An Introduction to Quantile-Quantile Plots for the Experimental Physicist

Allen J. Andersen and JR Dennison
Department of Physics, Utah State University, 4415 Old Main Hill, Logan, UT, USA 84322-4415

(Revised 11/2/2018 14:47:00)

An important question often encountered in experimental physics is, are two observables for the same independent variable related or not? Quantile-quantile (Q-Q) analysis compares the cumulative distributions of two sets of observations or one set of observations and a theoretical curve in a sophisticated and versatile statistical way that is easy to visualize. If the two observables follow the same trend, the associated Q-Q plot will be linear; if they are identical, the plot will have unity slope with zero intercept. A non-unity slope results from a scaling factor and a non-zero intercept results from an offset. Non-linear Q-Q plots indicate that the two observables do not follow the same trend. Examples given demonstrate that the Q-Q analysis method is applicable to a wide range of scenarios in experimental physics. While Q-Q analysis is presently not typically taught in physics curricula, it can prove a useful statistical tool for comparing experimental data sets.

I. INTRODUCTION

In experimental physics, measurements are made to test theoretical predictions. Ideally, diverse independent measurements will provide results consistent with either a confirmation or rejection of the theoretical model in question. However, in practice, laboratory measurements are rarely straightforward. Questions of measurement precision, accuracy, or inherently complex or stochastic systems frequently result in distributions of outcomes from repeated tests. One may also want to determine whether two distributions of different measurements are at all related. The typical university physics curriculum equips the experimentalist with many statistical tools for evaluating how well measurements are known and how well distributions of measurements follow a theoretical curve, including linear and non-linear regression. This paper does not seek to minimize the value of such tools, but rather add an additional statistical tool to the reader’s mathematical toolkit. The purpose of this paper is to offer a brief introduction to quantile-quantile (Q-Q) plots for physicists at a level that is accessible to most physics students. While Q-Q plots are a part of the statistician’s standard mathematical toolkit, they are not presented as part of the standard physics curriculum at any level. There are examples of Q-Q plots in the physics literature; however, the uninitiated physicist may not be familiar with them.

The authors’ primary research area is the interaction of spacecraft materials with the space plasma environment. Specifically, our tests explore electrostatic discharges of dielectrics under applied high electric fields. One of our experiment configurations records measurements of two distinct phenomena, one with a sample size roughly two orders of magnitude larger than the other. It became clear that if a relationship—not necessarily causal—between the two populations of measurements could be demonstrated, it would have important ramifications for our research. Measurements of the higher rate non-destructive tests could more efficiently predict the distribution of the lower-rate destructive tests. Our initial efforts to compare the two populations seemed to indicate a relationship, but lacked the clarity needed to draw well defined conclusions. Subsequent conversations with statisticians led us to Q-Q plots. Q-Q plots became a very useful graphical tool in our research efforts for comparing distributions of measurements.

II. CUMULATIVE DISTRIBUTION FUNCTIONS

An important prerequisite to preparing a Q-Q analysis is the selection of two cumulative distributions for comparison. This section briefly reviews the notion of cumulative distribution functions.

A probability density function (PDF) describes the likelihood of some event to occur as a function of an independent variable, \( x \), over a range of \( x \). A well-known example of a PDF is a Gaussian distribution. A cumulative distribution function (CDF), \( C(x) \), is simply the integration of a PDF, \( P(x) \), from a lower bond of the independent variable, \( x_{\text{min}} \), to some value of \( x \):

\[
C(x) \equiv \int_{x_{\text{min}}}^{x} P(x) \, dx \quad . \tag{1}
\]

For example, the corresponding normal CDF to a Gaussian PDF is related to the error function as

\[
\Phi(x) = \frac{2}{\sqrt{\pi}} \text{Erf}(x/\sqrt{2}) + 1 \quad .
\]

A useful feature of CDFs as probability distributions is that they always increase monotonically from zero to unity as a function of some variable of interest. A discrete CDF of data can be represented by an empirical cumulative distribution (ECD). Note that ECDs are a subset...
of CDFs. The ECD describes what fraction of a population of events has occurred up to a given value of independent variable. For large discrete sample size \( j \), the likelihood of occurrence \( P \) as a function of some variable \( V \) is

\[
P(V) = \frac{1}{j} \sum_{i=1}^{j} \mathbb{1}\{x_i \leq V\}
\]

where

\[
\mathbb{1}\{x_i \leq V\} = \begin{cases} 
1 & \text{if } x_i \leq V \\
0 & \text{otherwise}
\end{cases}
\]

(3)

Two such CDFs are plotted together in Fig. 1. It can be useful to know if the ECD of a data set follows some known empirical or physics-based trend. Alternatively, it may be useful to know if two populations of observables follow the same trend. This may not be obvious from the ECD plot due to relative shifts or relative scaling factors between the two distributions.

III. Q-Q PLOTS

Q-Q plots directly compare two CDFs. Figure 2 is the Q-Q plot corresponding to the two ECD shown in Fig. 1. For each value of cumulative probability, or quantile (the vertical axis in Fig. 2), there are two corresponding values of a variable of interest (horizontal axis in Fig. 2), one from each CDF. One CDF provides the \( x \)-values of the resulting Q-Q plot, while the other CDF provides the \( y \)-axis values. Each \( x-y \) pair in the Q-Q plot corresponds to the same quantile value.

It is evident that for any two identical CDFs, the resulting Q-Q plot would follow \( y = x \). Q-Q plots following a linear trend, such as Fig. 2, are an indication that the two populations are correlated. Unity slope indicates a linear relation between the CDF. Non-unity slopes reflect a relative scaling factor between the distributions. A non-zero intercept indicates that one population is shifted by a constant offset relative to the other. See, for example, Fig. 3 which is explained further in Section IV.

A non-linear Q-Q plot is an indication that the CDFs follow different trends. The two Q-Q plots shown in Fig. 3 are examples of Q-Q plots that do not follow a linear trend and therefore indicate the underlying CDFs are not correlated. These Q-Q plots are clearly not well approximated by a linear fit. In these cases, all four CDFs are also ECDs.

In both plots in Fig. 3, the independent variables that correspond to the ECDs are normalized to 1 at the maximum observed values at which the cumulative probability is unity.
This eliminates any relative scaling factor, such as may result from a choice of units. Q-Q plots can be used to compare any two CDFs. If two CDFs in question share a common variable, such as in Fig. 1, the Q-Q plot compares not only the shapes of the two CDFs, but their relative response to the common variable of interest. Use of normalized independent variables also facilitates comparisons of CDF like those shown in Fig. 1, even when different independent variables are used for the two CDF.

Might add the section here from the end of the paper on log-log plot analysis.

For experimentalists investigating whether two populations of measurements might be related, the 2-sample Q-Q plot can be a valuable mathematical tool. 2-sample Q-Q plots compare the quantiles of the ECDs of two populations. For populations of different sample size, interpolation within one or both data sets is required so that Q-Q x,y pairs can be determined at the same quantile values; such interpolation is most often done for the less dense data set. Note that the maximum number of quantiles that can be plotted will correspond to the sample size of the smaller population, unless this data set is interpolated.

A 1-sample Q-Q plot compares one population of data to a known function. Most physicists will already have tools for evaluating the goodness of a fit of a model function, such as a \( \chi^2 \) test; however, Q-Q plots can provide a convenient visual representation of the goodness of a fit to a data set. A 0-sample Q-Q plot comparing two known functions is of course straightforward to create, but is unlikely to be very useful except perhaps for demonstration purposes.

The advantage of the Q-Q plot method is that it results in a non-parametric plot that is easy to interpret qualitatively—if the distributions are correlated, the Q-Q plot will be linear; otherwise, it will not. The drawback is that for a two-sample Q-Q plot, quantifying the results becomes more complicated than a simple linear correlation, especially for a Q-Q plot comparing two data sets rather than a single data set to a known distribution function. The linear correlation coefficients for Q-Q plots, such as in Fig. 4, provide a qualitative way to compare relative agreement between two Q-Q analyses. The higher correlation coefficient for the blue curve in Fig. 4(c) \( (r=??) \) than for the red curve \( (r=??) \), clearly indicates the Lorentzian fit is superior; but it cannot easily quantify how much better the fit is. The linear correlation coefficients are not readily converted to absolute measures of the quality of a fit, such as confidence limits. Calculating a linear correlation coefficient gives artificially good results due to the sorting in Eq. 2 required when creating ECDs for the Q-Q plot, even for Q-Q plots that clearly deviate from linear.

The discussion in this introduction to Q-Q plots, is limited to presenting the qualitative advantages of using Q-Q plots. It is obvious which of the two Q-Q plots superimposed in Fig. 4(c) provide a better model to measured data. Q-Q plots can be quantified using statistical confidence limits such as the K-S statistic; however, one should bear in mind that confidence limits are not necessarily a reflection of the measurement uncertainty. The interested reader can consult references [7-10] regarding confidence intervals and quantifying Q-Q plots.

IV. EXAMPLE

In Section III, Q-Q plots were presented using 2-sample examples. 2-sample Q-Q plots will reveal whether two data sets follow the same distribution. What that distribution is will not be identified by the Q-Q plot, but there is also no need to assume any functional form of the underlying distribution.

As an additional example of how Q-Q plots may be used, consider the following illustrative problem. Suppose an experiment results in a distribution of data and one wants to compare fits with both a Gaussian function and Lorentzian fits. (b) Q-Q plot comparing the fitting functions to each other (c) Q-Q plots comparing the data to each fitting function.
\[ f(x) = Ae^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \]  
\[ f(x) = B \left[ \frac{1}{(x-x)^2+\gamma^2} \right]. \]

In Fig. 4 (a), the PDFs are shown. One can compare the two fitting functions to see that they are indeed significantly different, based on the corresponding 0-sample Q-Q plot. Figure 4 (b) is a Q-Q plot comparing the two fitting functions. The obvious nonlinearity indicates that they are clearly different functions, as expected.

Of course, one would be more interested in which fit is better. Figure 4 (c) shows two 1-sample Q-Q plots, comparing the data to each fit. Given that the Lorentzian Q-Q plot is more linear and closer to \( y = x \), it is clearly a better fit than the Gaussian fit. This is apparent from even a cursory visual comparison of the Q-Q plots.

V. CONCLUSIONS

Distribution functions are found in many branches of physics including quantum mechanics, statistical mechanics, plasma physics, and others. Quantile-quantile (Q-Q) analysis is not part of the standard physics curriculum; however, it is a useful statistical tool for comparing any two distributions. Q-Q plots are an easy-to-visualize representation of the relationship between any two distributions. For the experimental physicist, Q-Q plots are especially useful for comparing different populations of measurements. Q-Q plots also can compare data to a fitting function.

ACKNOWLEDGEMENTS

This work was supported by a NASA Space Technology Research Fellowship. We also acknowledge the helpful discussions we had with David Brown, Richard Lambert, and Mark Riffe.

6. A. Andersen, Utah State University, 2018.