NormalityAssessment: An Interactive Classroom Tool for Testing Normality Visually

Christopher J. Casement and Laura A. McSweeney
Department of Mathematics, Fairfield University

Abstract

Many inferential and predictive statistical procedures possess underlying theoretical conditions that should be met in order for the results of those procedures to be considered reliable. One condition associated with methods for population means, including linear regression coefficients, is that of normality of a population(s). When assessing normality, two graphical tools that are often utilized are normal quantile-quantile (QQ) plots and histograms. However, while these tools are popular, they still present challenges for many who use them due to the subjectivity oftentimes involved when examining them. In this article, we describe a free, interactive Shiny application, downloadable as an R package, which implements two procedures recently developed for graphical inference with a specific emphasis on assessing normality. As the application was created and designed with a focus on statistics education, we also discuss a student assessment of it and, in the appendix, provide a suggested classroom activity.

Keywords: statistics education, graphical inference, assessing normality, Shiny

1. INTRODUCTION

Inferential statistical procedures, as well as certain predictive methods such as prediction intervals, have associated conditions for the statistical theory underlying the procedures to hold, and consequently for the results of an analysis to be considered reliable. One of the conditions for parametric hypothesis testing and confidence interval procedures for population means is that of normality. In the case of one- and two-sample procedures, the condition is that the variable is normally distributed in the population(s), while for ANOVA and linear regression, the condition is that the errors of the population model are normally distributed (although for ANOVA, it is also common to check the conditional distribution of the response variable given the explanatory variable(s)).

Both analytical and graphical methods exist for assessing normality. Commonly-used analytical methods include the Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling tests, among others, as well as measures involving skewness (e.g., the skewness ratio). For more in-depth discussions on tests for normality, and comparisons between them, see Dufour et al. (1998), Ghasemi and Zahediasl (2012), Razali and Wah (2011), Stephens (1974), Thode (2002), and Yap and Sim (2011). In terms of graphical tools, normal quantile-quantile (QQ) plots and histograms are widely used for checking normality. Less commonly used graphical tools for assessing normality, which will not be addressed in this paper, include probability plots, confidence bands,

and detrended normal QQ plots. See Gan et al. (1991) and Wilk and Gnanadesikan (1968) for discussions on probability plots; Aldor-Noiman et al. (2013) and Loy et al. (2016) for discussions on normal QQ plots with confidence bands; and Loy et al. (2016) for a discussion on detrended normal QQ plots.

Histograms and normal QQ plots are taught in many statistics courses and covered in textbooks such as those by De Veaux et al. (2020), McClave and Sincich (2021), Peck et al. (2020), Starnes and Tabor (2018), and Triola (2022). While both types of plots are especially popular and prominent among a wide array of statistical software programs, they do possess a drawback in that they can be difficult to assess, especially for individuals who have not worked with them extensively. In particular, for normal QQ plots, it can be challenging to determine whether the points in a plot overall fall "close enough" to the diagonal reference line regularly added to the plot, or whether they deviate "far" from the line. In fact, due to the natural variability of data, some deviations from the line are to be expected. With histograms, it can be challenging to decide whether or not the distribution plotted appears roughly normal, especially due to (1) the reliance of histograms on the number of bins used, which can alter the perceived shape of the distribution, and (2) the natural variability of sampled data, especially for small samples. The ability to understand the overall nature of deviations from what is expected is paramount for an accurate inspection of these two types of plots.

A substantial amount of literature has been devoted to the ability of students to correctly read and interpret histograms. In fact, various studies have been run to examine this, including those by Batanero et al. (2005), Chaput et al. (2021), delMas et al. (2005), Kaplan et al. (2014), and Lem et al. (2013, 2014). The studies found that students (and, in one case, some faculty) possess misconceptions about histograms. While assessing normality was not a part of most of the studies, the results indicate students struggle with various concepts related to histograms. And, through different methods of assessment, Batanero et al. (2005) found that students had difficulty identifying variables that were normal. An informative review of difficulties encountered when making or interpreting histograms, including an impressive list of sources (some of which are mentioned above), is provided by Boels et al. (2019).

We are not aware of any studies that examine the ability of students to correctly read and interpret individual normal QQ plots. However, through our own experiences in the classroom, we have found that some students struggle to assess normality using such plots. Loy et al. (2016), Loy (2021), and Stine (2017) attempt to address difficulties students have when working with normal QQ plots. In fact, Loy et al. (2021) and Stine (2017) even created software applications that implement the methods they describe. Still, we believe there remains a need for additional methods and tools for assessing normality, particularly for students.

The graphical tool described in this article was created to address common difficulties in evaluating both normal QQ plots and histograms. The tool, a Shiny application, is run locally in R (R Core Team 2021)—a free, widely-available statistical programming language—as a popup, point-and-click graphical user interface. To run the app, the user simply types two short lines of code (found later in this article) in R—one to load the package containing the app and the other to open the app—making the application widely accessible to many individuals, including those who

do not have any experience using R. This straightfoward design also makes it an ideal app to use in statistics classes.

The article proceeds as follows. We first discuss two graphical procedures for statistical inference—the Rorschach and line-up procedures—that form the foundation of the Shiny application. We then describe the *NormalityAssessment* app and provide an example of its use, followed by a discussion of a student assessment of the app. Lastly, we conclude with a summary of the article, and additionally provide a suggested classroom activity in the appendix.

2. GRAPHICAL INFERENCE METHODS

In this section we discuss two fairly recently-developed graphical inference methods, the Rorschach and line-up procedures (Buja et al. 2009; Wickham et al. 2010), upon which our Shiny application is based. The Rorschach procedure is aimed at assisting its users in developing a better understanding of the natural variability often present when working with real data, whereas a common goal of the line-up procedure is to help its users assess whether or not data plausibly came from a particular model, and it thus often acts as a graphical goodness-of-fit test. Wickham et al. (2010) also provide examples of types of plots and questions those plots can be used to answer. For instance, the authors mention how scatterplots can help one determine if a relationship exists between two variables, choropleth maps can help one assess if a spatial pattern exists, and QQ plots can help one determine if data plausibly came from a particular distribution (the focus of the tool described in this article), among others. We now discuss the Rorschach and line-up procedures more in depth.

2.1. Rorschach procedure

The objective of the Rorschach procedure is to improve a user's ability to understand the natural variability that often arises with data. When utilizing the procedure, P data sets of the same size are randomly generated from the same model. An appropriate plot (e.g., a bar chart or histogram) is then created for each data set, with the plots displayed side-by-side. The person running the procedure is tasked with examining each plot and comparing them, focusing on both expected and unexpected characteristics in each. Unexpected features displayed are the result of the natural variability in data, as each data set was randomly generated from the same model.

As an example, suppose someone wants to develop a better understanding of the variability they might see when working with a sample of size 50 from a normal distribution with a mean of 100 and a standard deviation of 20, commonly denoted $N(100, 20)$. Also suppose the individual specifies P, the number of plots (and thus the number of data sets of the same size), to be 12. The Rorschach procedure involves randomly generating 12 samples, each of size 50, from the $N(100, 20)$ distribution using, for instance, Monte Carlo simulation. A histogram (or some other appropriate type of plot) is then created for each simulated data set, and the 12 resulting histograms are placed in a grid format, as can be seen in Figure 1. None of the histograms displayed in Figure 1 exhibits the smooth bell shape one might expect, yet each sample was randomly obtained from the $N(100, 20)$ distribution. This is due to the natural variability present when drawing 50 observations from this distribution. Instead, deviations from the expected bell shape—bars that

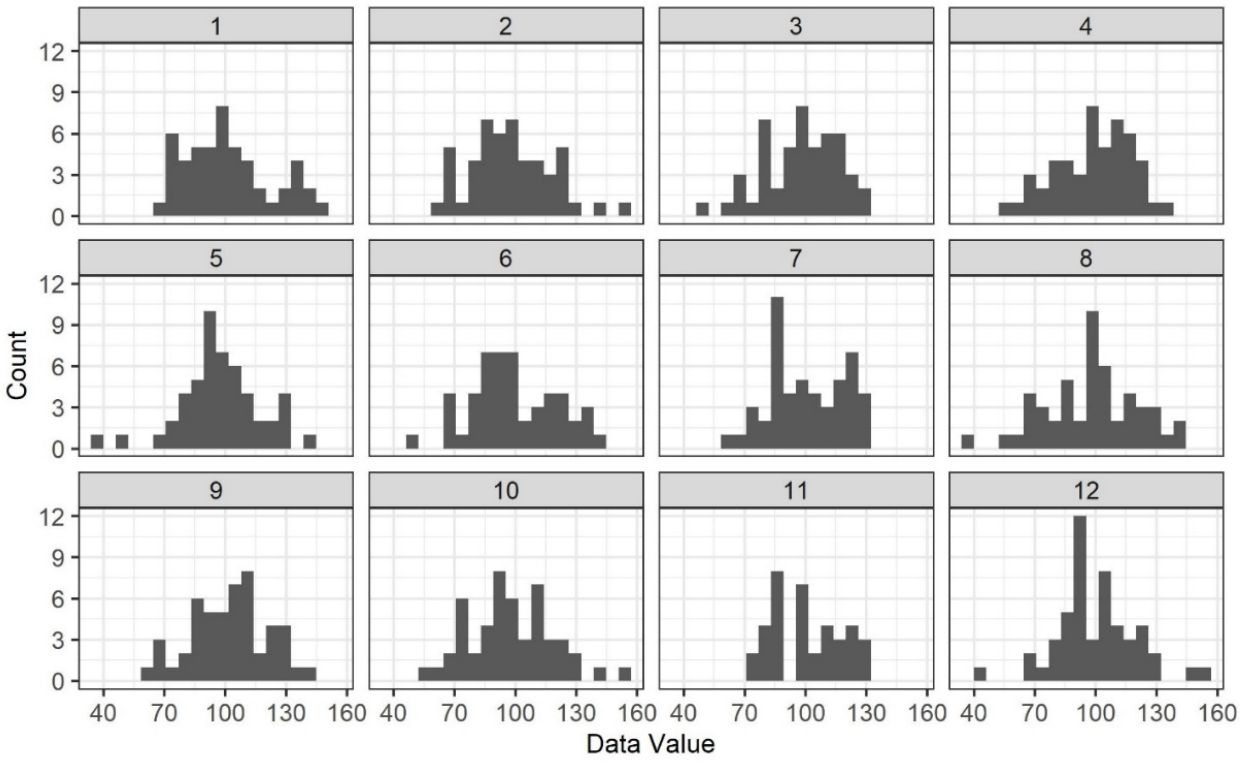

Figure 1. Rorschach histograms corresponding to 12 random samples of size 50, each from the $N(100, 20)$ distribution

extend either higher or lower than would be necessary to create a smoother bell shape—enable the user to improve their feel for the types of behavior they might see when randomly sampling a certain number of observations (e.g., 50) from a particular distribution (e.g., $N(100, 20)$). Further, the user becomes less likely to incorrectly conclude data do not come from a particular distribution when they see such variability. In fact, practicing with the Rorschach procedure prepares users for the line-up procedure, which relies on the user's ability to understand the natural variability in data.

2.2. Line-up procedure

A second graphical inference procedure, the line-up procedure, can operate as a graphical goodness-of-fit test. In such a scenario, like any other goodness-of-fit test, the goal is to test whether or not data came from a particular model. The null hypothesis states that the data came from a particular model, and the alternative hypothesis states that the data did not come from that model. As they do for traditional hypothesis tests, concepts such as power, errors, and p-values apply to the line-up procedure when it is used as a formal test. In this article, we will focus on p-values and the significance level when using the line-up procedure to draw conclusions based on user data. However, readers interested in thorough discussions of other statistical components (e.g., power) can refer to other sources, such as Hofmann et al. (2012), Loy et al. (2016), and Majumder et al. (2013).

While the Rorschach procedure focuses entirely on simulated data, the line-up procedure incorporates both real and simulated data. The first step in the line-up procedure involves fitting a model of interest to the real data (e.g., normal), including estimating any parameters the model requires (e.g., mean and standard deviation) using an applicable estimation method (e.g., maximum likelihood). Many textbooks that cover the mathematics of statistical inference, such as Casella and Berger (2002) and Hogg et al. (2020), provide discussions of estimation methods. After a model is fit, $P - 1$ data sets, each with the same number of observations as the real data, are randomly obtained from that model. An appropriate plot is created for each of these simulated data sets as well as one for the real data set, with the plot corresponding to the real data randomly mixed in when all P plots are presented in a grid to the user. The user is tasked with identifying the graph that corresponds to the real data. If they successfully do so, then they have evidence suggesting the real data did not come from the fitted model. On the other hand, if they are unable to choose the graph of the real data, then they do not have evidence suggesting the real data did not come from the fitted model, and it is plausible that the data did, in fact, come from that model.

As an example, suppose there are 250 real measurements of a quantitative variable and P is set to 12. The line-up procedure involves fitting a model to that data (in this case, the normal distribution with the same sample mean and standard deviation as the real data, though an alternative estimation method such as maximum likelihood could have been used instead), randomly generating 11 data sets of size 250 from the fitted model, and then creating a histogram (or another appropriate plot) of each simulated data set. The histogram of the real data is randomly mixed in with the 11 histograms of the simulated data, and the 12 plots are displayed in a grid, as can be seen in Figure 2. The user is then tasked with selecting the histogram of the real data from the 12 candidate histograms presented to them. For the solution to this example, see footnote [1].

2.3. Choosing the number of plots

While the Rorschach and line-up procedures both require the user to choose the total number of plots, P, selecting an appropriate P is particularly important for the line-up procedure. Buja et al. (2009) and Wickham et al. (2010) discuss this choice in terms of statistical and practical considerations, with Wickham et al. suggesting P be set to 20 in order to strike a reasonable balance. This value is sufficiently large in that it leads to a suitable p-value for the inferential procedure (as seen in an example later), while also being sufficiently small in that the user should be able to examine 20 plots without being overwhelmed by the number of plots.

2.4. Existing software for graphical inference

In addition to their discussion of the Rorschach and line-up procedures, Buja et al. (2009) implement both procedures in the R package *nullabor*. This valuable package provides users with various options for models that can be used and even allows for the computation of p-values and power. However, the package likely is inaccessible to many introductory statistics students due to its complexity and the fact that many such students do not have a background in R.

[1] Histogram 5 contains the real data. That distribution is right skewed, whereas the distributions displayed in the other 11 plots appear relatively symmetric.

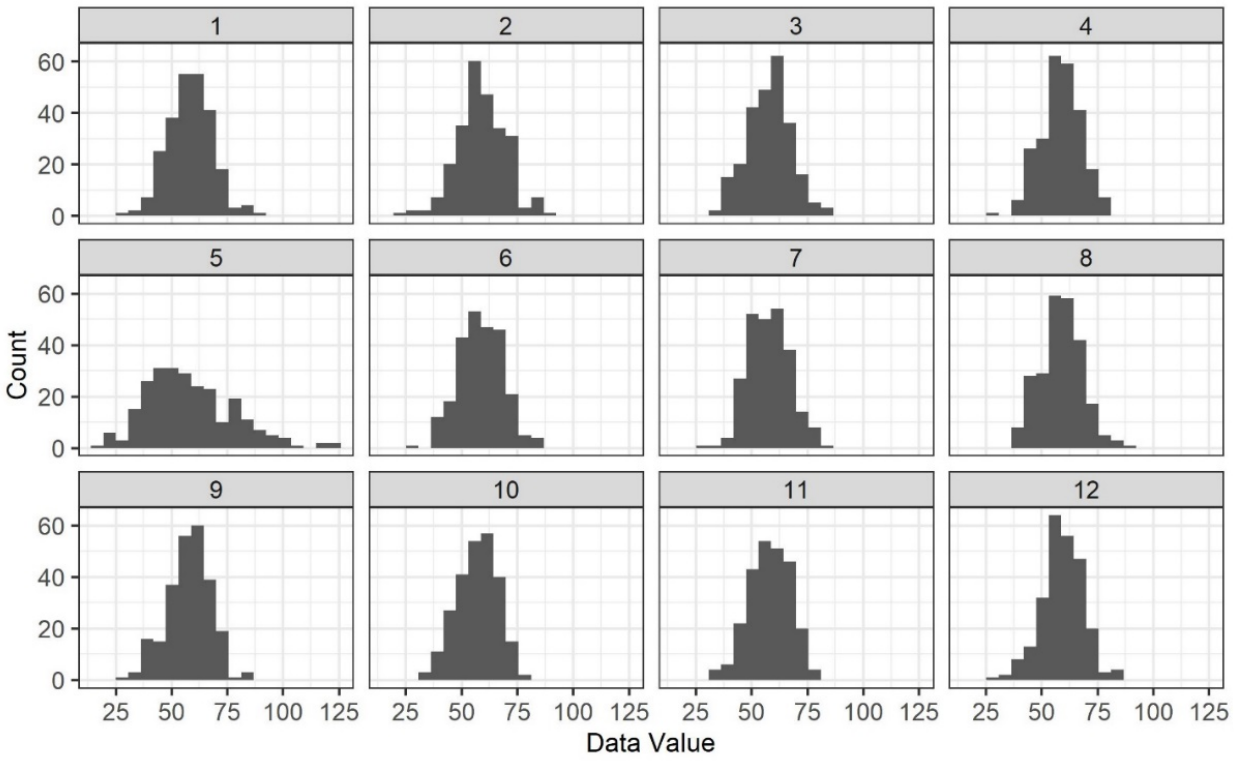

Figure 2. Line-up histograms: 11 of simulated data and one of real data randomly placed in the line-up. The goal is to determine which plot contains the real data.

Another R package that assists with graphical inference is *TeachingDemos* (Snow 2020), which, among other functionalities, enables users to run the line-up procedure using either histograms or normal QQ plots. While it has built-in functions aimed at simplifying the process for running the line-up procedure when single users are involved, *TeachingDemos* still requires users to be familiar with R. For both single- and multiple-user scenarios, students likely need to know how to import data; and for multiple-user scenarios, where plot reproducibility is imperative, they need to know how to set a seed, store objects, and access the contents of a list, a certain kind of data structure in R. Additionally, the functions that pertain to graphical inference focus on the line-up procedure but not the Rorschach procedure.

Loy (2021) also developed a collection of Shiny apps that enable users to run the line-up procedure with different types of plots (e.g., mosaic plots, side-by-side boxplots, and normal QQ plots). While these interactive apps are certainly useful, when it comes to assessing normality specifically, the app designed to do so (that which creates normal QQ plots) is limited to the line-up procedure for single users and does not provide an option to use histograms.

NormalityAssessment addresses various drawbacks to the existing tools for assessing normality via graphical inference procedures, particularly for students in introductory statistics classes, as we discuss further in Sections 3 and 4.

3. DESCRIPTION AND EXAMPLE OF THE NORMALITYASSESSMENT APPLICATION

We now describe the *NormalityAssessment* application and its ability to perform both the Rorschach and line-up procedures when assessing normality. The application, which possesses an interactive graphical user inference using the *Shiny* package (Chang et al. 2021), is downloaded and installed as an R package and runs locally on the user's machine. Strengthened by R packages such as *ggplot2* (Wickham 2016) and *rio* (Chan et al. 2021), among others, it enables users to import data sets from a variety of file formats associated with popular software used in wide-ranging disciplines (e.g., .csv, .mat, .rda, .RData, .rds, .sas7bdat, .sav, .txt, .xls, and .xlsx). The app was developed for use with two types of plots, normal QQ plots and histograms, that are commonly used for checking the normality condition associated with various parametric inferential and predictive methods. Additionally, the app enables either single or multiple users (working together as a group) to synergistically run the two graphical inference procedures discussed earlier in this article.

The user has the option of first running the Rorschach procedure to practice visualizing the natural variability in data randomly sampled from a chosen distribution (e.g., standard normal) as plotted in normal QQ plots and/or histograms. Once they have developed a better understanding of such variability, the user proceeds to the line-up stage, where they can either input their own data or use one of the data sets included in the app, and are tasked with selecting the normal QQ plot (or histogram) that corresponds to their data from a grid that contains multiple plots, all but one of which are of simulated data. This two-stage process enables educators and students at all levels and in wide-ranging disciplines to assess normality when teaching statistics and running statistical analyses.

When further describing the app, we first explain its capabilities when working solely with simulated data (the Rorschach stage) and follow that with a discussion of the app's usage with both simulated and real data (the line-up stage). We also demonstrate the usage of the app by providing a step-by-step example where we move through the full training and selection processes harmoniously using data from the *Highway1* data set that is a part of the *carData* package (Fox et al. 2020). This data set contains data related to automobile accidents for 39 segments of large highways in Minnesota in 1973. We use linear regression to model the accident rate per million vehicle miles (the *rate* variable) based on the speed limit for the segment of the highway (the *slim* variable) for those 39 segments. We then calculate the 39 residuals[2] based on the fitted linear model, add them to the data set, and save the new version of the data set as a CSV file (*Highway1_new.csv*) in the working directory. We will then assess these residuals for normality. To ensure reproducibility, the following R code accomplishes this, though any statistical package can be used to generate the residuals:

```
> # load carData package
> library(carData)

> # fit linear model and store residuals
> Highway1$residuals <- lm(rate ~ slim, data = Highway1)$residuals
```

[2] The data set contains no missing data.

```
> # save data set
> write.csv(Highway1, file = "Highway1_new.csv", row.names = FALSE)
```

3.1. Installing and launching the application

Users must install R, and optionally RStudio, on their computer in order to use the *NormalityAssessment* application. Experienced R users can install the *NormalityAssessment* package, which contains the application, from the Comprehensive R Archive Network (CRAN). However, for those who either have not used R before or are still developing their level of comfort with it, instructions for installing and opening R (and RStudio), as well as others for installing and loading packages, can be found through various sources. A free version of an R programming book written by Grolemund (2014), which can be found on the book's website located at https://rstudio-education.github.io/hopr/starting.html, is one such source, with detailed instructions found in its appendices. After installing the *NormalityAssessment* package on their machine, followed by opening either the R graphical user interface or RStudio, and then loading the package into the current R session, the user launches the Shiny application locally by running the following function from the package, with the code typed exactly as it appears (aside from the italics): *runNormalityAssessmentApp()*. The user is immediately taken to the "Home" tab, which describes the features contained in the app's other tabs. This tab can be seen in Figure 3.

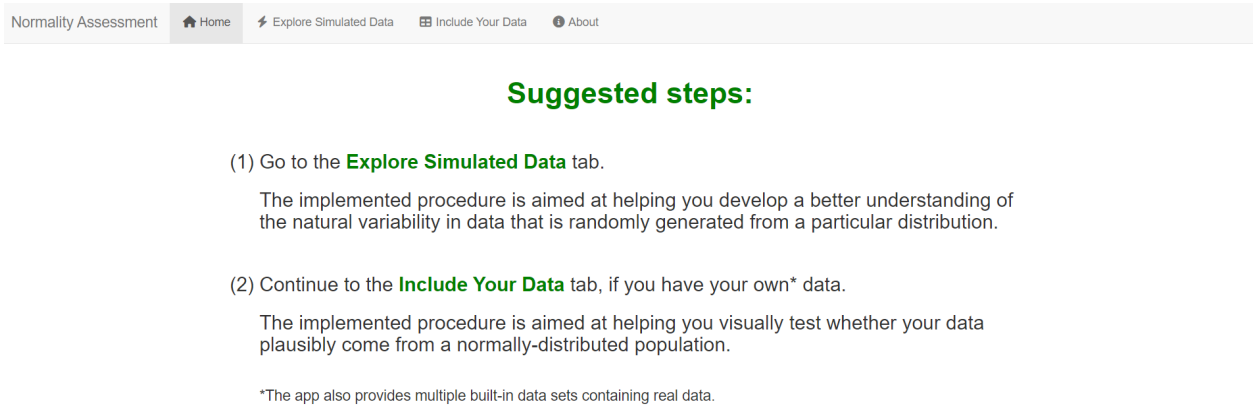

Figure 3. The "Home" tab of the *NormalityAssessment* app includes an overview of the features offered in the two main tabs

3.2. Working with simulated data only

To begin, it is recommended that the user navigates to the "Explore Simulated Data" tab, where they can run the Rorschach procedure with simulated data. However, they can proceed directly to the "Include Your Data" tab to run the line-up procedure for their own data, if they so choose.

To run the Rorschach procedure, the user makes multiple selections. First, they choose the distribution of the population from a list of options. For anyone who is relatively new to statistics, we suggest choosing the "basic" option, which asks the user to select a general shape for the population distribution. The possible shapes range from severely left skewed, to bell-shaped

("normal" in the app), to severely right skewed. For those with a deeper knowledge of distributional families, the "advanced" option enables the user to select from a list of common families and input the respective parameter values. These families include normal, *t*, *F*, uniform (continuous), χ^2, exponential, gamma, and beta. After selecting the population distribution, the user inputs the sample size, the total number of plots that will be displayed, and the type of plot (normal QQ plot and/or histogram). If histograms are selected, the app provides the user with a slider input to change the number of bins. This dynamic feature is important, as the number of bins used can alter the shape of the resulting histogram, a potential drawback to using histograms to check normality in general, especially for small samples. Once the plots are displayed in a grid, the user is tasked with identifying both expected and unexpected features in each plot, as described previously. The user has the option to view the grid in a larger window.

In the linear regression example introduced before, recall there are 39 segments of highways studied, leading to 39 residuals after fitting the linear model. The normality condition will be assessed using these residuals. The user would then choose the following in the app: a normally distributed population, a sample size of 39, and the number of normal QQ plots or histograms to be displayed. Assuming they choose 20 normal QQ plots, 20 samples, each of size 39, are then randomly generated from the standard normal distribution, and a normal QQ plot corresponding to each random sample is created. A grid of 20 such plots can be seen in Figure 4. None of the plots contains points that all fall perfectly along the diagonal line, yet each sample was, in fact, generated from a normal population. The expected features that appear in the plots (points directly on or quite close to the line) and unexpected features (points that deviate substantially from the line) help the user develop a better understanding of natural variability in the data values when 39 observations are randomly sampled from the standard normal distribution. The user is also able to continue to practice with the Rorschach procedure by repeating the process, including changing inputs to try different distributions if they wish.

3.3. Working with simulated and real data

After training with the Rorschach procedure, the user moves to the "Include Your Data" tab, which allows them to run the line-up procedure. There, in the "Input Data and Users" tab, the user inputs their own data, either manually or by uploading a file, and selects whether they are working individually (one user) or with others (multiple users). We provide built-in data sets for the user, if needed.

Returning to the linear regression example, suppose the user uploads the *Highway1_new* data set. Figure 5 shows the full data set uploaded, with the exception of any non-numeric variables (such as *htype* in this case). The user then continues to the "Make Plots" tab.

Single-user scenarios

When working individually, the user selects (1) the variable of interest, (2) the total number of plots, P, and (3) the type of plot (normal QQ plot or histogram), as seen in Step 1 in Figure 6. We'll assume that the user selected the normal QQ plot option. Next, $P - 1$ samples, each with the same number of observations as the user's data, are randomly drawn from the normal distribution with a mean and standard deviation that are the same as the sample mean and standard

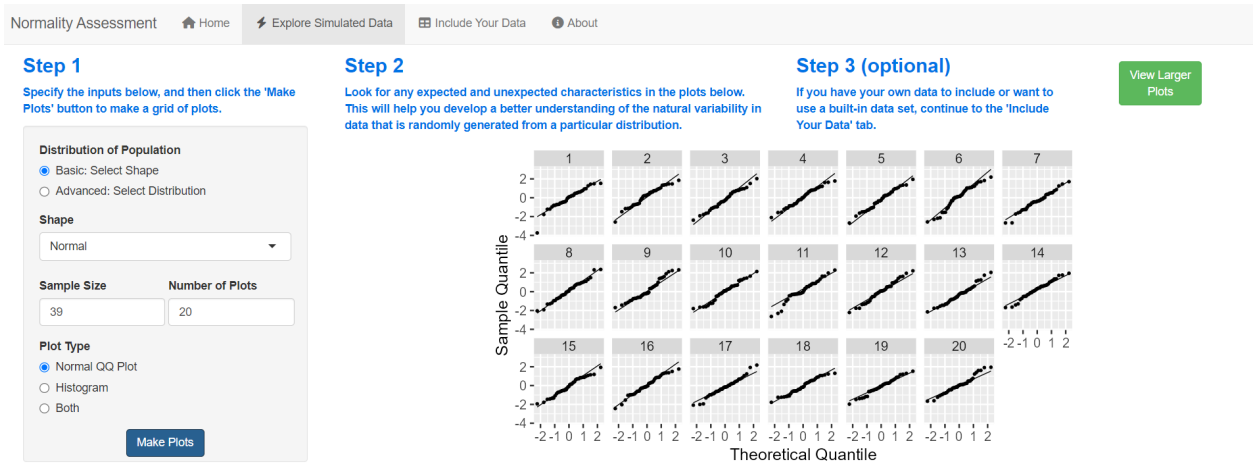

Figure 4. Rorschach normal QQ plots corresponding to 20 random samples of size 39, each from the standard normal distribution

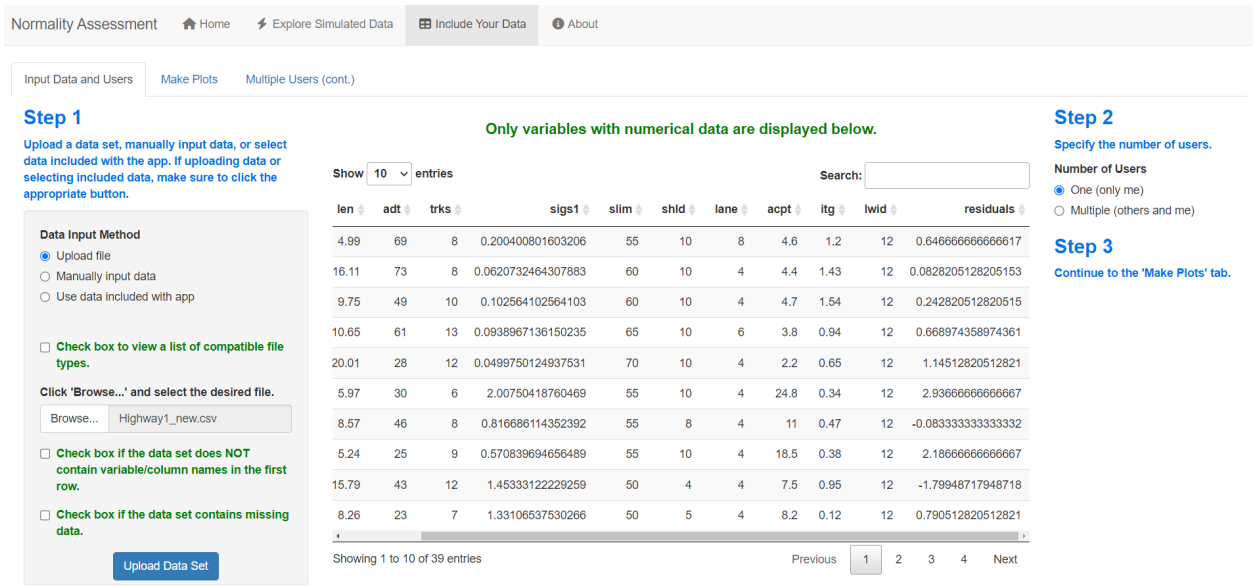

Figure 5. The *NormalityAssessment* app after the user has uploaded the *Highway1_new* data set and specified they are working individually as opposed to in a group setting

deviation[3] of the user's data. If the simulated data did not come from the model with the same mean and standard deviation as the actual data, then the sample quantiles (found on the y-axis) of the normal QQ plot corresponding to the actual data would likely be on an entirely different scale than that of the simulated data, making it a trivial exercise for the user to select the graph corresponding to their data. In fact, this could potentially invalidate any conclusions drawn using the line-up procedure as a graphical goodness-of-fit test, as the user might incorrectly conclude

[3] The choice of statistics to use, in particular the standard deviation here, is an important consideration when estimating parameters. We remark on this in Section 5.

Figure 6. Inputs and output for 20 assumed normal QQ plots and a single user. One plot corresponds to the real residuals, while the other 19 correspond to data randomly generated from the $N(\text{mean}_{\text{real}}, \text{SD}_{\text{real}})$ distribution.

their data do not plausibly come from a normal population simply due to the difference in scales used.

Returning to the example from before, suppose the user chooses 20 normal QQ plots with the *residuals* variable (the residuals from the user's fitted linear model that are used to assess the normality condition). The app creates 19 plots, each corresponding to 39 randomly sampled observations from the $N(\text{mean}_{\text{real}}, \text{SD}_{\text{real}})$ distribution, where $\text{mean}_{\text{real}}$ and SD_{real} represent the sample mean and standard deviation of *residuals*. The sample mean and standard deviation of *residuals* are 0 (as expected based on statistical theory) and 1.45, respectively, so each of the 19 plots of simulated data is a data set randomly generated from the $N(0, 1.45)$ distribution. The app also produces the normal QQ plot corresponding to the real residuals and then randomly places it among the 19 other plots, as can be seen in Figure 6. If the user correctly identifies the plot that corresponds to the real residuals (see footnote [4] for the solution), then, with a p-value of $1/P = 1/20 = 0.05$, they have evidence suggesting the errors of the linear regression model are not normally distributed. On the other hand, if they are unable to correctly identify that plot, then it is plausible the errors are normally distributed.

Multiple-user scenarios

When users are working in groups (such as in a classroom setting), each person makes identical selections for the same three inputs as when working individually (the variable of interest, the total number of plots, and the type of plot) as well as a fourth input—a seed that ensures all users will be presented with the exact same set of plots in the line-up even if working on different devices. Figure 7 displays these inputs as well as the updated instructions for the multiple-user scenario.

[4] Plot 12 is the normal QQ plot corresponding to the user's residuals.

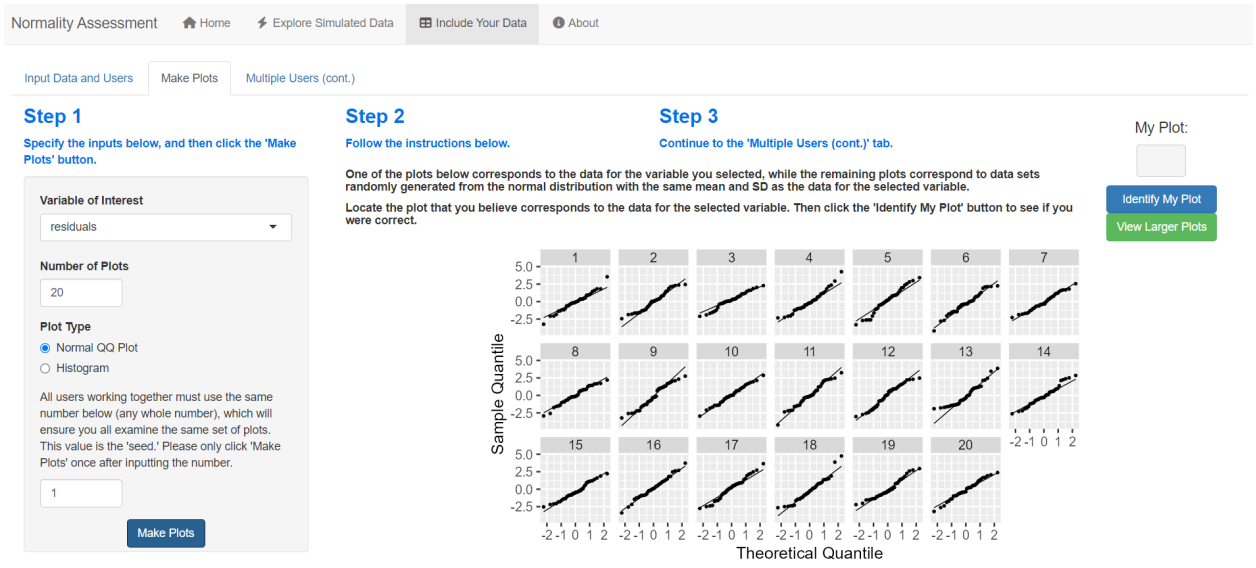

Figure 7. Inputs and output assuming 20 normal QQ plots and multiple users working as a group. One plot corresponds to the real residuals, while the other 19 correspond to data randomly generated from the $N(\text{mean}_{\text{real}}, \text{SD}_{\text{real}})$ distribution.

Each person independently completes the line-up procedure. After identifying the correct plot, the users navigate to the "Multiple Users (cont.)" tab. Once they have discussed how many of the group members correctly identified the plot of the real data, they input the number of users in their group, the number of those users who correctly identified the plot, and the number of plots each user viewed. An adjusted *p*-value is then calculated based on the binomial model, something discussed further in Wickham et al. (2010) and Majumder et al. (2013). The users would then use this *p*-value in the typical way when testing whether or not the real data plausibly came from a normal population. A potential outcome assuming three users can be seen in Figure 8.

We now briefly discuss the choice of model used when calculating the *p*-value. The binomial model discussed in Wickham et al. (2010) and Majumder et al. (2013) requires strong conditions, including that of independence of the selections made by the users. VanderPlas et al. (2021) argue, however, that the selections are not, in fact, independent, and that the dependence must be accounted for in order to obtain a more conservative and subsequently more accurate *p*-value. They propose using a beta-binomial model instead of the binomial model for the *p*-value calculation when multiple users examine the same set of line-up plots. Hofmann et al. (2020) have implemented the beta-binomial model, as well as the more general Dirichlet-multinomial model for additional situations that can arise in line-up testing, in the *vinference* R package. While we recognize the value of the beta-binomial model, especially for use in practice (such as research settings), the full procedure for calculating a *p*-value requires a combination of choices and steps that are well beyond the scope of most introductory statistics courses and which might cause confusion for students and educators for whom the *NormalityAssessment* application was developed. Additionally, due to the application's intended use in low-stakes educational settings, such as small-group projects, where there is no intent to publish any findings, we instead implement the binomial model and use that to calculate a *p*-value when multiple users are involved.

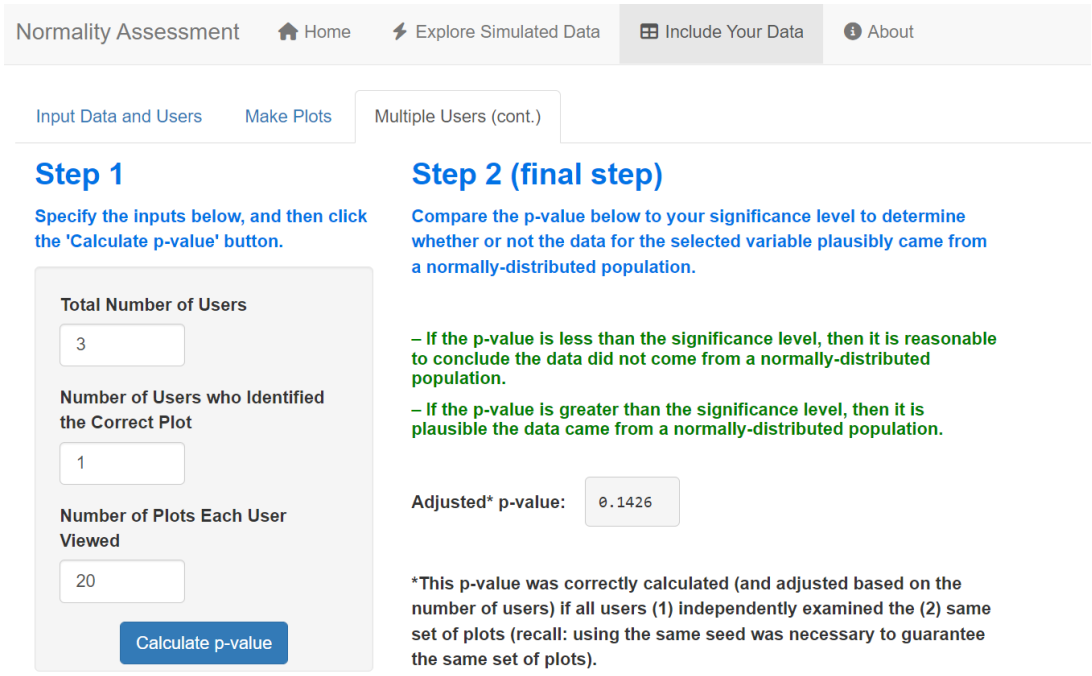

Figure 8. Inputs and output assuming three users (working as a group) independently examined the same set of 20 plots, and only one identified the correct plot

4. ASSESSMENT OF THE APPLICATION

We now describe the results of a survey that was given to students who did an activity using the *NormalityAssessment* app. This activity was similar to the first sample provided in the appendix. The students assessed the normality of data provided to them using a variety of methods in the *NormalityAssessment* app, including one histogram, one normal QQ plot, and the line-up method with 20 histograms or 20 normal QQ plots, after training using the "Explore Simulated Data" tab with the Rorschach plots.

Six students from a math major Applied Statistics I class and 24 Master's students from a similar Applied Statistics I class completed the activity and survey in fall 2020. The survey contained a combination of Likert-scale questions and open-ended questions, and was exempt from review by the Fairfield University Institutional Review Board. It also included a required question that asked participants if they give consent for their responses to be used anonymously for this study to evaluate the application.

The survey contained the following Likert-scale statements about the app:

(1) The Normality Assessment app was easy to use.
(2) I encountered problems when using the Normality Assessment app.
(3) The "Explore Simulated Data" tab was helpful to see the natural variability in the plots of simulated data drawn from the same distribution.
(4) Using the set of 20 histograms in the "Include Your Data" tab was helpful when assessing normality.

(5) Using the set of 20 Normal QQ plots in the "Include Your Data" tab was helpful when assessing normality.
(6) I do not see myself using the app in the future when I need to assess normality.

The results are summarized in Table 1. Note that Questions 2 and 6 are phrased negatively so that disagreement responses are actually positive (*). Question 3 asks about the Rorschach plots, and Questions 4 and 5 ask students to assess the line-up method.

Question	Strongly Disagree	Somewhat Disagree	Neither Disagree nor Agree	Somewhat Agree	Strongly Agree
Q1: Easy to use	0	0	13.3	53.3	33.3
Q2*: Encountered problems	50.0	23.3	10.0	16.7	0
Q3: Explore tab helpful	0	0	23.3	53.3	23.3
Q4: 20 histograms helpful	3.3	3.3	20.0	40.0	33.3
Q5: 20 normal QQ plots helpful	0	0	16.7	46.7	36.7
Q6*: Do not see myself using app	6.7	30.0	40.0	20.0	3.3

Table 1. Percent of student responses ($n = 30$) to Likert-scale statements about the *NormalityAssessment* application

The survey then provided free response questions to obtain additional information about the student experience. Responses to the questions can be found below.

What students liked about the app:

- App was easy to use [$n = 7$]
- User friendly [$n = 3$]
- "It worked as expected."
- "The interface."
- "Usability."
- "It's [sic] simplicity."
- "Being able to input my own data and to generate the visuals of the data is useful."
- "I did like how it had mult[iple] options to change between distribution shape and testing them with different graphs."
- "I like how visual the app is."
- "Helpful to compare other shapes."
- "I liked the amount of practice that I could get in a short period of time using it, as well as the visual learning."

- "It shows how histograms can look not normal even if they do come from a normal distribution."
- "The graphics are informative."
- "While using the program, I saw that I did not deal with calculations and that I could reach the result easily."
- "Quick visual response."
- "Ease of identifying the plots."
- "I liked that I could click buttons see the results change."
- "That it was easy to find how to filter the data you wanted."
- "Just a few clicks and the data was clean, organized, and interesting."

What students did not like about the app and suggestions for improvement:

- Nothing or N/A [$n = 12$]
- "I honestly do not know what to improve the app on, as long as the instructions are clear and concise it should be fine."
- "Could be a little complicated without the step by step explanations."
- "I didn't find the 20 chart display all that helpful because the charts were so small."
- "I feel having to view the 20 plots together looked a bit clustered and hard to read."
- "I feel like bigger look at the histograms and plots."
- "I can't tell the difference in the graphs."
- "I was a little unclear about what I was looking for when it switched to the multiple graphs, and how the variables were changing by increasing the numbers each time."
- "Theres [sic] some features that do not work like the density curve on histograms."
- "Lack of creativity."

The results of the survey indicate that a majority of students found the app easy to use and felt that both the Rorschach and line-up plots were helpful when assessing the normality criterion. Most students did not have difficulty with the app, though a few did, and in response to their feedback we made updates to the app to enhance the user experience. Namely, we expanded the size of the grid of plots to improve viewing and added more-detailed instructions on each tab to assist the user through the process.

5. CONCLUSION

Normal QQ plots and histograms are widely used for checking the condition of normality associated with various inferential and predictive methods. However, assessing such plots remains a challenging task, particularly for those lacking experience using them. In this article, we describe the free *NormalityAssessment* Shiny application, which is aimed specifically at addressing this shortcoming by building upon recent advances made in graphical inference. We also provide feedback left by students who used the app. While the feedback was quite positive and helped improve the app, we see potential areas for future work. One is a study of the impact of the Rorschach and line-up procedures on a student's ability to accurately check for normality. Another

is an assessment in which different methods for estimating the variability of the hypothesized model in the line-up procedure are used and compared, something Loy et al. (2016) consider in a study.

For use of the app in the classroom, we provide a suggested interactive activity, which can be completed on an individual or group basis, in the appendix. Our hope is that the app and activity are helpful for instructors and students of statistics at all levels and in wide-ranging fields.

6. ACKNOWLEDGEMENTS

We thank the editor and anonymous reviewers for their thoughtful and helpful comments. We also thank the students who used the app and provided valuable feedback.

7. REFERENCES

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

Batanero, C., Tauber, L. M., & Sánchez, V. (2004), "Students' Reasoning about the Normal Distribution." In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 257-276), Dordrecht: Springer.

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2019), "Conceptual Difficulties When Interpreting Histograms: A Review," *Educational Research Review*, 28.

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

Casella, G., & Berger, R. L. (2002), *Statistical Inference (2nd ed.)*, Boston, MA: Cengage Learning.

Chan, C. H., Chan, G., Leeper, T., & Becker, J. (2021), rio: A Swiss-army Knife for Data File I/O, R package version 0.5.27.

Chang, W., Cheng, J., Allaire, J.J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021), shiny: Web Application Framework for R, R package version 1.7.1.

Chaput, J. S., Crack, T. F., & Onishchenko, O. (2021), "What Quantity Appears on the Vertical Axis of a Normal Distribution? A Student Survey," *Journal of Statistics and Data Science Education*, 29, 192-201.

delMas, R., Garfield, J., & Ooms, A. (2005), "Using Assessment Items to Study Students' Difficulty Reading and Interpreting Graphical Representations of Distributions," *Proceedings of the 4th International Research Forum on Statistical Reasoning*, Auckland, New Zealand: University of Auckland.

De Veaux, R., Velleman, P., & Bock, D. (2020), *Stats: Data and Models (5th ed.)*, Hoboken, NJ: Pearson.

Dufour, J., Farhat, A., Gardiol, L., & Khalaf, L. (1998), "Simulation-based Finite Sample Normality Tests in Linear Regressions," *The Econometrics Journal*, 1, C154-C173.

Fox, J., Weisberg, S., & Price, B. (2020), carData: Companion to Applied Regression Data Sets, R package version 3.0-4.

Gan, F., Koehler, K., & Thompson, J. (1991), "Probability Plots and Distribution Curves for Assessing the Fit of Probability Models," *The American Statistician*, 45, 14-21.

Ghasemi, A., & Zahediasl, S. (2012), "Normality Tests for Statistical Analysis: A Guide for Non-Statisticians," *International Journal of Endocrinology and Metabolism*, 10, 486-489.

Grolemund, G. (2014), *Hands-On Programming with R: Write Your Own Functions and Simulations*, O'Reilly Media, Inc. https://rstudio-education.github.io/hopr/starting.html

Hofmann, H., Follett, L., Majumder, M., & Cook, D. (2012), "Graphical Tests for Power Comparison of Competing Designs," *IEEE Transactions on Visualization and Computer Graphics*, 18, 2441-2448.

Hofmann, H., VanderPlas, S., Röttger, C., & Cook, D. (2020), vinference: Inference Under the Lineup Protocol, R package version 1.0.0. https://github.com/heike/vinference

Hogg, R. V., Tanis, E. A., & Zimmerman, D. (2020), *Probability and Statistical Inference (10th ed.)*, Pearson.

Kaplan, J. J., Gabrosek, J. G., Curtiss, P., & Malone, C. (2014), "Investigating Student Understanding of Histograms," *Journal of Statistics Education*, 22.

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013), "On the Misinterpretation of Histograms and Box Plots," *Educational Psychology*, 33, 155-174.

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2014), "Interpreting Histograms. As Easy As It Seems?" *European Journal of Psychology of Education*, 29, 557-575.

Loy, A. (2021), "Bringing Visual Inference to the Classroom," *Journal of Statistics and Data Science Education*, 29, 171-182.

Loy, A., Follett, L., & Hofmann, H. (2016), "Variations of Q-Q Plots: The Power of Our Eyes!" *The American Statistician*, 70, 202-214.

Majumder, M., Hofmann, H., & Cook, D. (2013), "Validation of Visual Statistical Inference, Applied to Linear Models," *Journal of the American Statistical Association*, 108, 942-956.

McClave, J., & Sincich, T. (2021), *Statistics (13th ed.)*, Hoboken, NJ: Pearson.

Peck, R., Short, T., & Olsen, C. (2020), *Introduction to Statistics & Data Analysis (6th ed.)*, Boston, MA: Cengage.

R Core Team. (2021), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Razali, N., & Wah, Y. (2011), "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lillefors and Anderson-Darling Tests," *Journal of Statistical Modeling and Analytics*, 2, 21-33.

Snow, G. (2020), TeachingDemos: Demonstrations for Teaching and Learning, R package version 2.12.

Starnes, D. S., & Tabor, J. (2018), *The Practice of Statistics (6th ed.)*, New York: W. H. Freeman and Company.

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

Stine, R. A. (2017), "Explaining Normal Quantile-Quantile Plots Through Animation: The Water-Filling Analogy," *The American Statistician*, 71, 145-147.

Thode, H. (2002), *Testing for Normality*, New York: Marcel Dekker, Inc.

Triola, M. (2022), *Elementary Statistics (15th ed.)*, Hoboken, NJ: Pearson.

VanderPlas, S., Röttger, C., Cook, D., & Hofmann, H. (2021), "Statistical Significance Calculations for Scenarios in Visual Inference," *Stat*, 10, e337.
Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis (2nd ed.)*, New York: Springer.
Wickham, H., Cook, D., Hofmann, H., & Buja, A. (2010), "Graphical Inference for Infovis," *IEEE Transactions on Visualization and Computer Graphics*, 16, 973-979.
Wilk, M., & Gnanadesikan, R. (1968), "Probability Plotting Methods for the Analysis of Data," *Biometrika*, 55, 1-17.
Yap, B. W., & Sim, C. H. (2011), "Comparisons of Various Types of Normality Tests," *Journal of Statistical Computation and Simulation*, 81, 2141-2155.

A. APPENDIX

A.1. Individual version of activity

While the Shiny application is freely available for anyone to use, it was designed with a particular focus on supporting statistics education for students and instructors. We suggest instructors devise an activity using the app to aid students in developing their ability to examine normal QQ plots (and/or histograms) when assessing normality, after having introduced such plots. One potential in-class activity, which is described below, follows the first five of the six recommendations made in the GAISE report (Carver et al. 2016), and can even follow the sixth if carefully incorporated into an assessment. While it was designed to be an in-class activity completed (mostly) individually, the instructor can modify it to be completed outside of class or as a group assignment. The group version of this activity, which includes an essential addition regarding *p*-values, follows. The individual version of the activity includes the following steps:

- *Class session 1 – instructions for the instructor:*
  (1) Ask each student to collect real data for an analysis where using normal QQ plots or histograms to check normality is important. This might be a hypothetical analysis or one they will actually conduct. (In this activity we focus on normal QQ plots, though the instructions can be adapted for histograms.)
  (2) Ask each student to send you their data (preferably as a CSV file) so that you can access it during the next class session.
  (3) Ask each student to download and install R (and RStudio, if desired) and the *NormalityAssessment* package in R.

- *Class session 2 – instructions for the students:*
  (1) Open the R graphical user interface or RStudio on your computer.
  (2) Load the *NormalityAssessment* package in R by typing the following code, exactly as it appears, and then hitting "Enter": library(NormalityAssessment)
  (3) Open the *NormalityAssessment* application by running the following code from the *NormalityAssessment* package, with the code typed exactly as it appears, and then hitting "Enter": runNormalityAssessmentApp()

  You'll start by working with normal QQ plots for simulated data only in steps 4 through 8. This will help you learn how to examine this type of plot.

(4) Go to the *Explore Simulated Data* tab.
(5) Don't change any of the default inputs for now, but note the population distribution and sample size used. Then click the "Make Plots" button. What expected and unexpected features do you see in each plot?
(6) Change the shape of the population distribution to "Severely Left Skewed" and leave all other inputs unchanged. Again keep track of the sample size, and make a new set of normal QQ plots. What expected and unexpected features do you see in each plot?
(7) Repeat step 6 for two more shapes of the population distribution.
(8) Repeat steps 5 through 7, but now for different sample sizes, such as 10, 50, and 100.

You'll now work with the data you found as well as simulated data in steps 9 through 13. This will help you check the normality condition for your analysis.

(9) Click the *Include Your Data* tab and input your data in the *Input Data* section.
(10) Make sure "One (only me)" is selected for "Number of Users" on the right-hand side of your screen.
(11) Click the *Make Plots* tab. Keep the number of plots at 20 and the "Normal QQ Plot" option selected for the plot type, and click the "Make Plots" button.
(12) Follow the instructions above the plots, but don't click "Identify My Plot" yet.
(13) After trying to identify the plot that corresponds to your data, click the "Identify My Plot" button. Is the plot you chose the one that corresponds to your data? If so, what does that indicate? If not, what does that indicate? You've now checked the normality condition for your analysis.

After everyone has completed step 13, we'll come together and practice more as a class.

- *Class session 2 – instructions for the instructor:*
  (1) Have each student's data set readily available to work with as a class.
  (2) Answer any questions students have during steps 1-13 of the activity.
  (3) After all students have completed step 13, come together as a class.
  (4) Using the data collected by a few students, one data set at a time, go through the line-up stage as a class. Each time, ask all students in the class to think about which graph they believe corresponds to the actual data and why. Discuss as a class.
  (5) Remind students that if they were not able to correctly identify the plot of their data, that does not mean they would conclude their data definitively did not come from a normal population. Or on the other hand, if they were able to correctly identify the plot of their data, that does not mean they have evidence suggesting their data came from a normal population.

A.2. Group version of activity

The suggested activity described above was developed to be completed on an individual basis. However, it also can be adjusted to be completed in groups. While most of the steps remain quite similar or even the same, an important change must be made when calculating the *p*-value for the

line-up procedure, since this value depends on the number of users. For details on the calculation of the *p*-value, see Section 3.5. The full list of steps for the group activity is as follows:

- *Class session 1 – instructions for the instructor:*
  (1) Ask each group to collect real data for an analysis where using normal QQ plots or histograms to check normality is important. This might be a hypothetical analysis or one they will actually conduct. (In this activity we focus on normal QQ plots, though the instructions can be adapted for histograms.)
  (2) Ask each group to send you their data (preferably as a CSV file) so that you can access it during the next class session.
  (3) Ask each student to download and install R (and RStudio, if desired) and the *NormalityAssessment* package in R.

- *Class session 2 – instructions for the students:*
  (1) Break into your groups.
  (2) Open the R graphical user interface or RStudio on each group member's computer.
  (3) Load the *NormalityAssessment* package in R by typing the following code, exactly as it appears, and then hitting "Enter": library(NormalityAssessment)
  (4) Open the *NormalityAssessment* application by running the following code from the *NormalityAssessment* package, with the code typed exactly as it appears, and then hitting "Enter": runNormalityAssessmentApp()

  You'll start by working with normal QQ plots for simulated data only in steps 5 through 9. This will help you learn how to examine this type of plot. We recommend that all group members do these steps together on one group member's computer.

  (5) Go to the *Explore Simulated Data* tab.
  (6) Don't change any of the default inputs for now, but note the population distribution and sample size used. Then click the "Make Plots" button. What expected and unexpected features do you see in each plot? First think about this on your own, and then discuss with your groupmates. Note that each group member might have different plots if students work on their own computer.
  (7) Change the shape of the population distribution to "Severely Left Skewed" and leave all other inputs unchanged. Again keep track of the sample size, and make a new set of normal QQ plots. What expected and unexpected features do you see in each plot? First think about this on your own, and then discuss with your groupmates.
  (8) Repeat step 7 for two more shapes of the population distribution.
  (9) Repeat steps 5 through 7, but now for different sample sizes, such as 10, 50, and 100.

  You'll now work with both your group's data and simulated data in steps 10 through 18. This will help you check the normality condition for your group's analysis. It is important that each group member does the steps independently on their own computer.

(10) Click the *Include Your Data* tab and input your group's data in the *Input Data* section.
(11) Click "Multiple (others and me)" under "Number of Users" on the right-hand side of your screen.
(12) Click the *Make Plots* tab. Keep the number of plots at 20 and the "Normal QQ Plot" option selected for the plot type. For the seed input, you may choose any whole number (minimum of 1), but it's important that each group member uses that same number. Then click the "Make Plots" button.
(13) Before beginning this step, understand it's very important to complete it on your own. Follow the instructions above the plots, but don't click "Identify My Plot" yet.
(14) After doing your best to identify the plot that corresponds to your group's data, click the "Identify My Plot" button. Is the plot you chose the one that corresponds to your group's data? If so, what does that indicate? If not, what does that indicate? After each person in your group has completed this step, discuss your findings as a group.
(15) Decide on the significance level (e.g., 0.05) that you will use for the upcoming hypothesis test. Record this number. The test will involve the following two hypotheses:
H_0: The data came from a normal population.
H_a: The data did not come from a normal population.
(16) Click the *Multiple Users (cont.)* tab. Input three values: (i) the number of people in your group who completed the previous step, (ii) the number of people in your group who successfully identified the plot corresponding to your group's data when doing so individually, and (iii) the number of plots each person used (recall you were asked to use 20). Then click the "Calculate p-value" button.
(17) The output from step 16 is your group's *p*-value when testing. Record this number.
(18) Compare your group's *p*-value from step 16 to the significance level from step 15. If the *p*-value is smaller than the significance level, then there's evidence suggesting your data did not come from a normal distribution (and that the normality condition for your analysis is violated). If the *p*-value is larger than the significance level, then it is plausible your group's data came from a normal distribution, and it is safe to continue as if the normality condition is reasonably met. (Note: that doesn't mean you conclude your data came from a normal distribution, as that would be the equivalent of accepting the null hypothesis from step 15, which you know is never an option when running a hypothesis test.)

After each group has completed step 18, we'll come together and practice more as a class.

- *Class session 2 – instructions for the instructor:*
  (1) Have each group's data readily available to work with as a class.
  (2) Answer any questions students have during steps 1-18 of the group activity.
  (3) After all groups have completed step 18, come together as a class.

(4) Using each group's data (or perhaps the data for a few groups if there are many), one data set at a time, go through the line-up stage as a class. Each time, ask all students in the class to think about which graph they believe corresponds to the actual data and why. Discuss as a class.

REFERENCES

Carver, R. H., Everson, M., Gabrosek, J., Horton, N. J., Lock, R. H., Mocko, M., Rossman, A., Roswell, G. H., Velleman, P., Witmer, J. A., and Wood, B. (2016), Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016. Available at https://commons.erau.edu/publication/1083.