MATH 6358

Probability Models and Statistical Computing Final Project

Fall 2019

Assessing Q-Q Plot Variations and Testing Normality

Stephanie Dinh

Jordan Robles

Dayana Sosa

Kishore Tumarada

University of Houston

Abstract

When assessing a set of data, we would like to see if it fits a theoretical distribution. Having graphical tools, such as the Quantile-Quantile (Q-Q) plot, allows us to test our data against various distributions with visual ease. In our paper, we will focus on assessing normality through formal distribution tests and lineup tests to see which performs better. We will be utilizing lineup tests to compare the Q-Q plot against other variations. With our results, we will apply our findings of the best test of normality to various data sets that claim to be normal. This paper will be based on “Variations of Q-Q Plots: The Power of Our Eyes” by Adam Loy, Lendie Follett, and Heike Hofmann (THE AMERICAN STATISTICIAN 2016 Vol 70).

Contents

1. Introduction: The Normal Distribution

In statistics, we wish to model data collected with a known distribution using certain parameters. One of the most well-known and used distributions in statistics is the normal, or Gaussian, distribution. A continuous random variable X with mean μ and standard deviation σ is said to have normal distribution if its probability density function is defined as

$$f(x:\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)},-\infty<x<\infty$$

.

Knowing the type of distribution a data set has allows us to understand and further investigate our data. For example, if we know that our data is normally distributed, we can parameterize our data with its mean and standard deviation. We can also perform accurate statistical computations and analysis given that we know the proper distribution of our data.

In this project, we will explore the use of lineup protocol to improve graphical distributional assessment of normal distribution like QQ plots and compare its power to classical distributional tests such as Shapiro-Wilk test.

2. Classic Tests To Assess Normality

There are various tests we can use to check whether a particular set of data follows a normal distribution. Some of the classical tests for assessing normality include the Kolmogorov-Smirnov (K-S test), Lilliefors corrected K-S test, Anderson-Darling test, and Shapiro-Wilk test. These tests are interpreted numerically, and compare the empirical distribution function from our data to the theoretical distribution function. The K-S test, Lilliefors corrected K-S test, and Anderson-Darling test look at the absolute difference between the two distribution functions for each sample point. A limitation of the K-S test is its sensitivity to extreme values, to which the Lilliefors corrected K-S test can accommodate. From classical tests of normality, the Shapiro-Wilk test, which focuses on the linearity of a normal Q–Q plot, is preferred (Loy 2016).

3. Q-Q Plots

We can use graphical methods, such as Q-Q plots to assess normality, to visually assess normality. Q-Q plots are "a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution" (Ford 2015). We will primarily focus on utilizing Q-Q plots to test for normality, but we shall also see the extension of the Q-Q plot for other distributions. Q-Q plots are generated by plotting the quantiles of a theoretical distribution against the quantiles of our data. Furthermore, we can add a line of identity, defined by a "theoretical" quantile-quantile plot which passes through the first and third quartiles, to our plot for aiding in our visual assessment. If the Q-Q plot appears to fall on or closely to the identity line, our data may likely have a normal distribution. Otherwise, if the points in the Q-Q plot appear to deviate from the identity line, then it is less likely that our data may have a normal distribution. While the Q-Q plot assessment is criticized for being too subjective, we shall use lineup tests for different variations of Q-Q plots to inject rigor into our assessment of normality.

4. Q-Q Plot Variations

Not including the control, we have 4 different combinations of QQ plot designs as shown in figure 1. The Davison–Hinkley (DH) 95% pointwise confidence band and the 95% Tail-Sensitive (TS) simultaneous confidence bands each have a standard and de-trended version, giving us 5 designs in total. As the names suggest, these two confidence bands are the region in which we are 95% confident that the data contained within will be normal. Consequently, if any of the points fall outside of the band, we reject the notion of normality.

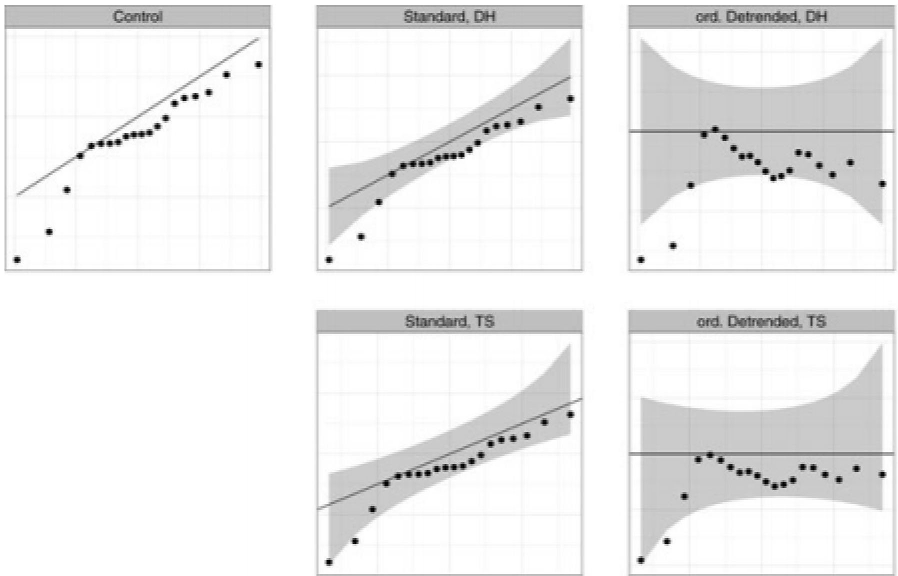

Figure 1. Q-Q plot variations: control, standard and ordinary Detrended Q-Q plots, with Davison-Hinkley (DH) or Tail Sensitive (TS) 95% confidence bands.

Compared to DH, the TS confidence bands are slightly wider in the middle and at the tail ends of the plot. With a larger area, there should be fewer rejections of normality and as a result, the TS confidence band may decrease the amount of false-positives in the lineup test indicating a deviation from normality, increasing the power.

The standard versions are the plots as explained above with the line of identity as well as a confidence band superimposed on top. The de-trended versions of the QQ plots change the y-axis so that the plot visualizes the difference between the observed and theoretical points rather than the overall values of the data. Here the identity line is now horizontal, and viewers can more easily see the raw deviations from the normal distribution.

The goal is to analyze the differences in the designs and determine the most efficient plot in aiding viewers to identify the data plot in various lineups. Although the plots may look different, it should be noted that each of the designs are to be run on the same data to maintain an objective evaluation.

5. Lineup Test and Significance

Lineup tests proposed by Buja et al. (2009) aim to use an inferential framework to evaluate and measure the significance of visual statistical methods. In place of the test statistic are the plots, while the viewers' perception of the plots replaces the statistical test, and finally the statistical significance (p-value) is quantified by the viewers' comparison of the observed plot to the collection of simulated null plots.

5.1. Explanation

The term "lineup" comes from the association to a police lineup process, in which a witness picks out the accused among a group of visually similar people. In the statistical sense, the "accused" is our plot of data to be evaluated, randomly placed among a set of null plots which depict data created to be consistent with the null hypothesis. If the "witnesses" (or in our case the viewers), asked to pick the contrasting plot, are able to choose the observed plot from the lineup, then we have reason to believe that the observed plot is different from null and can use this as evidence against the null hypothesis.

We can also quantify the probability of rejecting the null hypothesis, or the power, of the lineup by using the equation $\widehat{Power} = Power_N = 1 - F_N(y_\alpha)$, where Y is the number of identifications of the data plot in N independent evaluations, and $Y \sim F_N$. In other words, this is the probability that more than y_α (critical value at α, $P(Y > y_\alpha) > \alpha$) of N viewers pick out the true data plot. Y is made up of the sum of the N viewers' decisions (binary: $Y_i \sim B_{1,p_i}$). The probability p_i that viewer i chooses the data plot depends on the strength of the signal (degrees of freedom and the sample size) in the data plot and individual i's visual ability to differentiate the true data plot. The visual ability needs be measured by having individual i evaluate multiple lineup plots.

In this way, lineup plots allow us to attach a quantifiable weight to the otherwise subjective perception of the viewer in determining the normality of a data set.

5.2. Construction of Lineup Plots

For each of our lineup plots, we created the true data plot from a t-distribution with varying sample size and degrees of freedom. This was done using the function rt(size,df) in R. In order to create the lineup plots we had to use the Nullabor package. Using the lineup() function from this package allowed us to get 19 separate normally distributed samples. Then we were able to plot our true data set with the 19 other plots into one lineup using ggplot. Within ggplot we were able to have lineups with confidence bands by adding the code: stat_qq_band(bandType = ' '). We were also able to create the detrended lineup plots by adding to the code: detrend=TRUE.

6. Simulation Setup

6.1. Assumptions

All participants are assumed to share the same visual ability.

6.2. Design

Below are the two hypotheses used in the paper.

	Hypothesis 1	**Hypothesis 2**
Null Hypothesis	H_0 : F=N(0,1)	H_0 : F=N(0, S^2)
Alternative Hypothesis	H_a : F≠N(0,1)	H_a : F≠N(0, S^2)

Table 1: Hypothesis for simulation

If multiple people surveyed can choose the true plot that is not normally distributed, then this leads to the rejection of the null hypothesis. We can claim that the data is not normally distributed. Otherwise, we fail to reject the null hypothesis.

We were able to survey 28 people in total across 4 different surveys, however, the amount of people was not evenly distributed. Each member of the group created two tail sensitive (TS), detrended DH, and detrended TS.

We made use of simulation study with sample size : {20,75} and t distribution with degrees of freedom : {2,10}. Each of the sample is inserted among 19 null plots with same sample size from normal distribution. Each of the 4 samples has 5 lineup plots for each kind of QQ plot, a total of 4X5 = 20 plots. Further plots are also created against sample distribution N(0,S^2) as explained above. So a total of 40 lineup plots are simulated for this experiment.

From there we compiled the results, determining the proportion of people able to correctly identify the true plot and created the line plots as seen below.

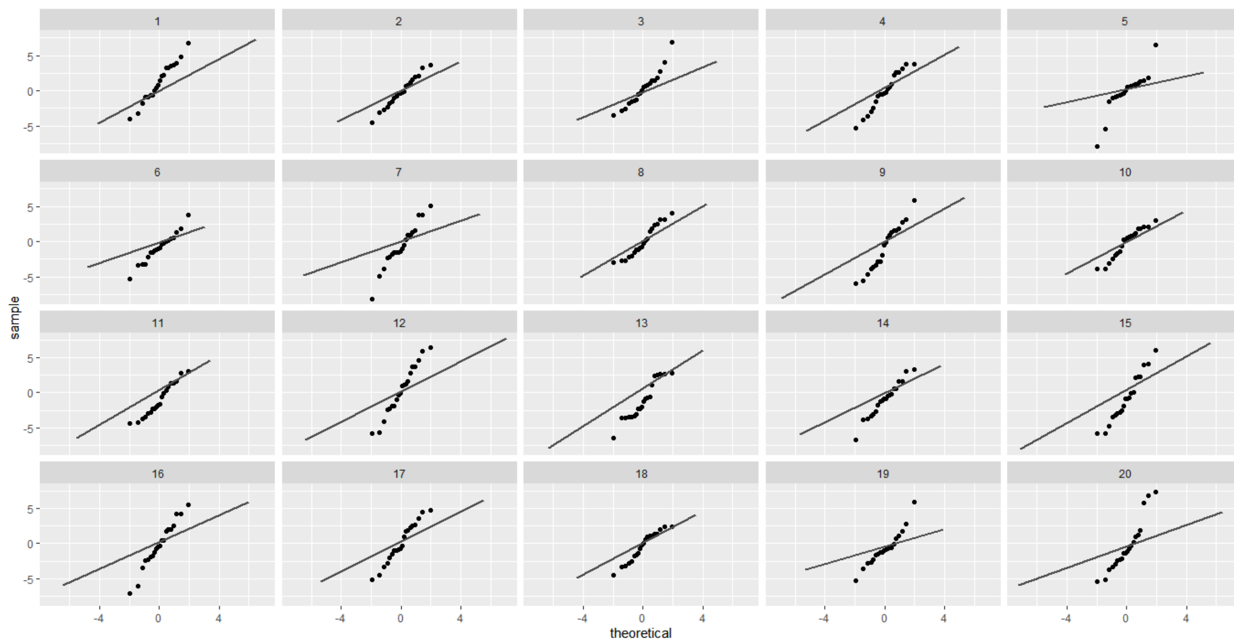

Figure 2. Standard lineup plot sample

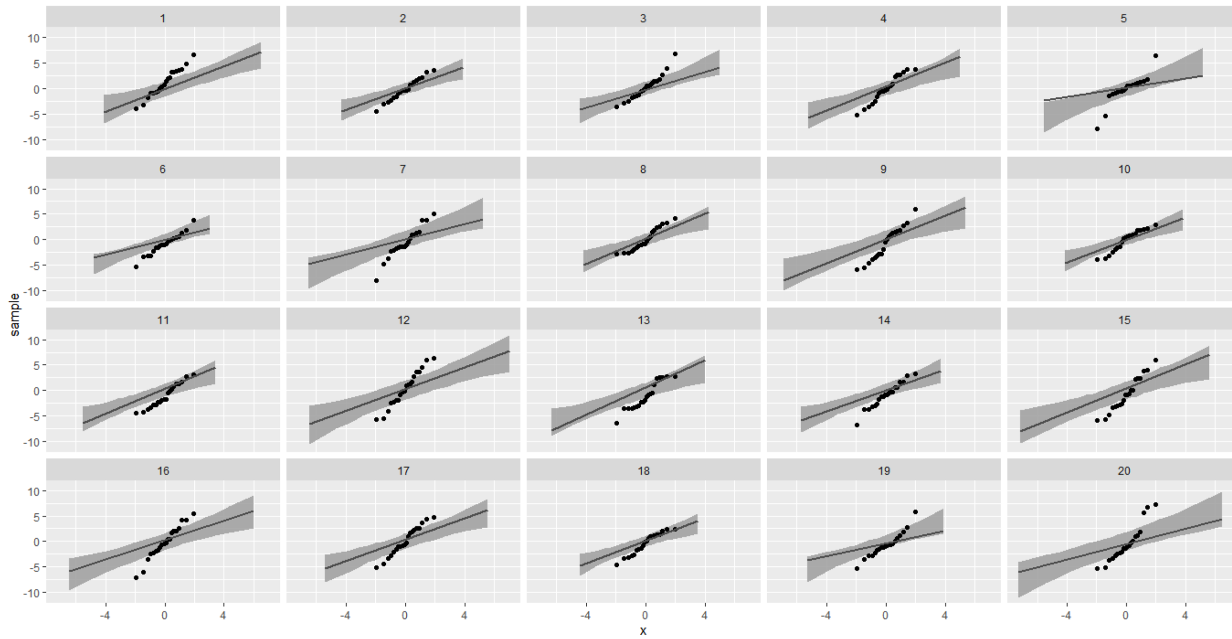

Figure 3. DH confidence lineup sample

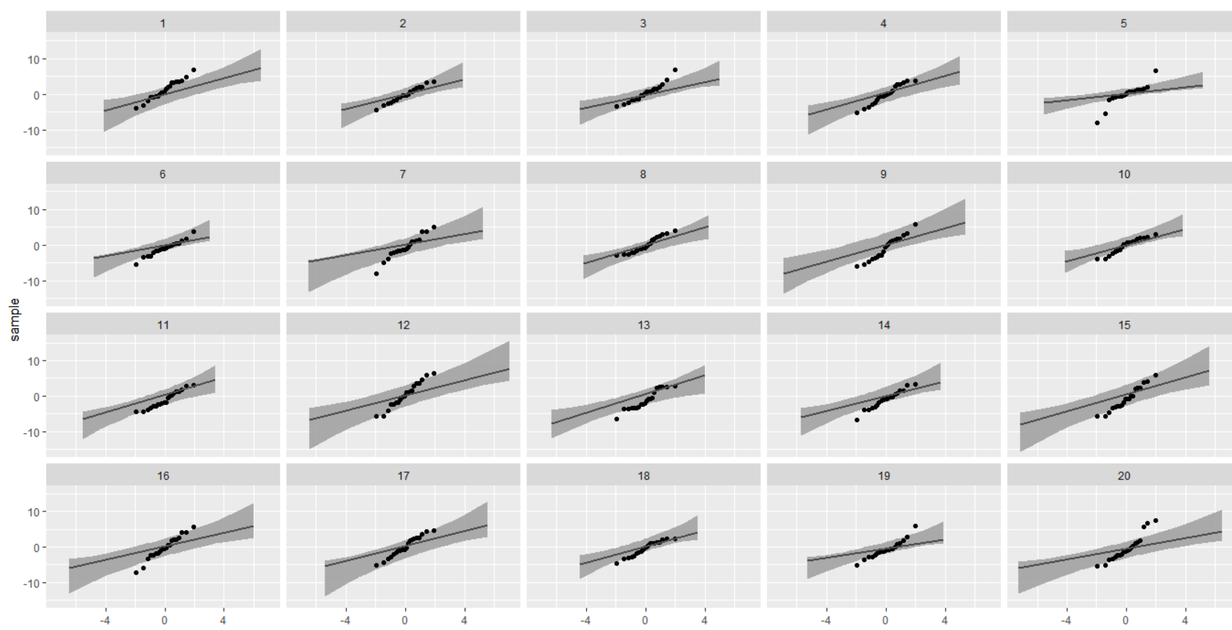

Figure 4. TS confidence lineup sample

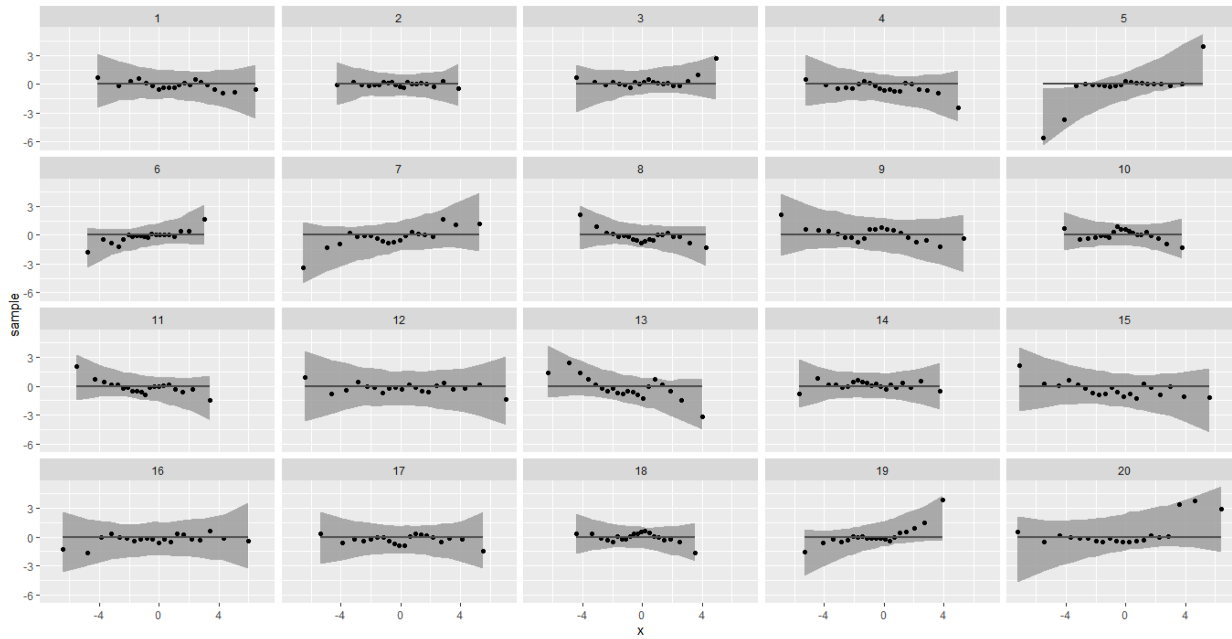

Figure 5 . Original detrended DH lineup plot sample.

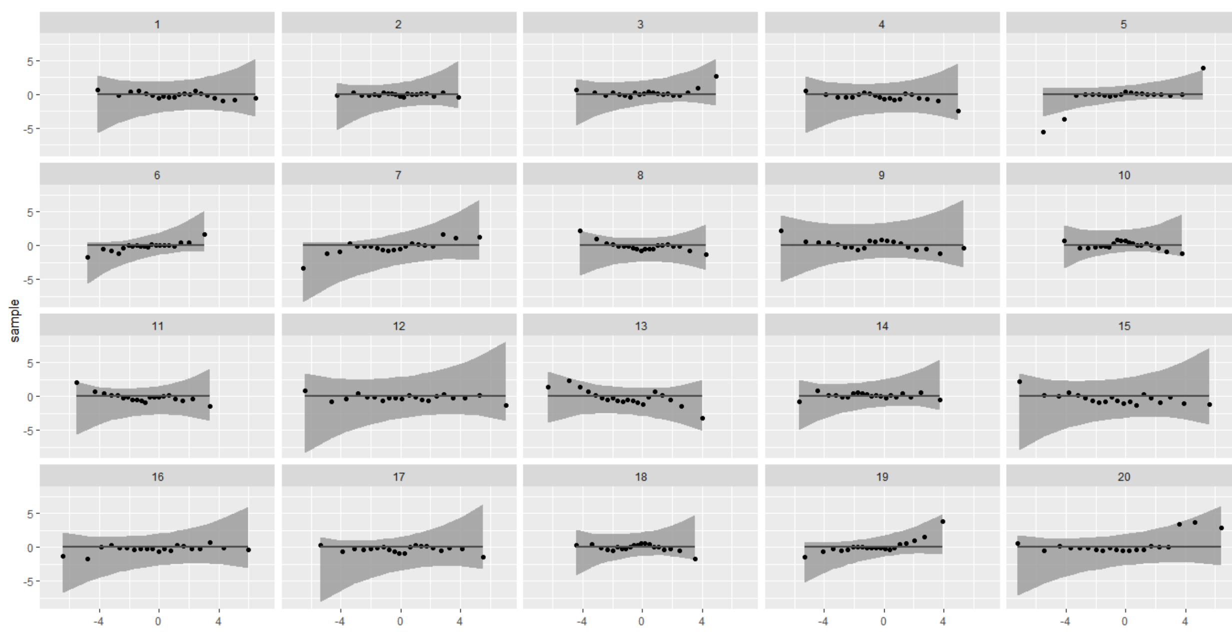

Figure 6 . Original detrended TS lineup plot sample.

6.3. Percentage of correct identifications under the different variations of Q–Q plots

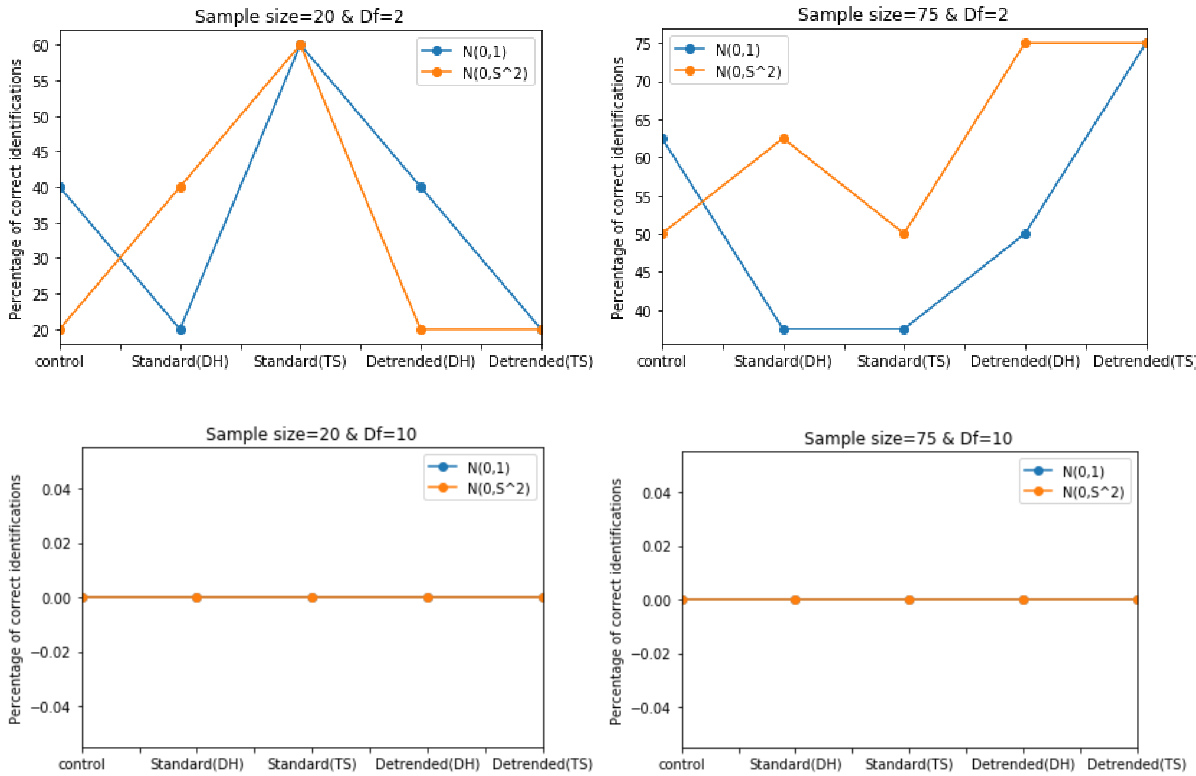

Figure 7. Faceted parallel coordinate plots of percentage of participants identifying t sample from lineups for different variations of QQ plots.

7. Analysis of Survey outcomes

7.1. Within plots performance

a. **Performance of unit variance Vs sample variance:** For lower degrees of freedom(here df =2) and smaller sample size, unit variance has better prediction capacity than sample variance plots. But for higher sample size, the sample variance plots have higher power of prediction.

b. **Performance of different types of plots:** For lower sample size, standard Tail sensitive plots have higher power than detrended plots, but detrended plots have better performance for higher sample sizes. This is significant as sample size increases, the distribution approaches normal distribution and becomes difficult to differentiate from standard QQ plots. Even in such critical conditions, detrended plots have higher performance.

c. **Performance of DH band vs TS band:** TS bands are found to be more powerful than DH bands in detecting heavy-tailed distribution.(Aldor Noiman et al 2013 - 4). Tail sensitive bands have higher performance for both lower and higher sample sizes as shown in figure 6.

d. **Performance for increase in sample size:** As sample size increases, there is an increase in signal strength, hence there is a higher chance for observers to pick out the sample.

e. **Performance for increase in degrees of freedom:** As degrees of freedom increases, the prediction rate has come down to nil for both smaller and larger sample size. This might be because the corresponding t distribution is closer to normal distribution.

    i. Comparison of sample (75,2) with sample(75,10) for detrended TS plot, where there is a wide margin between correct plot prediction and incorrect prediction

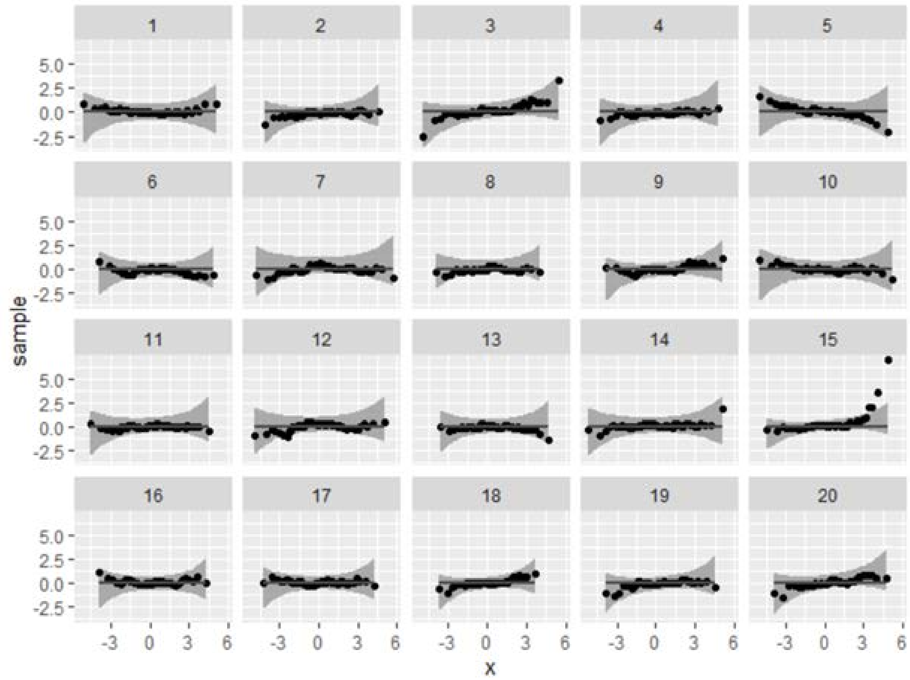

Figure 8 : De-trended TS band plot with null hypothesis for N(0,1) sample for Sample size 75 and df 2 with 75% prediction rate

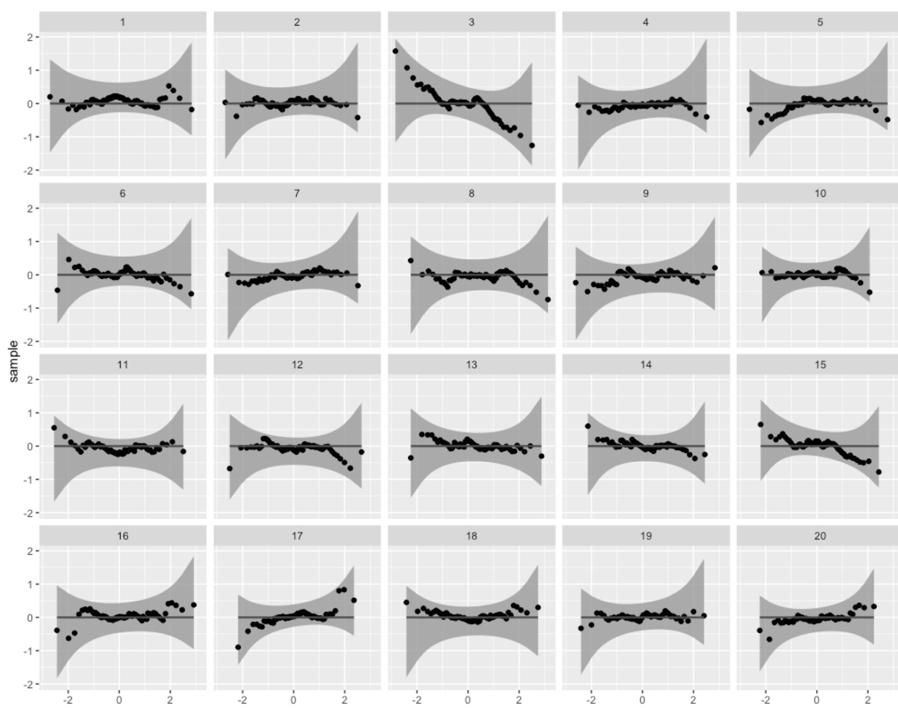

Figure 9 : Detrended TS band plot with null hypothesis for N(0,1) sample for Sample size 75 and df 10 with 0% prediction rate

For figure 8 , correct plot position of 15 has been predicted by 75% observers, as it is the only plot with points outside the confidence band. The reason is also mentioned as ‘outlier’. However for figure 9, the correct plot at position 18 has not been predicted by anyone but plot 3 has been selected by 83% observers with reason 'Curve’. It might be because observations on both tails are away from identity line unlike any other plot.

f. **Reasons for predictions :** For correct predictions, Observers gave outliers and curve as the reasons for choosing correct non-normal plots. Observers, who gave wrong predictions, are more biased towards plots with left and right deviations of QQ plots. Hence reasons other than outliers decrease the chance of predicting correct plots.

7.2. Plots vs classical test

a. Shapiro wilk test has been applied to all 4 samples and following are the results :

			shapiro wilk	shapiro wilk
S No	**Sample size**	**df**	**W**	**p value**
1.	20	2	0.90055	0.04224
2	20	10	0.94971	0.3628
3	75	2	0.9436	0.002257
4	75	10	0.99079	0.8677

Table 2: Results of Shapiro – Wilk test on above samples

b. For a confidence level of 95%, alpha value is 0.05. Here Hypothesis testing is done for:
   i. H0 : Sample is from normal distribution
   ii. Ha : Sample is from non-normal distribution.
c. If p value is less than alpha value, we reject the hypothesis of normality. As per above table of p-values, null hypothesis can be rejected for the combinations (20,2) and (75,2) and cannot be rejected for (20,10) and (75,10).

d. Here all the samples are non-normal distributions (t-distribution), so the test should have rejected null hypothesis for all 4 samples. But the test could not reject for 2 samples.
e. Unlike the reference paper, the QQ plot lineup survey also gave same results as the Shapiro Wilk test. Ideally, with more observers for the survey, we would have been able to reject null hypothesis for remaining 2 samples also.

8. Limitations of QQ plot lineup analysis

When compared to classical tests, lineup tests take more time to test normality, as a lineup of 20 plots have to be made even for a single sample. However availability of R packages like Nullabor decreases the time taken for lineups substantially.

Moreover, the efficiency of this method can be increased by increasing the number of observers in the survey.

9. Application of Lineup plots to Exponential Distribution:

Q-Q plots are not limited to only testing for normality. These plots can be used to visually assess data against any theoretical distribution. We were able to construct Q-Q plots for testing against the exponential distribution. The variations of these Q-Q plots included a control exponential Q-Q plot, a standard exponential Q-Q plot with confidence bands, and a de-trended exponential Q-Q plot with confidence bands. Similar to our lineup Q-Q Plots for assessing normality, we created lineup plots for the exponential distribution as shown below non-exponential plot position #4. Out of these three types of plots, the best performing was the de-trended plot with confidence bands.

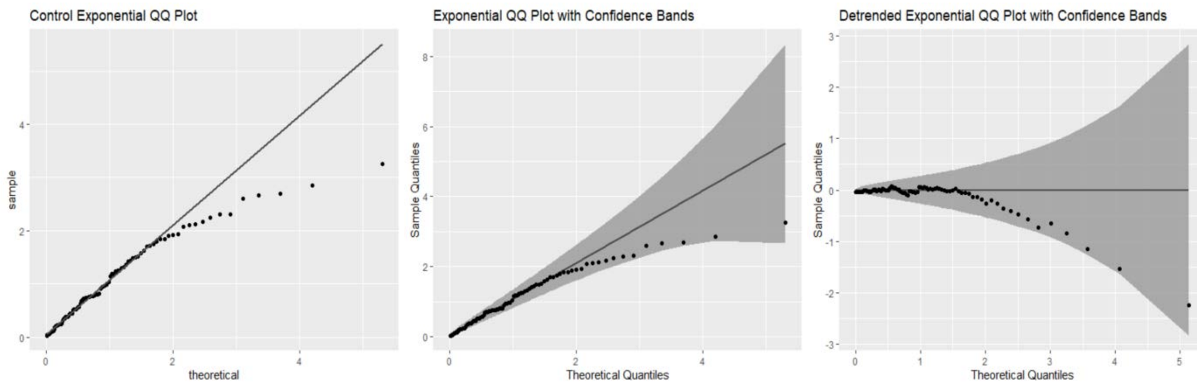

Figure 10: Application Q-Q Plot Variations to Exponential Distribution

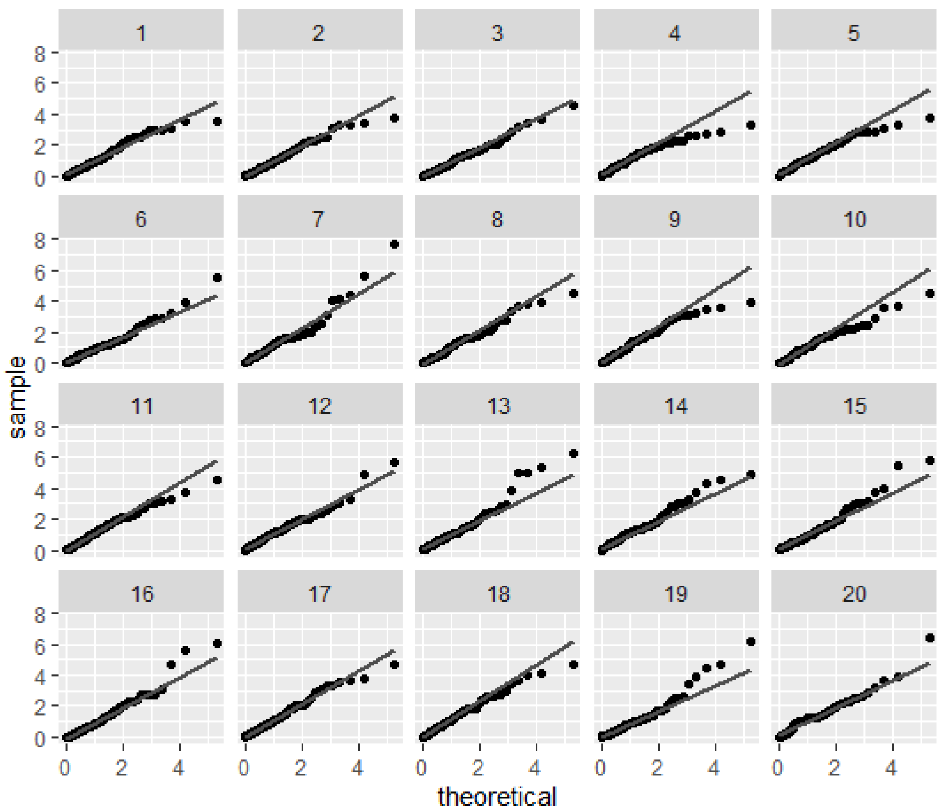

Figure 11: Control Lineup Plots for Exponential Distribution

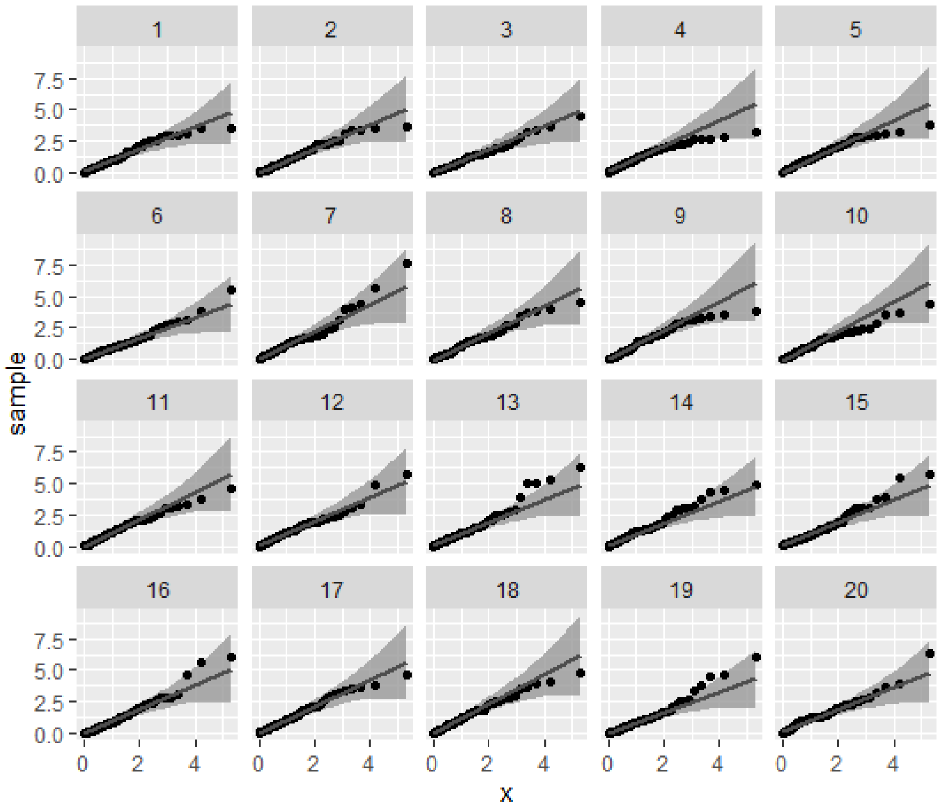

Figure 12: Standard Lineup Plots with Confidence Bands for Exponential Distribution

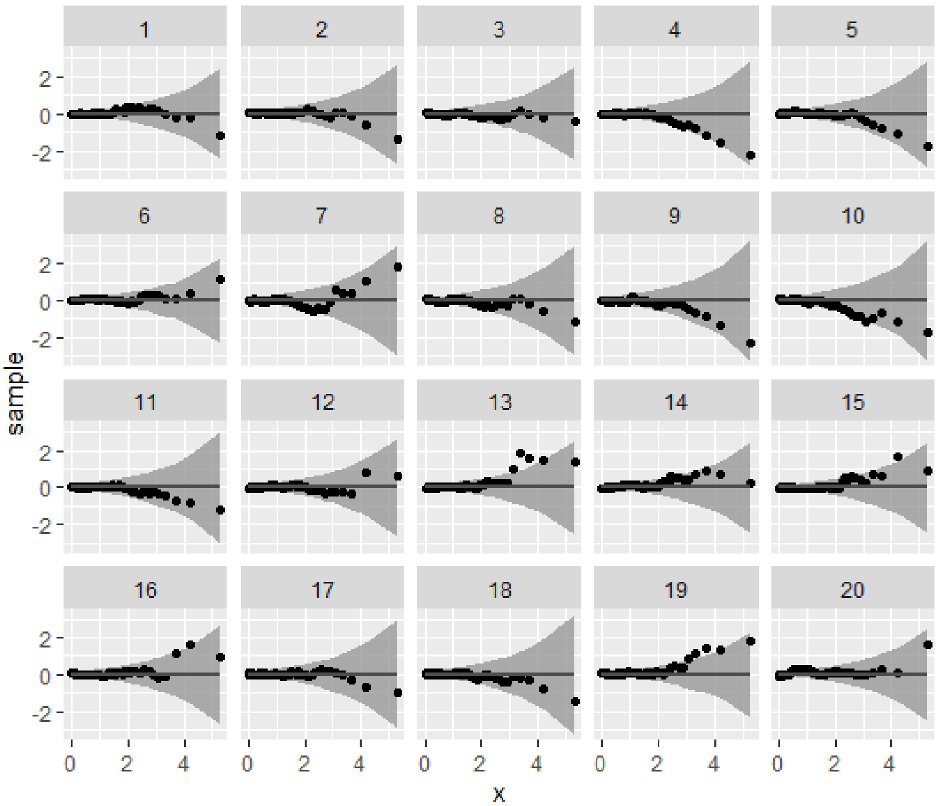

Figure 13: Detrended Lineup Plots with Confidence Bands for Exponential Distribution

10. Conclusion

In conclusion, all enhanced versions of Q-Q plots are more effective than the standard Q-Q plot in detecting non-normality in a distribution. Confidence bands, especially tail-sensitive(TS) bands are more powerful. Among the five kinds of Q-Q plots, detrended plot is found to be more effective. Although, we could not get better results than classical normality test like the Shapiro-Wilk test, we have been able to extend lineup Q-Q plots of different variations to the exponential distribution.

APPENDIX

A.1. Survey results

S No	Alternative Hypothesis	Sample size	df	Plot type	true position	majority vote	%	Reason	second rank	%	Reason	Correct Prediction
1	N(0,1)	20	2	standard	5	5	40	outlier	13-18-20	20	right and curve	TRUE
2	N(0,1)	20	2	DH confidence	5	none	20		5-13-15-16-20	20	outlier	FALSE
3	N(0,1)	20	2	TS confidence	5	5	60		20	40	curve	TRUE
4	N(0,1)	20	2	DH orig detrended	5	5	40	outlier	12	40	curve	TRUE
5	N(0,1)	20	2	TS orig detrended	5	none	20		5-7-9-13-20	20	outlier and curve	FALSE
6	N(0,S^2)	20	2	standard	3	none	20		3-5-6-15-16	20	curve	FALSE
7	N(0,S^2)	20	2	DH confidence	3	3	40	outlier and curve	1-15--16	20	right and outlier	TRUE
8	N(0,S^2)	20	2	TS confidence	3	3	60	outlier	13-1	20	right and outlier	TRUE
9	N(0,S^2)	20	2	DH orig detrended	3	5	60	curve and outlier	8--3	20	outlier	FALSE
10	N(0,S^2)	20	2	TS orig detrended	3	5	40	outlier and other	1--3--11	20	outlier	FALSE
11	N(0,1)	20	10	standard	20	1	60	curve	8,4	40	left	FALSE
12	N(0,1)	20	10	DH confidence	20	13	60	outlier	8,15	40	curve	FALSE
13	N(0,1)	20	10	TS confidence	20	8	60	curve	1,10	40	right and outlier	FALSE

14	N(0,1)	20	10	DH orig detrended	20	5	40	outlier	15,16,18	60	right and curve	FALSE
15	N(0,1)	20	10	TS orig detrended	20	5,16	80	curve	15	20	outlier	FALSE
16	N(0,S^2)	20	10	standard	14	8	40	left and outlier	1,17, 20	60	other	FALSE
17	N(0,S^2)	20	10	DH confidence	14	1	40	outlier	2,3,10	60	left and curve	FALSE
18	N(0,S^2)	20	10	TS confidence	14	1	60	outlier	2,16	40	left	FALSE
19	N(0,S^2)	20	10	DH orig detrended	14	15,20	80	outlier	17	20	curve	FALSE
20	N(0,S^2)	20	10	TS orig detrended	14	20	40	curve and outlier	8,15, 17	60	right	FALSE
21	N(0,1)	75	2	standard	15	15	62.5	curve	7, 12, 14	12.5	outlier	TRUE
22	N(0,1)	75	2	DH confidence	15	15	37.5	other	1	25	curve and outlier	TRUE
23	N(0,1)	75	2	TS confidence	15	15	37.5	outlier	7, 14	25	other	TRUE
24	N(0,1)	75	2	DH orig detrended	15	15	50	curve and outlier	5	25	curve and outlier	TRUE
25	N(0,1)	75	2	TS orig detrended	15	15	75	outlier	5	25	curve	TRUE
26	N(0,S^2)	75	2	standard	5	5	50	outlier	3, 17	25	other	TRUE
27	N(0,S^2)	75	2	DH confidence	5	5	62.5	outlier	17	25	other	TRUE
28	N(0,S^2)	75	2	TS confidence	5	5	50	curve and outlier	17	25	other	TRUE
29	N(0,S^2)	75	2	DH orig detrended	5	5	75	curve	15	25	outlier	TRUE
30	N(0,S^2)	75	2	TS orig detrended	5	5	75	outlier	3, 15	12.5	curve	TRUE
31	N(0,1)	75	10	control	1	10	8	left	18	1	outlier	FALS

							3.30	side		6.70		E
32	N(0,1)	75	10	standard ts	11	3,14,18	33.30	curve			left side	FALSE
33	N(0,1)	75	10	standard dh	6	15	66.70	left side	12,13	16.70	right,curve,outlier	FALSE
34	N(0,1)	75	10	detrended ts	18	3	83.30	curve	19	16.70	outlier,curve	FALSE
35	N(0,1)	75	10	detrended dh	4	1	50.00	outlier/right side	17	33.30		FALSE
36	N(0,S^2)	75	10	control	14	9	83.30	outlier/right/left	20	16.70		FALSE
37	N(0,S^2)	75	10	standard ts	1	8,13,16	33.30	left side			outlier	FALSE
38	N(0,S^2)	75	10	standard dh	12	5	66.70	left/right side	4,14	16.70	curve, outlier	FALSE
39	N(0,S^2)	75	10	detrended ts	11	5	50.00	outlier	2	33.30	right side	FALSE
40	N(0,S^2)	75	10	detrended dh	18	13,16	33.30	curve	5,20	16.70	outlier	FALSE

A.2. R code

```
library(nullabor)
library(ggplot2)
library(dplyr)
library(qqplotr)

#random t-dist sample, (change the size and df)
ts1 <- rt(75,10)

t.sample <-data.frame(ts1)

#this is the sample plot

ggplot(t.sample, aes(sample=ts1))+stat_qq()+stat_qq_line()
inf <- lineup(null_dist("ts1", dist = "normal",params = list(mean=0,sd=sd(ts1))),t.sample)
#position of true plot
attr(inf, "pos")
#standard lineup plot
ggplot(inf,aes(sample=ts1))+stat_qq()+stat_qq_line()+facet_wrap(~ .sample)
#ts confidence lineup
ggplot(inf,aes(sample=ts1))+ stat_qq_band(bandType = 'ts')
+stat_qq()+stat_qq_line()+facet_wrap(~ .sample)
 #DH confidence lineup
ggplot(inf,aes(sample=ts1))+ stat_qq_band(bandType = 'boot')
+stat_qq()+stat_qq_line()+facet_wrap(~ .sample)
 #Detrended/ TS lineup
ggplot(inf,aes(sample=ts1))+ stat_qq_band(bandType = 'ts',detrend = TRUE)
+stat_qq_point(detrend = TRUE)+stat_qq_line(detrend = TRUE)+facet_wrap(~ .sample)
 #Detrended/ DH lineup
ggplot(inf,aes(sample=ts1))+ stat_qq_band(bandType = 'boot',detrend = TRUE)
+stat_qq_point(detrend = TRUE)+stat_qq_line(detrend = TRUE)+facet_wrap(~ .sample)

###Shapiro Wilk test for all 4 samples###
size = 20
df =2
#random t-dist sample, (change the size and df)
ts1 <- rt(size,df)
shapiro.test(ts1)
```

```
size = 20
df =10
#random t-dist sample, (change the size and df)
ts1 <- rt(size,df)
shapiro.test(ts1)
size = 75
df =2
#random t-dist sample, (change the size and df)
ts1 <- rt(size,df)
shapiro.test(ts1)
size = 75
df = 10
#random t-dist sample, (change the size and df)
ts1 <- rt(size,df)
shapiro.test(ts1)

##EXPONENTIAL Distribution QQ Plot##

#random exponential distribution
#set.seed(1)
exp_data = rexp(100)
exp_df = data.frame(exp_data)

#distribution is exponential
di = "exp"
dp <- list(rate = 1)

#Control plot
ctrl_exp = ggplot(data = exp_df, aes(sample=exp_data)) +
  stat_qq_point(distribution = di, dparams = dp) +
  stat_qq_line(distribution = di, dparams = dp) +
  ggtitle("Control Exponential QQ Plot")

#Standard
stnd_exp <- ggplot(data = exp_df, mapping = aes(sample = exp_data)) +
  stat_qq_band(distribution = di, dparams = dp) +
  stat_qq_line(distribution = di, dparams = dp) +
  stat_qq_point(distribution = di, dparams = dp) +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
  ggtitle("Exponential QQ Plot with Confidence Bands")
```

```
#Detrended
de <- TRUE
det_exp <- ggplot(data = exp_df,mapping = aes(sample= exp_data)) +
  stat_qq_band(distribution = di, detrend = de) +
  stat_qq_line(distribution = di, detrend = de) +
  stat_qq_point(distribution = di, detrend = de) +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
  ggtitle("Detrended Exponential QQ Plot with Confidence Bands")
```

References

1. Adam Loy, Lendie Follett & Heike Hofmann (2016) Variations of *Q–Q* Plots: The Power of Our Eyes!, The American Statistician, 70:2, 202-214

2. Aldor-Noiman, S., Brown, L. D., Buja, A., Rolke, W., and Stine, R. A. (2013), "The Power to See: A New Graphical Test of Normality," The American Statistician, 67, 249–260

3. Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E. K., Swayne, D. F., and Wickham, H. (2009), "Statistical Inference for Exploratory Data Analysis and Model Diagnostics," Philosophical Transactions of the Royal Society, Series A, Mathematical, Physical and Engineering Sciences, 367, 4361–4383.

4. Ford C. (2015) Understanding Q-Q Plots. Research Data Services Sciences. https://data.library.virginia.edu/understanding-q-q-plots/