ggplot2 compatible Quantile-quantile plots in R

by Alexandre Almeida, Adam Loy, Heike Hofmann

Abstract An abstract of less than 150 words.

TODO:

- Abstract
- Intro
- Review Q-Q plots and P-P plots, including other arrangements, and what is implemented in other packages
- Package implementation
- Give examples
  - Heike: BRFSS example
- Conclusion

Introduction

Univariate distributional assessment is a common thread throughout statistical analyses during both the exploratory and confirmatory stages. When we begin exploring a new data set we often consider the distribution of individual variables before moving on to explore multivariate relationships. After a model has been fit to a data set, we must assess whether the distributional assumptions made were reasonable, and if they are not we then must understand the impact this has on the conclusions. Graphics provide arguably the most common way to carry out these univariate assessments. While there are many graphical methods that can be used for distribution exploration and assessment, probability plotting is one of the most common graphical approaches used.

Probability plotting refers to a family of methods based on the cumulative distribution function (CDF), most notably quantile (Q-Q) plots and probability (P-P) plots (Wilk and Gnanadesikan, 1968). In this paper, we focus on comparing an empirical distribution to a theoretical distribution. Let $Y_1, \ldots, Y_n$ denote a random sample from an unknown population, and let $\widehat{F}_y(q)$ be the empirical cumulative distribution obtained from the sample. Further, let $F(q)$ denote the CDF of a proposed dsitribution for the sample. A Q-Q plot is constructed by plotting the quantiles of the empirical distribution, $q_y(p) = F_y^{-1}(p)$, against the corresponding quantiles of the theoretical distribution, $q(p) = F^{-1}(p)$. This constuction is illustrated in Figure 1. A P-P plot is constructed by plotting $F(q)$ against $\widehat{F}_y(q)$ for various quantiles, q. This constuction is illustrated in Figure 2. Regardless of the plot constructed, if the two distributions are identical, then the scatterplots will be linear with slope 1 and intercept 0. Additionally, Q-Q plots are invariant to linear transformations, so if two random variables differ by a linear transformation a Q-Q plot showing draws from their distributions will still be linear, but with a different slope and intercept, as seen in Figure 1. P-P plots, in turn, are sensitive to linear transformations.

While the basic form of both the Q-Q and P-P plots is a scatterplot, additional graphical elements are often added to aid in distributional assessment. For Q-Q plots, a reference line is often drawn through the points $(q(.25), q_y(.25))$ and $(q(.75), q_y(.75))$. For P-P plots a reference line with slope 1 and intercept 0 is used. In both plots, pointwise or simultaneous confidence bands are often added around the reference line to further aid in the visual assessment.

Innovations to Q-Q and P-P plots have also been proposed. Loy et al. (2016) discuss the creation of detrended Q-Q plots, where the y-axis is changed to show the difference between q_y and the reference line. Consequently, the line representing the agreement with the theoretical distribution is the x-axis. Loy et al. (2016) find that detrended Q-Q plots are more powerful than other designs, so long at the y-axis limits are set so that the aspect ratio is kept the same as in the traditional Q-Q plot. In reliability and survival analysis, probability plots often refer to a hybrid probability plot, there the CDF of the proposed theoretical distribution is plotted against the empirical order statistics, and transformations are applied to each axis to linearize the CDF (cf. Meeker and Escobar, 1998, chapter 6). This hybrid probability plot is invariant to linear transformations.

Q-Q plots have been implemented in various forms in R, but none provide a complete implementation of the probability plotting framework. Normal quantile plots, where a sample is compared to the standard normal distribution, are implemented using the `qqplot` and `qqline` in **base** graphics (R Core

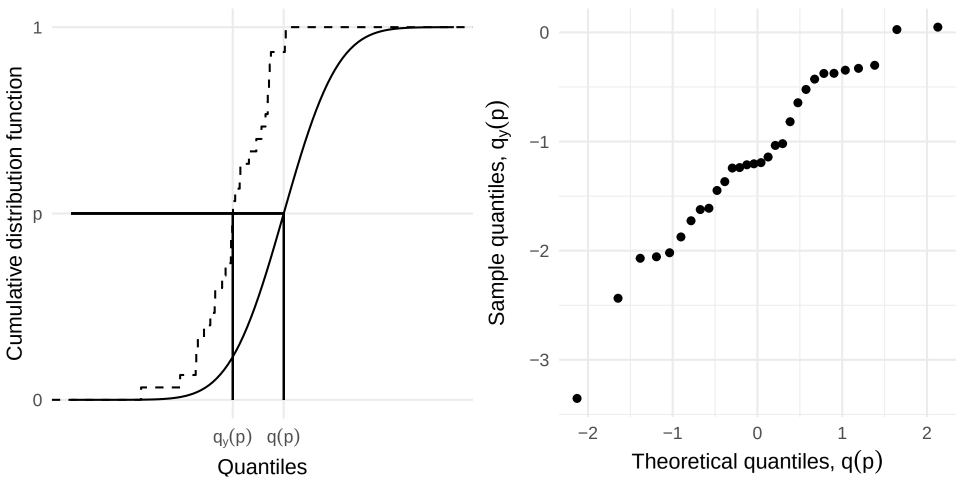

Figure 1: Illustrating what quantities are being plotted for Q-Q plots.

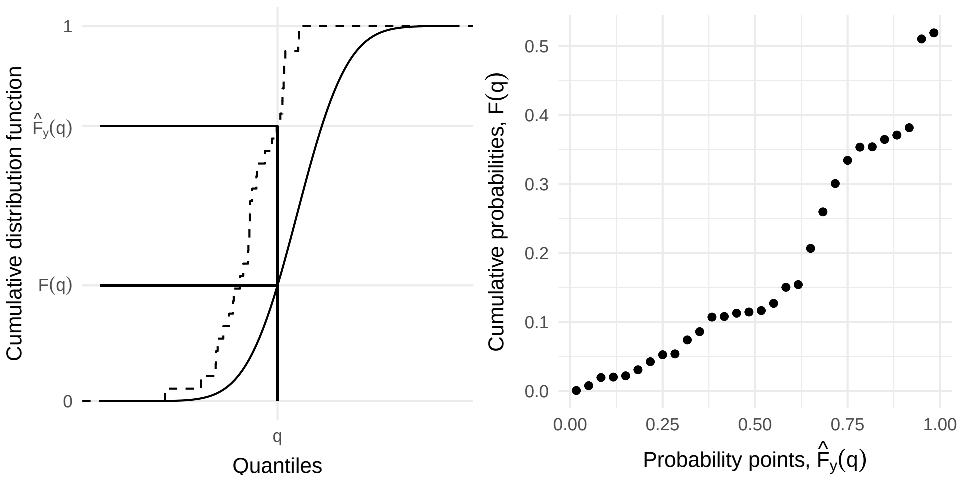

Figure 2: Illustrating what quantities are being plotted for P-P plots.

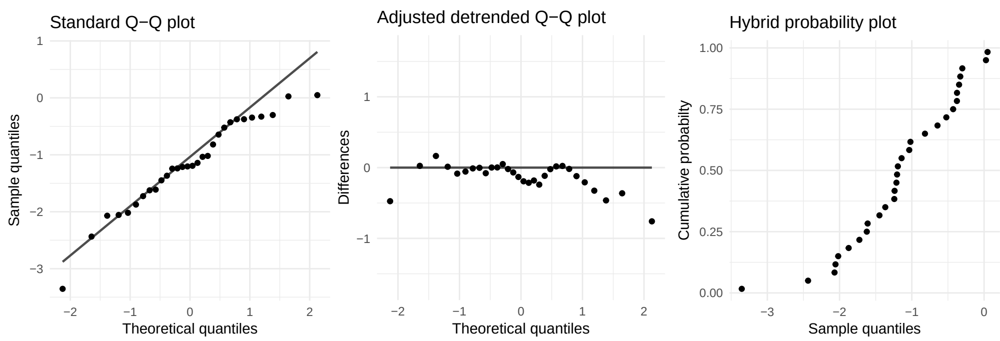

Figure 3: Illustrating different designs of probability plots.

Team, 2012). `qqmath` in **lattice** provides a general framework for Q-Q plots, comparing a sample to any theoretical distribution by specifying the quantile function (Sarkar, 2008). `qqPlot` in the **car** package also allows for the assessment of non-normal distribution and adds pointwise confidence bands based on the standard errors of the order statistics or the parametric bootstrap (Fox and Weisberg, 2011). **ggplot2** provides `geom_qq` and `geom_qq_line`, enabling the creation of traditional Q-Q plots with a reference line, much like those created using `qqmath`. **qqplotr** extends **ggplot2** to provide the most complete implementation of probability plotting.

XXX `qualityTools` Roth (2016)

In the remainder of this paper, we introduce the probability plotting framework provided by **qqplotr**… FILL THIS IN ONCE OTHER SECTIONS ARE WRITTEN…

TODO: FIGURE OUT WHERE TO INTRODUCE TS BANDS (Aldor-Noiman et al., 2013) TODO: FIGURE 3 HAS NO REFERENCES TO IT

Implementing probability plots in the ggplot2 framework

With **qqplotr** we extend some of the original **ggplot2** quantile plot functionatilites by permitting the drawing of Q-Q points, lines, and confidence bands. Our approach provides a **ggplot2** layering mechanism so that for each one of those plot elements we implemented a **ggplot2** "stat" (statistical transformation). In addition, we also implemented a **ggplot2** "geom" (geometrical object) specifically for the confidence bands. That geom permits a simpler way of handling graphical parameters, which will become clearer in the Examples section.

The Q-Q plot functions are divided into three statistical transformations:

- `stat_qq_point`: a modified version of `stat_qq` from **ggplot2** that plots the sample quantiles versus the theoretical quantiles (as in Figure 1). The novelty of this implementation is an option to detrend the plotted points (see Introduction). All other implemented functions in this package also allow the detrend adjustment.
- `stat_qq_line`: draws a reference line based on the sample data quantiles, defaulting to the first and third quartiles.
- `stat_qq_band`: draws confidence bands based on three methods: *Normal* theory, *Bootstrap* resampling, and *Tail-sensitive* simultaneous bands.
  - **Normal:** Specifying `bandType = "norm"` constructs pointwise confidence bands based on the normal approximation to the distribution of the order statistics. For example, an approximate 95% confidence interval for the ith order statistic is $\widehat{X}_{(i)} \pm \Phi^{-1}(.975) \cdot SE(X_{(i)})$, where $\widehat{X}_{(i)}$ denotes the value along the fitted line, $\Phi^{-1}(\cdot)$ denotes the quantile function for the standard normal distribution, and $SE(X_{(i)})$ is the standard error of the ith order statistic.
  - **Bootstrap:** Specifying `bandType = "bs"` constructs pointwise confidence bands using percentile confidence intervals from the parametric bootstrap.
  - **Tail-sensitive:** Specifying `bandType = "ts"` constructs the simulation-based tail-sensitive simultaneous confidence bands proposed by Aldor-Noiman et al. (2013).

Adding confidence bands

In this section we elaborate on the construction of each type of confidence band allowed in `stat_qq_band`.

Normal Specifying `bandType = "norm"` constructs pointwise confidence bands based on

```
+ **Normal**: constructs simultaneous confidence bands based on normal distribution confidence
intervals;
+ **Bootstrap**: creates pointwise confidence bands using parametric boostrap;
+ **Tail-sensitive**: builds tail-sensitive confidence bands, as proposed by
\citet{Aldor-Noiman2013-xw}.
```

Examples

In this section, we demonstrate the capabilities of the **qqplotr** package. We start by loading the package:

```
# also loads ggplot2
library(qqplotr)
```

Table 1: Summary of Iowa's residents height and standard deviation (in inch) by gender and total.

SEX	mean	sd
Male	70.55	2.97
Female	64.51	2.91
Total	66.99	4.18

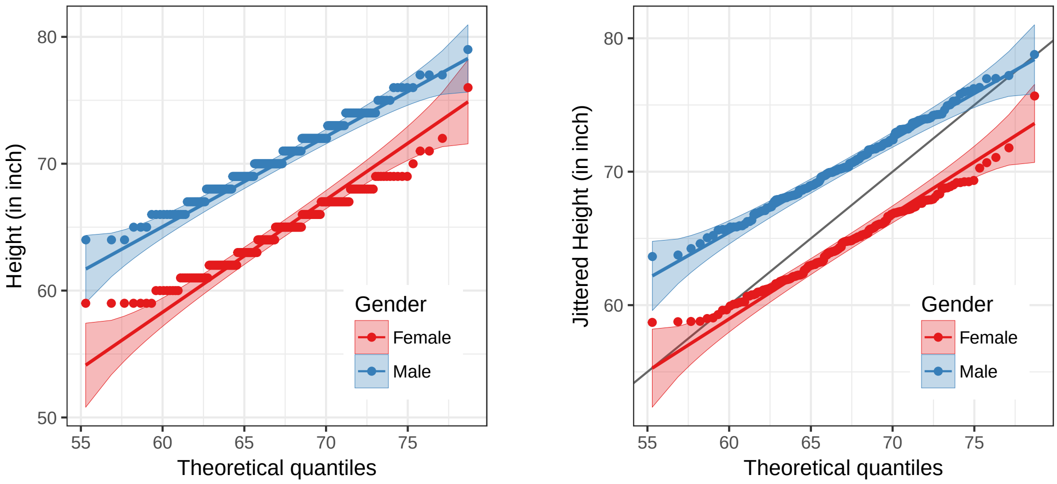

Figure 4: Sample (200 men and 200 women) of raw heights (left) and jittered heights (right). The distribution on the left is dominated by the discreteness of the data. On the right we see that except for some outliers an assumption of normality for people's height is not completely absurd.

BRFSS example

The Center for Disease Control and Prevention runs an annual telephone survey, the Behavioral Risk Factor Surveillance System (BRFSS), to keep track of the US populations' 'health-related risk behaviors, chronic health conditions, and use of preventive services'.

Close to half a million interviews are conducted each year. Here, we are focussing on the 2012 responses for Iowa. 7166 responses were gathered across 359 questions and derived variables. Among these, are people's height and weight, which we are going to assess in more detail.

Figure 4 shows two Q-Q plots side by side. For each of the plots, a sample of 200 men and 200 women is drawn from the overall number of responses. On the left hand side, individuals' heights are plotted in a Q-Q plot comparing raw heights to a normal distribution. We see that the distributions for both men and women (colour) is showing horizontal steps: this indicates that the distributional assessement is heavily dominated by the discreteness in the data, as most survey participants responded to the question of their height to the closest inch. On the right hand side of Figure 4, we use jittering; this means that we add a random number generated from a random uniform distribution on ± 0.5 inch to the reported height. By this mean we diminish the effect that discreteness might have on the distribution. This brings the observed distribution much closer to a normal distribution. Note that separate normal distributions were fitted for each gender, not surprisingly, the resulting distributions have different means (women are on average 6 inch shorter than men in this dataset). Interestingly, the slope of the two genders is similar, indicating that the same scale parameter fits both genders' distributions (the standard deviation of height in the data set is 2.97 inch for men and 2.91 inch for women, see Table 1). The dark line between the two groups is the identity line - indicating the theoretical distribution each of these groups are compared to. This distribution is based on parameters estimated from the whole population (see Table 1 for numbers) . While the mean is about half way between the gender means, we see from the higher slope of the line that in comparison to each group, the standard deviation of the height based on the whole population is larger.

Unlike respondents' heights, their weigths do not seem to be normally distributed. Figure 5 shows again two Q-Q plots. The Q-Q plot on the left uses raw weights and compares to a normal distribution. From the curved points we see that tails of the observed distribution are heavier than expected under a normal distribution. On the right, weights are log-transformed. We see that a normal distribution for each of the genders shows –with the exceptions of a few extreme outliers– a reasonable fit.

Instead of transforming the observed values, we can change the theoretical distribution against which we compare. Figure 6 shows two Q-Q plots where a log-normal distribution is chosen as the

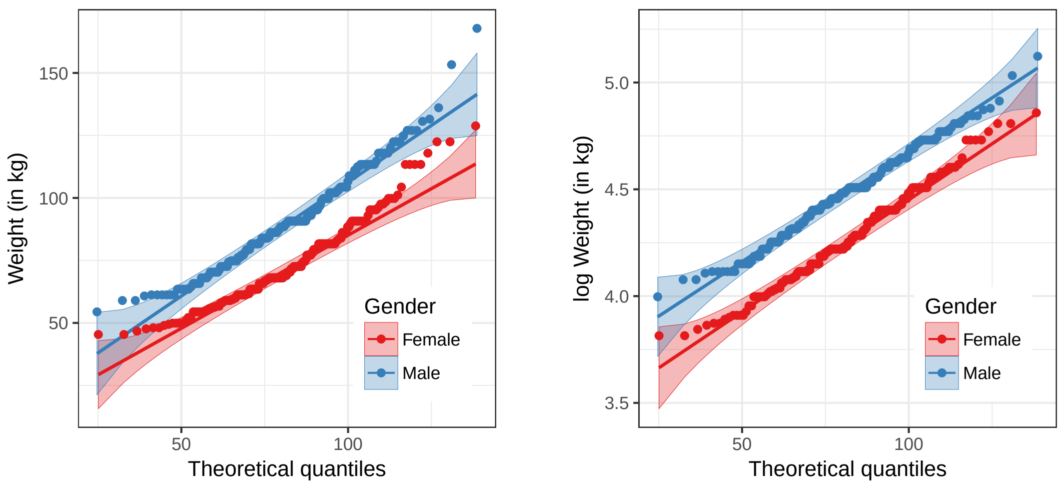

Figure 5: Sample (200 men and 200 women) of weights. Unlike people's height, weight seems to be heavily right skewed with some additional outliers on the extreme left (left plot). On the right, weight was log-transfomed before its distribution is compared to a theoretical normal.

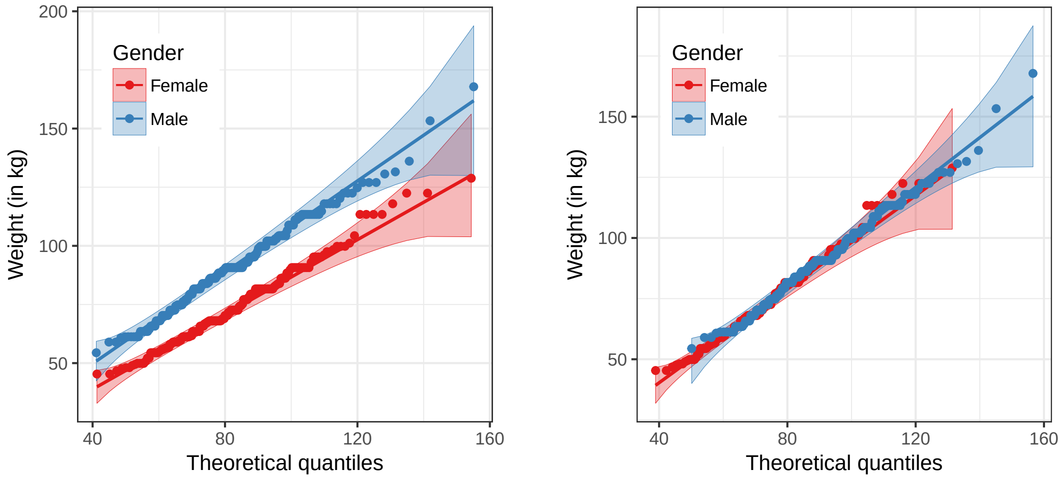

Figure 6: Sample (200 men and 200 women) of weights. On the left, the theoretical distribution is changed to a log normal. On the right, we additionally estimate shift and scale parameters for each of the genders separately before comparing distributions to a log-normal.

theoretical distribution. On the left, we compare against a log-normal distribution with mean 4.389 and standard deviation 0.223 (the log-transformed averages of average weight and standard deviation in Iowa's population). Again, the fits seem reasonable. On the right, parameters for the log-normal distribution are fit separately. The fits are slightly different from XXX

Using a other distributions

Using the capabilities of **qqplotr** with the distributions implemented in the **stats** package is relatively straightfoward, since the implementation allows you to specify the suffix (i.e. distribution and or abbreviation) via the `distribution` argument and the parameter estimates via `dparams` argument. However, there are times when the distributions in **stats** are not sufficient for the demands of the analysis. For example, there is no left-skewed distribution listed. User-coded distributions or distributions from other packages can be used with **qqplotr** as long as the distributions are defined

Table 2: this table is just for us at the moment

SEX	mean_wt	mean_log_wt	sd_wt	sd_log_wt
1	91.50342	4.495766	19.22981	0.2011216
2	74.37085	4.282780	17.90126	0.2254663

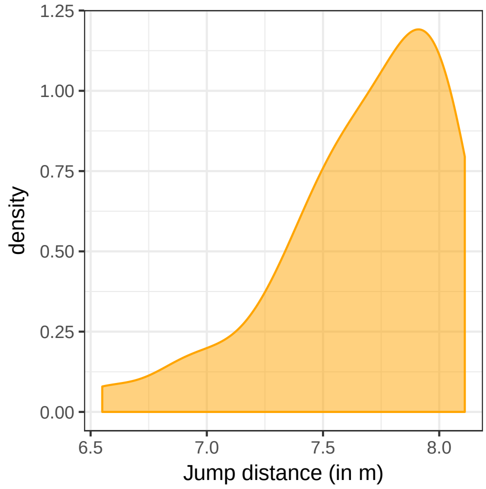

Figure 7: Density plot of the 2012 men's long jump qualifying round. The distances are clearly left skewed.

following the conventions laid out in the **stats** package. Specfically, for some distribution there must be density/mass (d prefix), CDF (p prefix), quantile (q prefix), and simulation (r prefix) functions. In this section we illustrate the use of the smallest extreme value distribution (SEV).

To qualify for the Olympics in the men's long jump in 2012, athletes had to either meet/exceed the 8.1 meter standard or place in the top twelve. During the qualification events, each athlete was able to jump three times, and their best (i.e. longest) jump is treated as the result. Figure 7 shows a density plot of the results, which are cleaerly left skewed.

In order to model the jump distances we must first define a left-skewed distribution. Below, we define the suite of distribution functions necessary to utilize the SEV distribution.

```
# CDF
psev <- function(q, mu = 0, sigma = 1) {
    z <- (q - mu) / sigma
    1 - exp(-exp(z))
}

# PDF
dsev <- function(x,  mu = 0, sigma = 1) {
  z <- (x - mu) / sigma
  (1 / sigma) * exp(z - exp(z))
}

# Quantile function
qsev <- function(p,  mu = 0, sigma = 1) {
  mu + log(-log(1 - p)) * sigma
}

# Simulation function
rsev <- function(n,  mu = 0, sigma = 1) {
  qsev(runif(n), mu, sigma)
}
```

With the *sev distribution functions in hand, we can create a Q-Q plot to assess the appropriateness of the SEV model (Figure 8). The Q-Q plot show that the distances do not substantially deviate from the SEV model, so we have found an adequate representation of the distances.

```
ggplot(longjump, aes(sample = distance)) +
  stat_qq_band(distribution = "sev", dparams=list(mu=0, sigma=1), alpha = 0.3) +
  stat_qq_line(distribution = "sev", dparams=list(mu=0, sigma=1)) +
  stat_qq_point(distribution = "sev", dparams=list(mu=0, sigma=1)) +
  xlab("Theoretical quantiles") +
  ylab("Jump distance (in m)") +
  theme_bw()
```

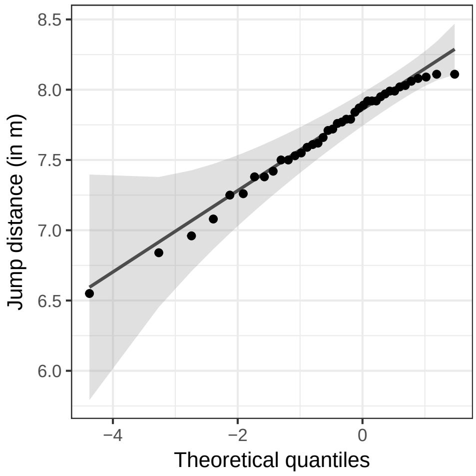

Figure 8: Q-Q plot comparing the long jump distances to the standard SEV distribution. The SEV distribution appear to adequately model the distances.

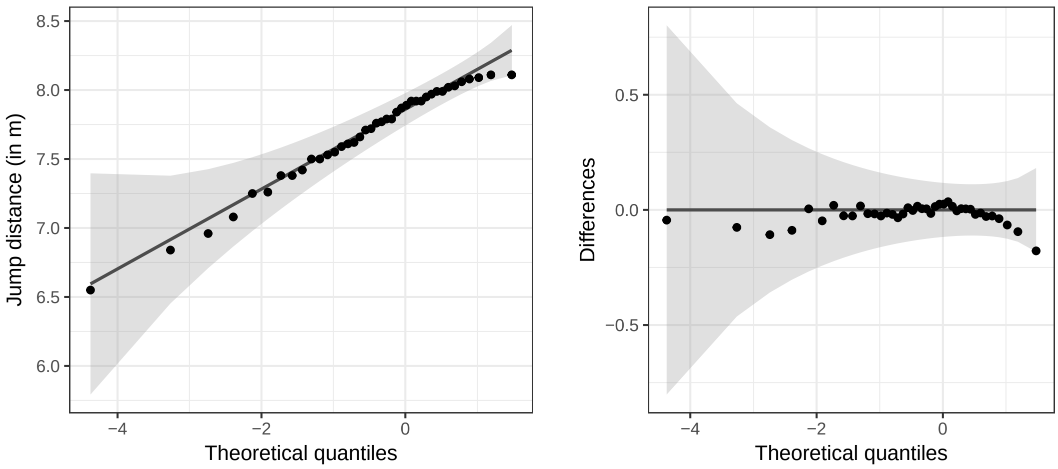

Figure 9: Q-Q plots assessing the appropriateness of the SEV distribution for the long jump data. On the left, a standard Q-Q plot is shown. On the right, we detrend the Q-Q plot by plotting the differences between the empirical quantiles and reference line on the y-axis.

Detrending Q-Q plots

qqplotr also allows for Q-Q plots to be *detrended*. In a detrended Q-Q plot, the y-axis shows the difference between the empirical quantile and the reference line (i.e. the theoretical distribution). This layout directly plots what we want viewers to assess—the difference between the distributions being compared—which Loy et al. (2016) found to be more powerful than other designs, so long at the y-axis limits are set so that the aspect ratio is kept the same as in the traditional Q-Q plot.

For example, Figure 9 compares the standard Q-Q plot shown in Figure 8 with a detrended version by adding the argument `detrend = TRUE` to the `stat_qq_band`, `stat_qq_line`, and `stat_qq_point` calls.

```
ggplot(longjump, aes(sample = distance)) +
  stat_qq_band(distribution = "sev", alpha = 0.3,
               detrend = TRUE, dparams=c(mu=0, sigma=1)) +
  stat_qq_line(distribution = "sev",
               detrend = TRUE, dparams=c(mu=0, sigma=1)) +
  stat_qq_point(distribution = "sev",
                detrend = TRUE, dparams=c(mu=0, sigma=1)) +
  xlab("Theoretical quantiles") +
  ylab("Differences") +
  theme_bw() +
  theme(aspect.ratio = 1)
```

In order to create a so-called *adjusted* detrended Q-Q plot (Loy et al., 2016) the aspect ratio must also be set to 1. If the aspect ratio is not adjusted in this way, an *ordinary* detrended Q-Q plot is created, which is known to lhave lower power than the standard Q-Q plot in some situations (Loy et al., 2016).

Summary

XXX P-P plots here Write this section once the rest of the paper is done.

Bibliography

S. Aldor-Noiman, L. D. Brown, A. Buja, W. Rolke, and R. A. Stine. The power to see: A new graphical test of normality. *The American Statistician*, 67(4):249–260, 1 Nov. 2013. [p3]

J. Fox and S. Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011. URL `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`. [p3]

A. Loy, L. Follett, and H. Hofmann. Variations of Q–Q plots: The power of our eyes! *The American Statistician*, 70(2):202–214, 2 Apr. 2016. [p1, 7, 8]

W. Q. Meeker and L. A. Escobar. *Statistical methods for reliability data*. John Wiley & Sons, 1998. [p1]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0. [p1]

T. Roth. *qualityTools: Statistics in Quality Science.*, 2016. URL `http://www.r-qualitytools.org`. R package version 1.55 http://www.r-qualitytools.org. [p3]

D. Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008. URL `http://lmdvr.r-forge.r-project.org`. ISBN 978-0-387-75968-5. [p3]

M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55 (1):1–17, Mar. 1968. [p1]

Acknowledgements

This work was partially funded by Google Summer of Code 2017.

Alexandre Almeida
University of Campinas
Institute of Computing
Campinas, Brazil 13083-852
`almeida.xan@gmail.com`

Adam Loy
Carleton College
Department of Mathematics and Statistics
Northfield, MN 55057
`aloy@carleton.edu`

Heike Hofmann
Iowa State University
Department of Statistics
Ames, IA 50011-1210
`hofmann@iastate.edu`