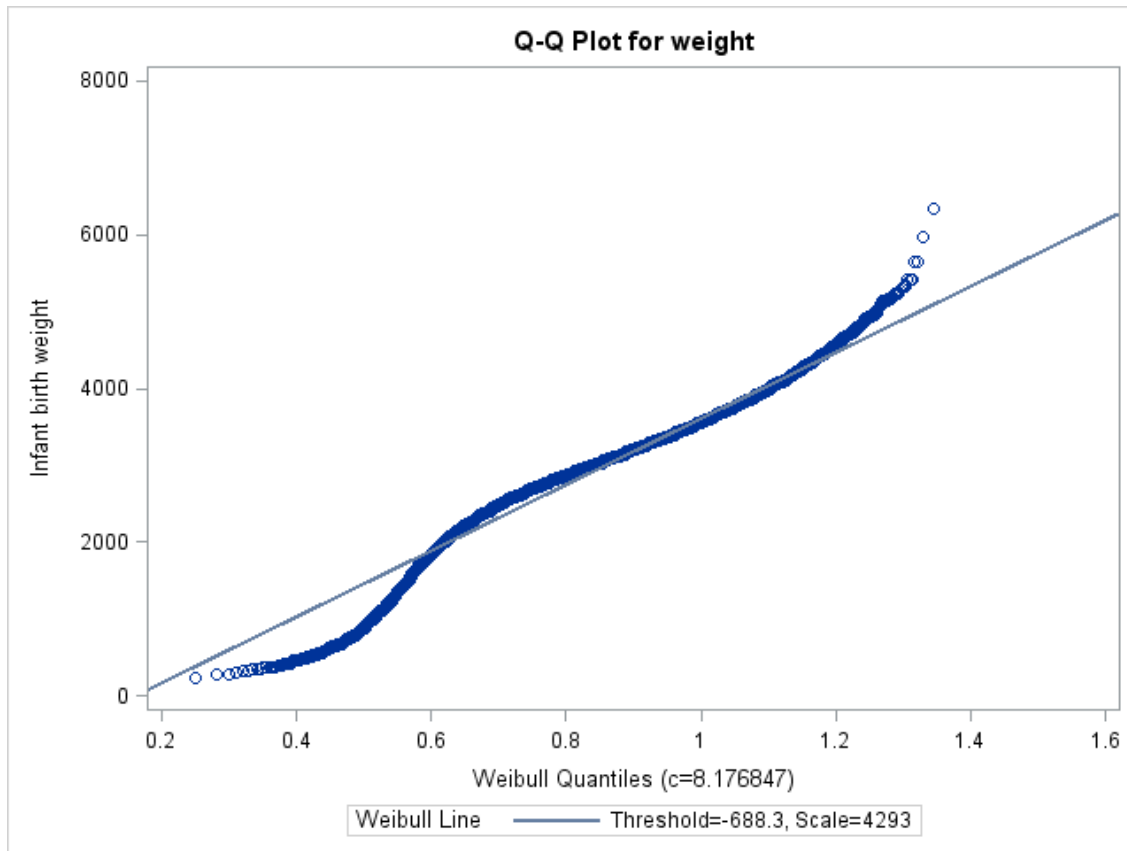


Figure 8: Q-Q plot for WEIGHT assuming a Weibull distribution and estimating parameter values.

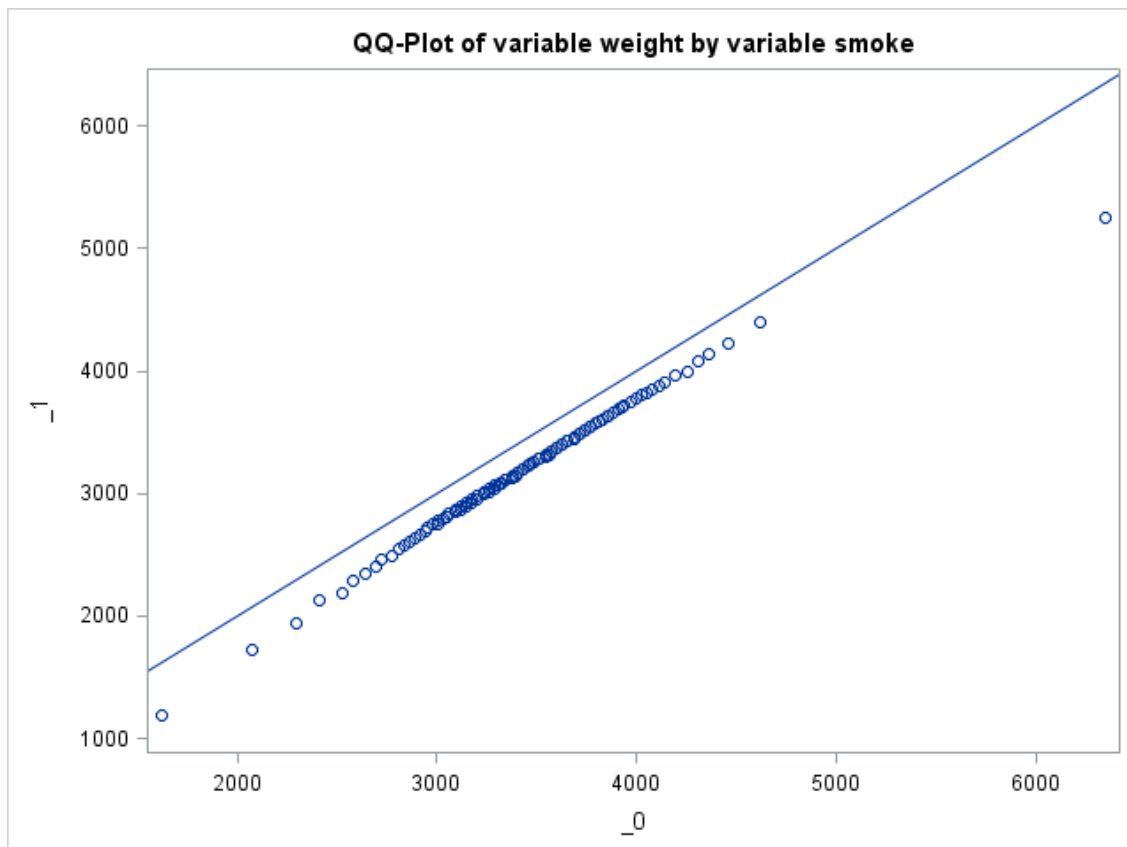


Figure 9: Q-Q plot for WEIGHT by values of SMOKE.

PhUSE 2016

Using this macro for our question results in the following macro call:

```
%qq(data=bweight, var=weight, class=smoke);
```

The resulting Figure 9 shows the quantiles of birth weight for non-smoking mothers on the x-axis and for smoking mothers on the y-axis. The reference line plotted is the first bisector. If the distribution was identical across groups, the plotting points would have to be scattered randomly around the reference line. This is obviously not the case for this data. Therefore, it can be concluded that the distribution of birth weight is not the same for smoking and for non-smoking mothers. However, since the plotting points still form a straight line, it can be concluded that the underlying distributional family is the same. The parallelism of the reference line and the plotting points indicates that both groups share the same underlying distribution with the same dispersion, only the level of birth weight is higher for non-smoking mothers for all quantiles (which is an expected result).

CONCLUSION

Quantile-quantile plots are a theoretically straightforward instrument to compare distributions. Briefly summarized, they are exploiting the to-be-expected similarity between quantiles of a theoretical distribution and quantiles of an empirical realization thereof. If theoretical assumption and empirical data (partially) correspond, certain patterns appear in quantile-quantile plots which can be used to assess the correctness of the presumed distribution. A special advantage of this method is the fact that it is a non-parametrical method because just the quantiles of two distributions are compared which doesn't necessitate any distributional assumption to be able to apply the method. This also results in the possibility to compare two empirical distributions directly without making any assumption on the underlying distribution. Quantile-quantile plots for the test of an empirical distribution against a theoretical one can easily be generated in SAS using PROC UNIVARIATE. For a comparison of two empirical distributions the macro provided in the appendix can be used.

The problems of the procedure are especially its subjectivity since no formal statistical test is performed and no strict rules for rejection of any hypothesis can be given. Furthermore, the method requires much experience from the user to yield meaningful results and interpretations. Nonetheless, quantile-quantile plots are a valuable instrument to arrive at hypothesis for further, more formal testing and an important mean to judge the validity of otherwise untestable assumptions.

REFERENCES

- J. M. Chambers, W. S. Cleveland, B. Kleiner, P. A. Tukey: Graphical Methods for Data Analysis. Wadsworth: Belmont 1983.
D. C. Montgomery: Statistical Quality Control: A Modern Introduction. John Wiley & Sons, Singapore 2013.

ACKNOWLEDGMENTS

This paper was prepared to a large extend during the author's employment at inVentiv Health Clinical.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Holger Langkabel
Hoffmann-La Roche Ltd.
Malzgasse 30
4052 Basel, Switzerland
holger.langkabel@roche.com

Brand and product names are trademarks of their respective companies.

APPENDIX

Note: While using this macro, please be aware that it has been developed under SAS 9.2 in a UNIX-environment.

```
%macro qq(data, var, class);  
    %local ngroups nobs value1 value2 increment;  
  
    proc sort data=&data.;  
        by &class.;  
    proc means data=&data. noprint;  
        var &var.;  
        by &class.;  
        output out=_n(where=(_stat_="N"));  
    run;  
  
    proc sql noprint;  
        select count(*), min(&var.), &class. into :ngroups, :nobs, :value1-:value2
```

PhUSE 2016

```
        from _n;
quit;

%if %sysfunc(anydigit(&value1.)) = 1 %then %let value1 = _&value1.;
%if %sysfunc(anydigit(&value2.)) = 1 %then %let value2 = _&value2.;

%if &ngroups ne 2 %then %do;
    %put ERROR: Macro qq: &class. does not have two distinct values.;
    %goto abort_macro;
%end;

* Choose number of displayed percentiles dependent on number of observations;

    %if &nobs. >= 100 %then %let increment = 1;
%else %if &nobs. >= 50 %then %let increment = 2;
%else %if &nobs. >= 25 %then %let increment = 4;
%else %if &nobs. >= 20 %then %let increment = 5;
%else %if &nobs. >= 10 %then %let increment = 10;
%else %if &nobs. >= 5 %then %let increment = 20;
%else %if &nobs. >= 4 %then %let increment = 25;
%else %if &nobs. >= 2 %then %let increment = 50;
%else %if &nobs. >= 1 %then %let increment = 100;

* Calculate percentiles;

proc univariate data=&data. noprint;
    var &var.;
    class &class.;
    output out=_qq1 pctlpts=1 to 100 by &increment. pctlpre=p;
    format &class.;
run;

* Prepare data;

proc transpose data=_qq1 out=_qq2;
    var p;;
    id &class.;
run;

* Plot;

proc sgplot data=_qq2 noautolegend;
    title "QQ-Plot of variable &var. by variable &class.";
    scatter x=&value1. y=&value2.;
    lineparm x=0 y=0 slope=1;
run;

proc datasets library=work;
    delete _n _qq1 _qq2;
run;

%abort_macro:
%mend qq;
```