Jared Yu

Computational Statistics

Problem Set 13

1. For univariate data quick insights can be obtained by a “4-plot” that consists of the following four plots

* Plot of versus to see if there is any trend in the way the data are ordered;
* Plot of versus to see if there are systematic lags;
* Histogram;
* A q-q plot of the data versus the normal distribution.
  1. Posted on the course Blackboard site is the data Univariate.txt. Perform a “4-plot” on the data and discuss the information gained from these preliminary plots.

The first plot of plotting against can be seen below in Figure 1 on the top left. The pattern is such that it appears quite scattered. There seems to be some heaviness towards the low end and fewer observations near the upper part. However, there’s no periodic behavior that indicates some time-dependency. The second plot of against can be seen on the top right of Figure 1. It looks densest on the bottom left corner with fewer observations in the top right area. It doesn’t seem to show any obvious patterns where it would be clear that there’s some first order differences. The third plot is the histogram and is seen in the bottom left. The data clearly has a strong skewness to it. This indicates that there’s a non-normal distribution. Looking at the Q-Q plot in the bottom right of Figure 1, there’s a strong curve rather than a straight line. It’s somewhat difficult to draw a straight line also since it could be drawn from different angles. This makes it difficult to determine whether there are heavy or light tails on the left and right edges.

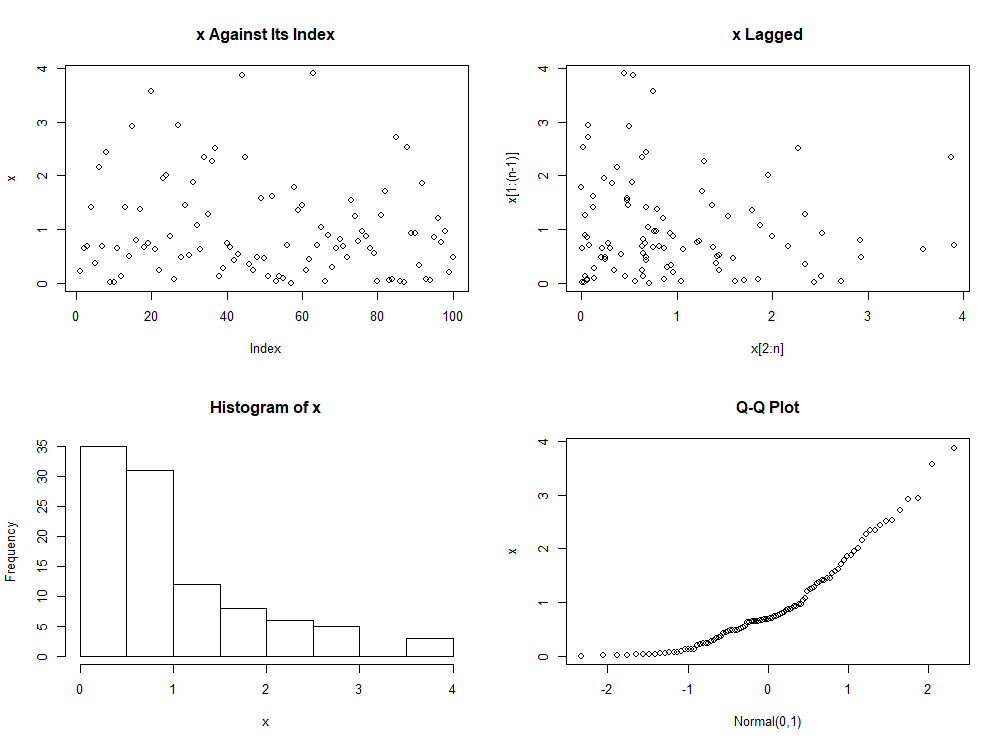


Figure 1 4-plot of the Univariate.txt data.

* 1. Give the Broken-Line ECDF and Mountain plot of the data. What do these plots tell you?

Below in Figure 2 is a grid of plots where the left-hand side is the Broken-Line ECDF and the right-hand side is the Mountain plot of the Univariate.txt data. The Broken-Line ECDF is clearly not a straight line, so it’s not from a Uniform distribution. The curve of the Broken-Line ECDF is concave, indicating that it’s a unimodal dataset. Also, there are more points to the left and fewer to the right, indicating some skewness. The Mountain plot also reinforces this notion that there’s a lack of symmetry in the data. The median is much closer to the smaller values of the data and so it lacks the isosceles triangle appearance of a Uniform dataset. It can be said then that the dataset has a unimodal and skewed distribution, like the histogram in Figure 1.

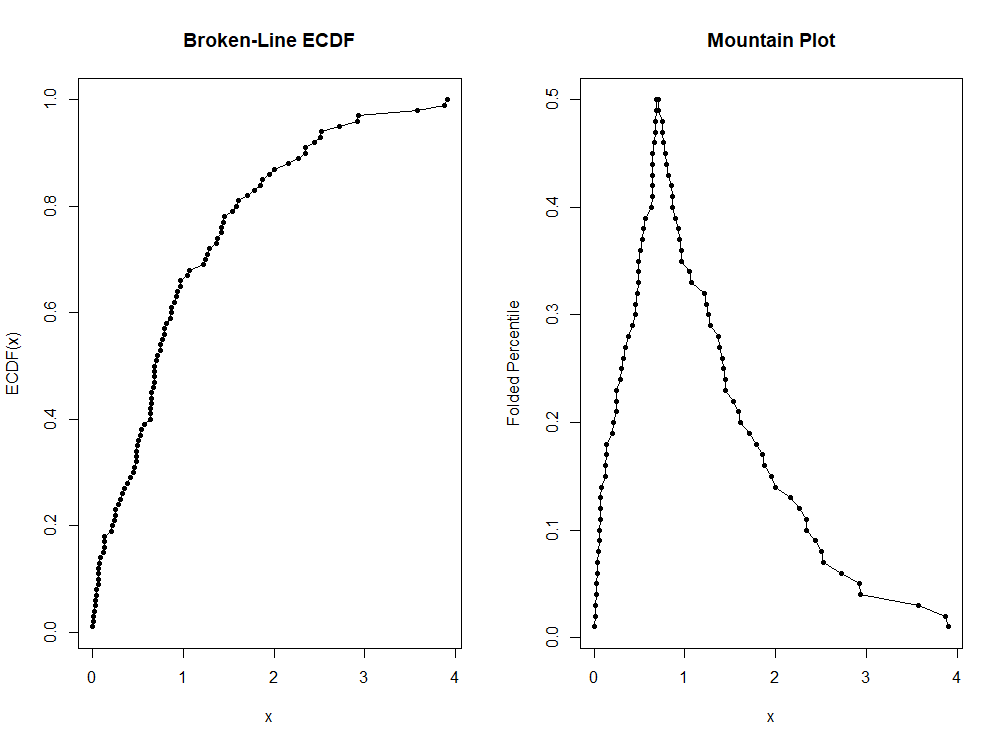


Figure 2 Broken-Line ECDF on the left-hand side and a Mountain plot on the right-hand side of the Univariate.txt.

1. Generate a sample of size 200 of pseudorandom numbers from a mixture of two univariate normal distributions. Let the population consist of 80% from a distribution and 20% from a distribution. Plot the density of this mixture. Notice that it is bimodal. Now plot a histogram of the data using 9 bins. Is it bimodal? Choose a few different numbers of bins and plot histograms.

The grid of plots can be seen below in Figure 3. The top left shows the density plot of the generated data. The top right shows the histogram using 9 bins. The bottom two histograms show histograms of 28 and 37 bins which were chosen after trying different values from 15 to 50 bins. The histogram with 9 bins looks somewhat unimodal, but also somewhat bimodal if looking closely at the small bump around 3. By increasing the number of bins, this smaller bump becomes much more distinct, matching what was seen in the density plot.

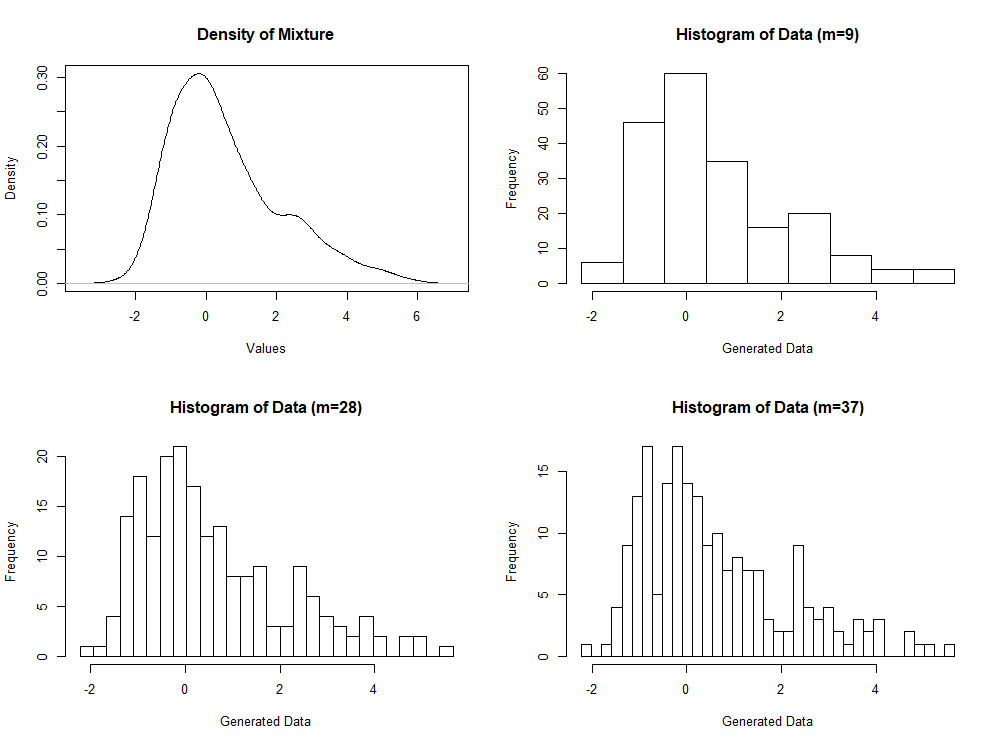


Figure 3 The top left shows the plot of the density of the mixture data. The top right shows the recommended histogram of 9 bins. The bottom two histograms are the selected histograms after plotting different bin numbers for a histogram.

1. Generate a sample of pseudorandom numbers from a distribution and produce a quantile plot of the sample against a distribution, similar to Q-Q plot presented in Lecture 13B. Do the tails of the sample seem light? (How can you tell?) If they do not, generate another sample and plot it. Does the erratic tail behavior indicate problems with the random number generator? Why might you expect often (more than 50% of the time) to see samples with light tails?

The plot below in Figure 4 shows a Q-Q plot where the data is generated from a random distribution using a seed of 4 in RStudio. In this plot, the small values are above the line and the large values are below the line. The line was drawn using linear regression in R with the lm() function (after dropping the last value due to Inf). This indicates that the distribution that the data came from has a lighter left tail and a lighter right tail than the reference distribution. This is interesting since the data itself was generated from the same reference distribution.

A possible reason is that the standard normal distribution has infinite bounds from the left and right side. It’s not possible for a computer to realistically generate exactly from this distribution since the computer isn’t precise to the point where it could evenly distribute from such a range. That’s sort of analogous to generating a purely random number from the real number line which is an impossible task for a computer and a philosophical debate as to what the meaning of random is. Therefore, it seems that the random number generator is doing fine sampling from within a specific range (e.g., ). However, outside of this range, it seems that it’s having difficulty getting a balanced distribution that is proportional to the theoretical distribution.

To be truthful, after plotting several sets of the Q-Q plots using different seeds, it didn’t seem that half of them would reflect this pattern. Out of the combination of left being heavy or light and right being heavy or light, it wasn’t a dominant occurrence compared to the other possibilities. However, when looking for either the left or right tail being light compared to the reference distribution, this was a common occurrence.

Imagining why this would be the case, it’s possible to think about what the theoretical distribution implies in an applied setting. It’s saying that as , the density of the samples would have a density which appears like the bell-shaped curve of the standard normal distribution. Generating infinite samples is an impossible task that would require infinite time. Furthermore, these values would need to be distributed proportionally across the real number line which is infinite in length. It seems then for the standard normal distribution, outside of the range of , that realistically representing these infinite tails in a simulation is not a practical task. Therefore, when comparing the results of the random number generator, it’s inevitable that the machine would often fail to represent this distribution fairly.

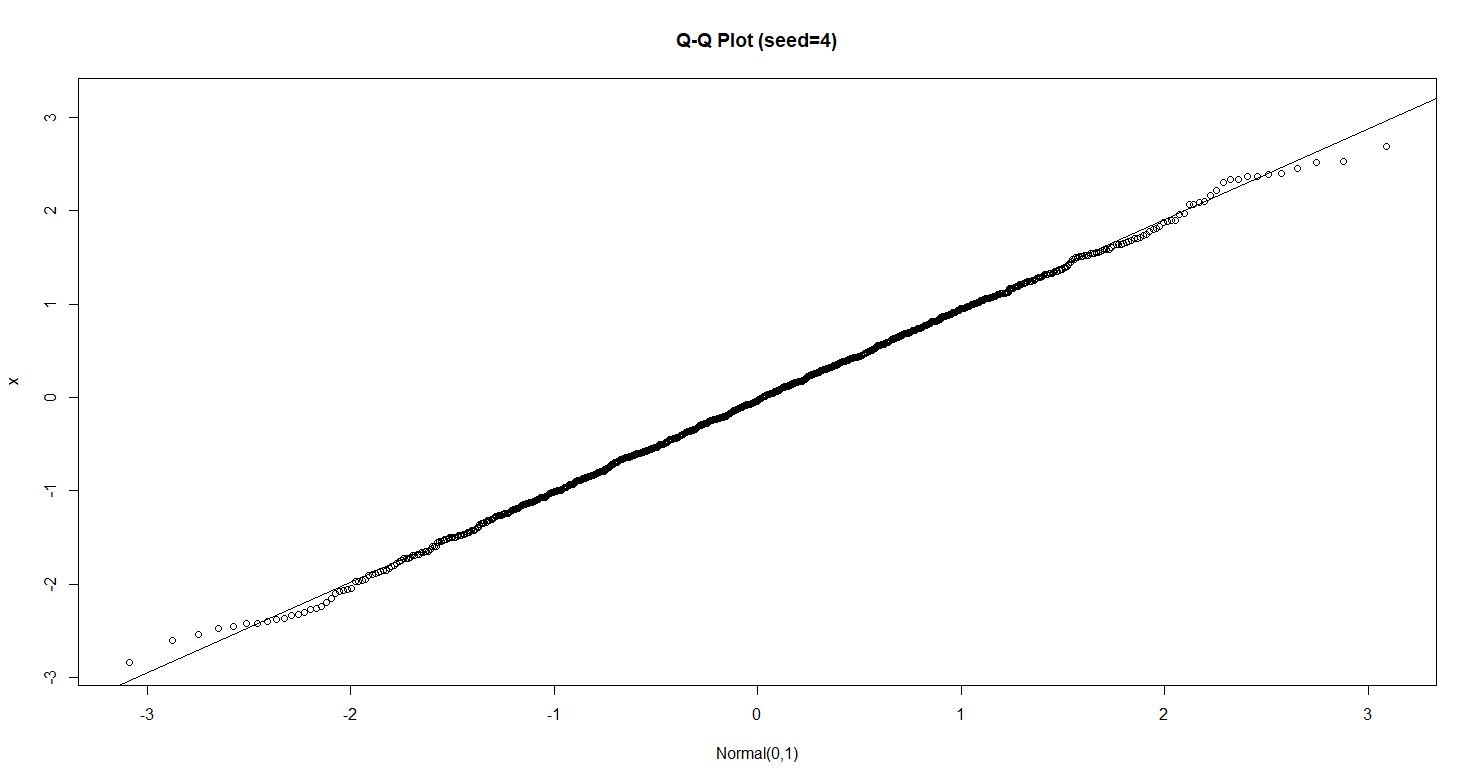


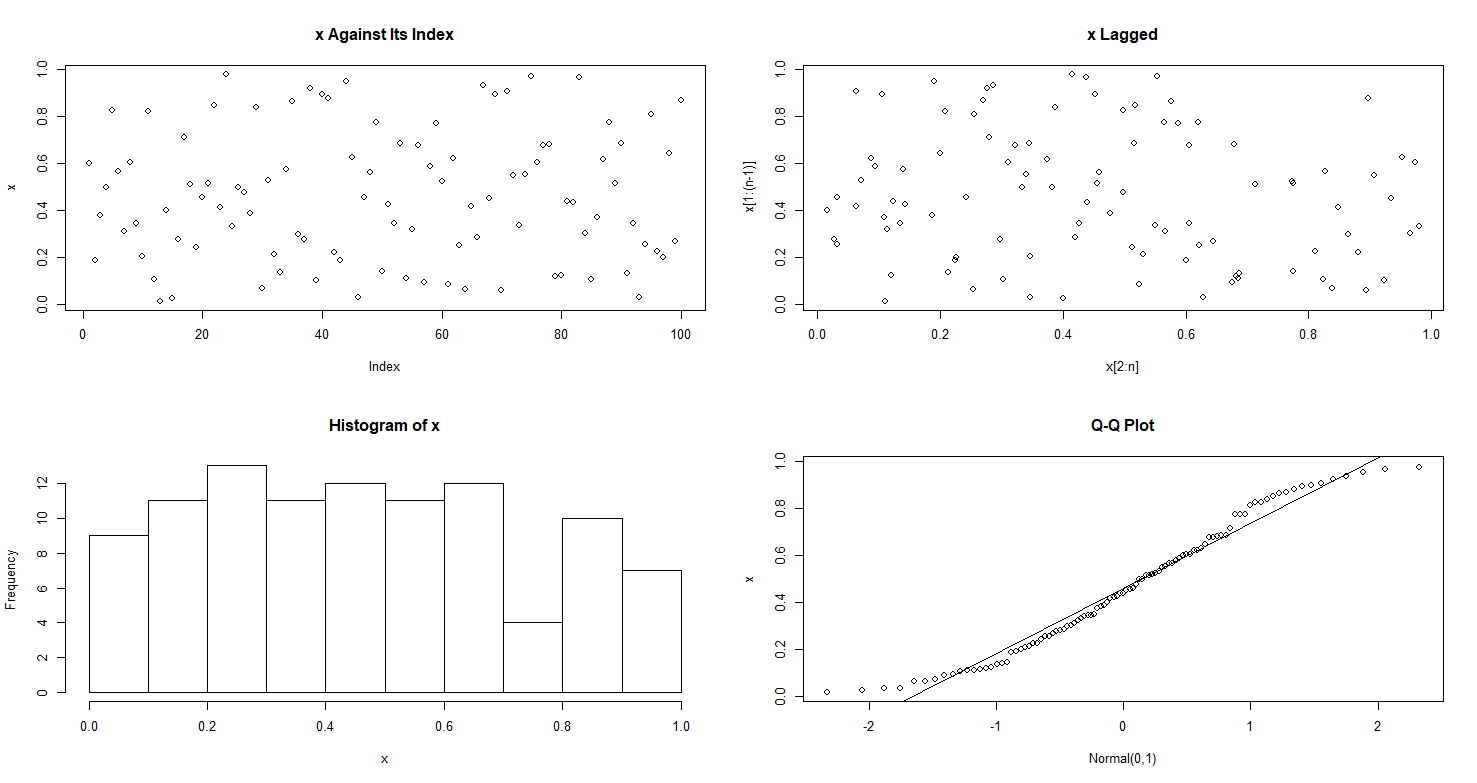
Figure 4 Q-Q plot of a sample of 1,000 points from a standard normal distribution using a seed of 4 in RStudio. It’s being compared to the reference distribution of a standard normal distribution. The tails indicate that the sampled data has lighter tails than the reference distribution. A line is drawn through using lm() in RStudio.

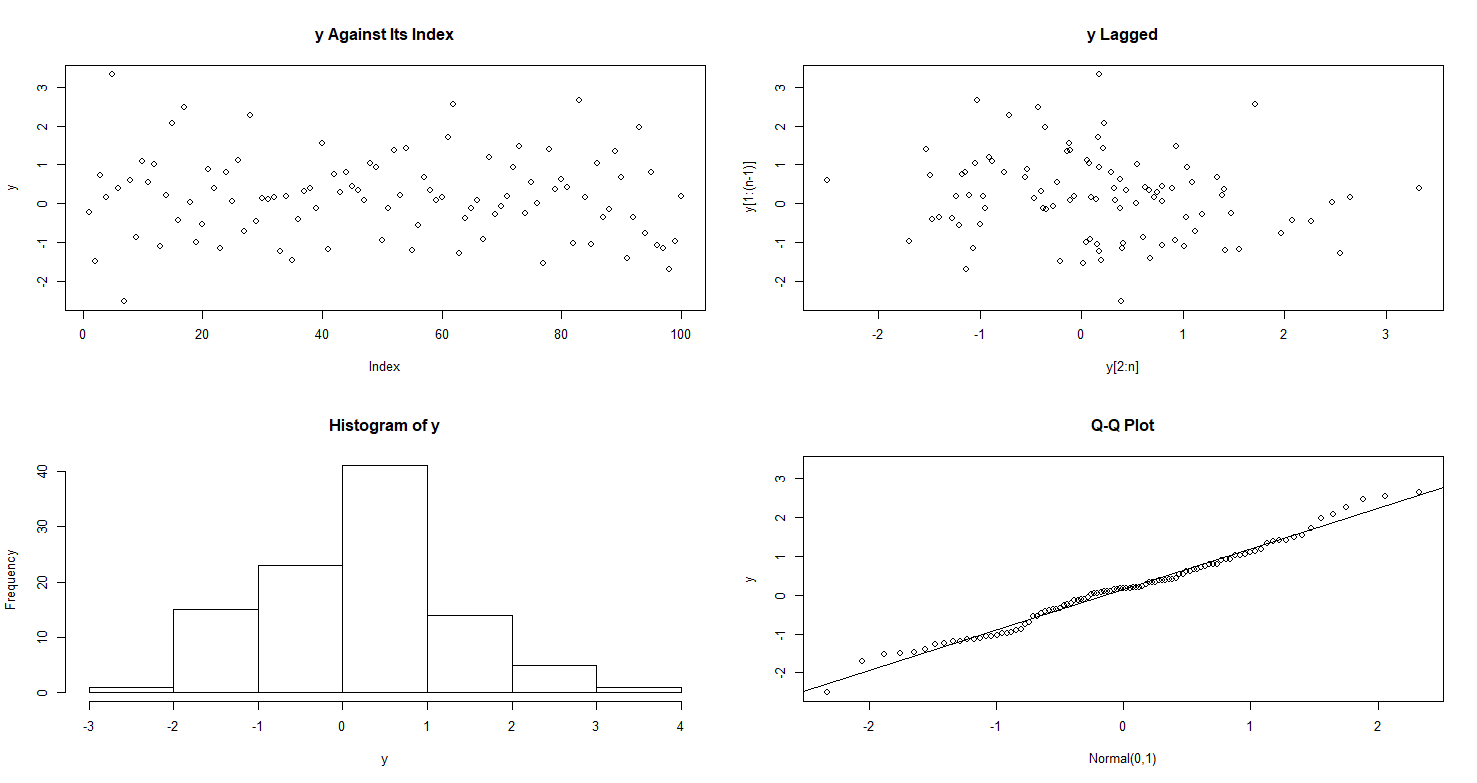
1. Posted on the course Blackboard site is the data MultiforGraph.txt. Use any of the techniques discussed in class to obtain information about this data set. What can you say about each of the variables individually? Do you see any dependencies among the variables? You may want to try techniques such as sorting the data or looking at subsets?

The first set of visualizations is the 4-plot for each of the variables: , , and (See Figure 5). In the 4-plot for , it seems that the top left plot of against its index is scattered. This indicates a lack of any time dependency. The top right plot also looks scattered, showing that there’s a lack of meaning for the first order difference between observations. The histogram on the bottom left looks almost uniform. Except there’s a dip around which makes it possibly something else. This may imply a bimodal distribution rather than the unimodal shape of the uniform distribution. Below in Figure 6 on the top left it shows the Q-Q plot when the theoretical distribution is the . In this plot it seems much more linear and so it makes sense to think that the original distribution of the data is something like a .

In the 4-plot for the variable, it seems that there’s a relatively scattered pattern in the top left plot of against its index. There are no periodic patterns to indicate that the variable is time dependent. Looking at the top right plot for the lagged observations, there are also no noticeable patterns other than it being somewhat dense around the center. In that case there possibly are no significant first order difference patterns. The histogram shows a distribution that looks like it possibly came from the normal distribution, but it’s heavier to the left and quite sharp in the center. The Q-Q plot seems to indicate however that it’s quite close to a normal distribution. In Figure 6 the Q-Q plot is shown for the normal distribution with mean and variance set to the sample mean and sample variance of the data. The line is much straighter, and the fit seems decent. This makes it plausible to think that the data comes from a type of normal density.

The last 4-plot can be seen in Figure 5 for the variable. The first plot showing against its index seems to indicate that there’s possibly some time dependency. At the beginning of the data, the observations have a greater variance. However, during later observations it’s apparent that the variance of the values decreases, it looks like a cone-shape. The top right plot of the lagged observations is interesting. It looks a bit dense around the center, but this seems to indicate more of a possible pattern rather than there not being a pattern at all. This seems to mean that there’s possibly something related to the first order difference between observations. However, there’s nothing certain yet. The histogram indicates something like a normal distribution, but with a low variance. The Q-Q plot shows that there’s some straightness, but it’s quite imbalanced on the left side. The Q-Q plot is remade in Figure 6 using the normal distribution with the sample mean and variance used for the parameters. The fit doesn’t look much better than what was seen initially, so it’s possible that the data comes from some non-stochastic distribution since there was evidently some time dependency.





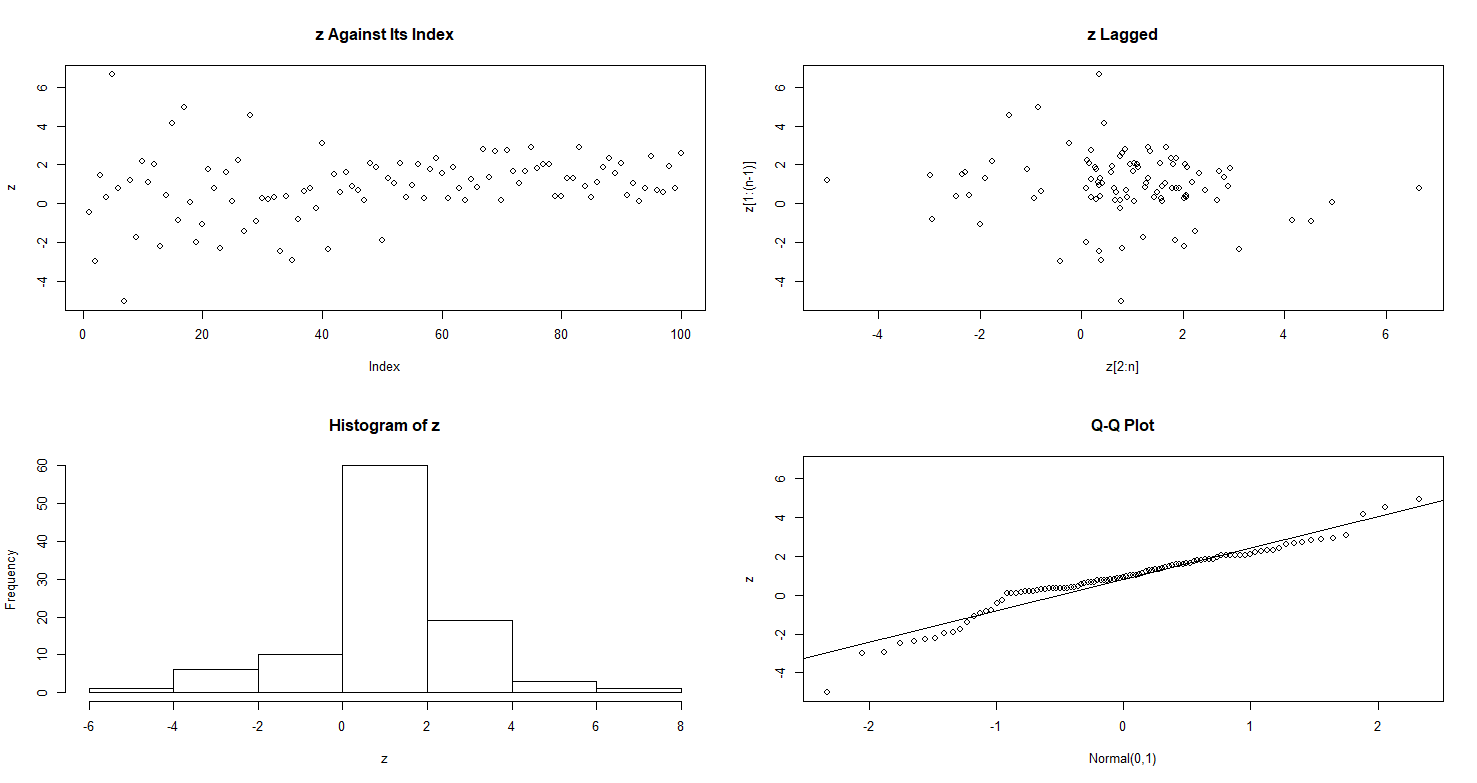


Figure 5 A series of 4-plots for the variables x, y, and z from the MultiforGraph.txt data.

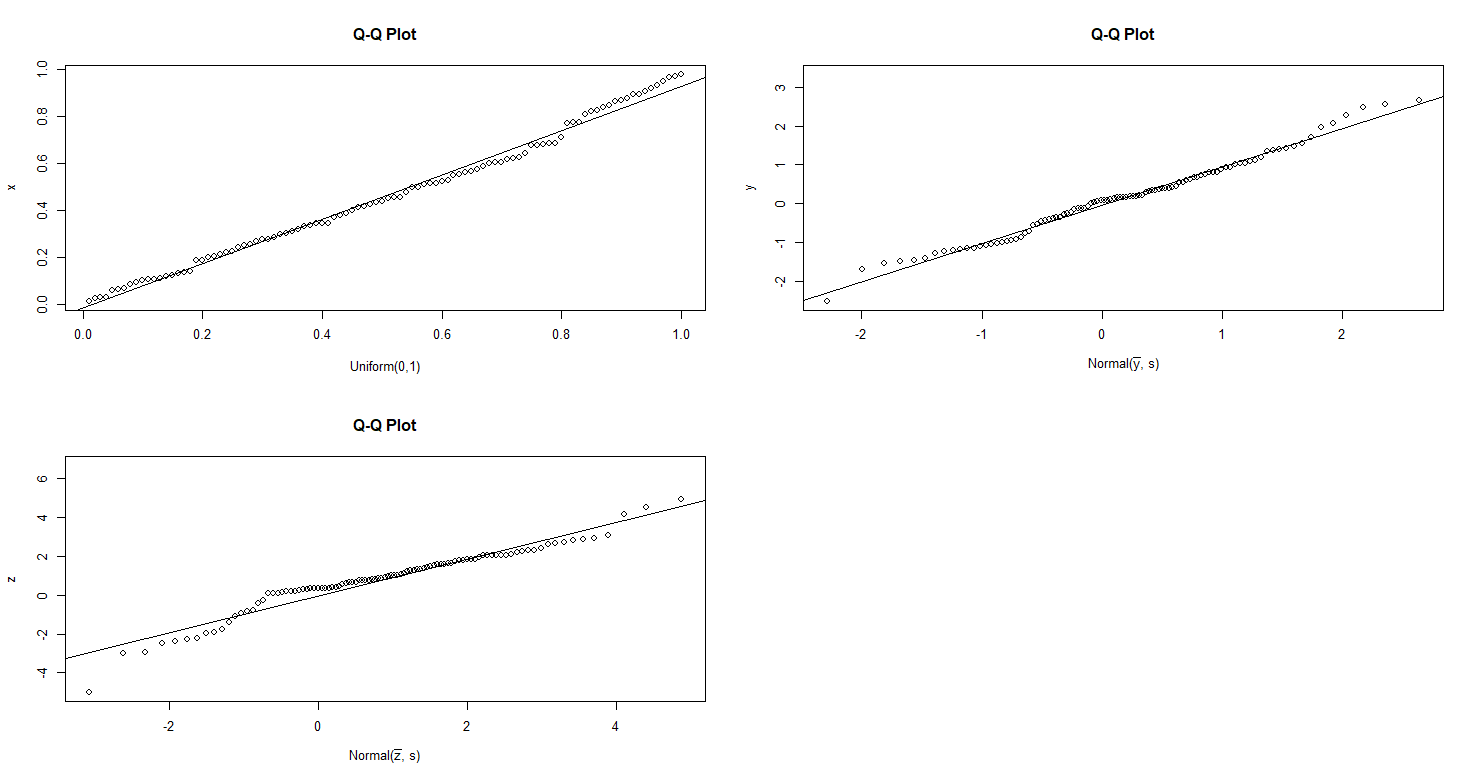


Figure 6 The above shows a set of Q-Q plots for variables x, y, and z using different reference distributions.

Below in Figure 7 it shows a scatter plot matrix for each of the pairs of variables. It seems to show that there’s a strong correlation between and the other two variables. For the scatter plots of and , there seems to be mostly a scatter of points that looks roughly random. However, for each of the pairs of scatter plots that include , it seems that there’s a strong linear relationship represented by a rather straight line along with some scatter around the line. In the scatter plot for and , it seems that the scatter is much more focused around the center of the line. When looking at the scatter for and , it seems that the randomness is around the entire line itself. This seems to indicate that is a function of both and . This makes sense based on the 4-plot, where seemed to have some time dependency which implies that the indices of the observations matter. This would be the case if each of the are functions of the and variables which themselves are random variables. In other words, the random combination of and observations would create each of the observations. If the order is scrambled, then the set of observations would have to change also.

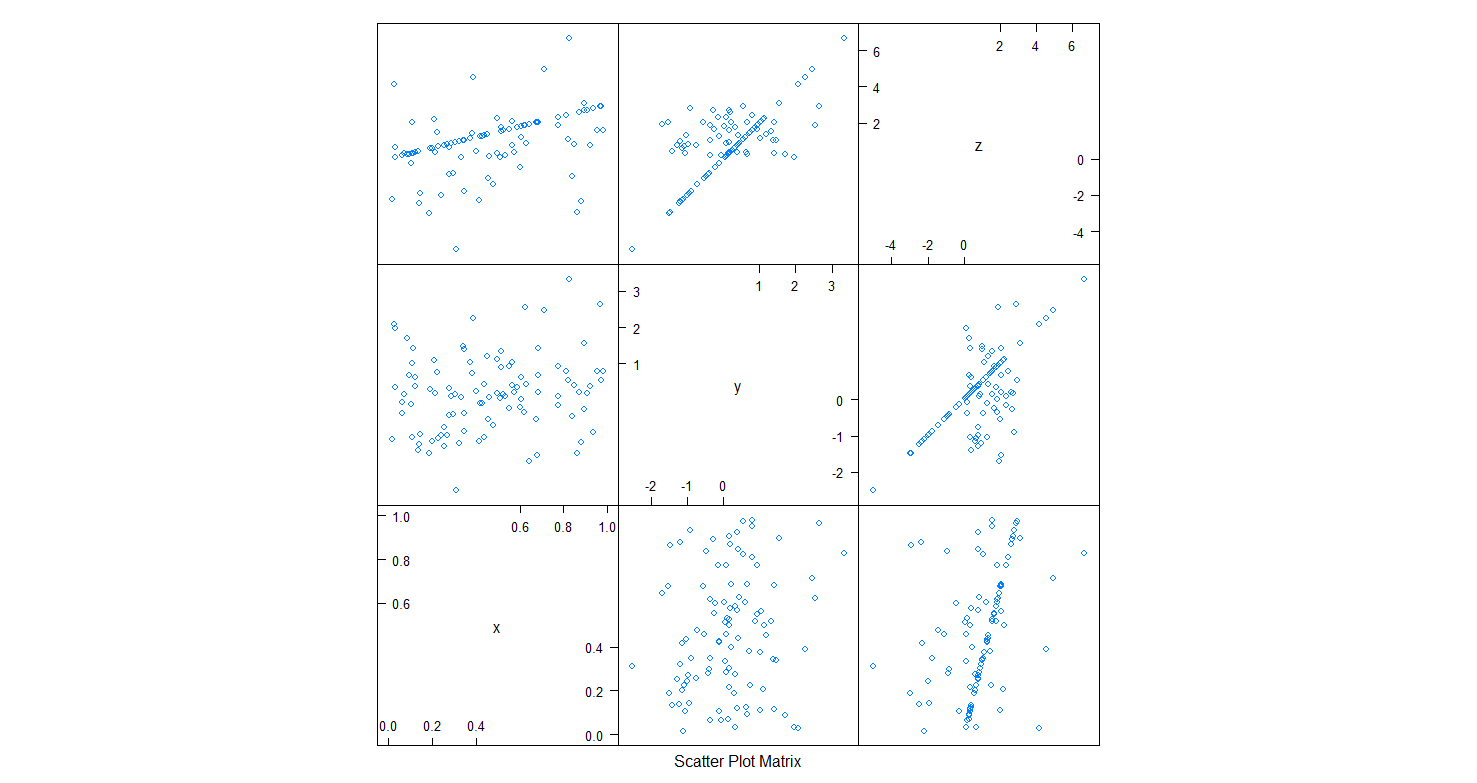


Figure 7 Scatter plot matrix of the three variables using splom() from {lattice}.

1. Posted on the course Blackboard site is the data Bears.txt a subset of a data set described in Reader’s Digest (April, 1979) and Sports Afield, (September, 1981). The data set consists of several measurements for bears that were captured, measured, and released.

The variables in the data set are:

1. Estimated age in months
2. Gender (1=male, 2=female)
3. Length of head in inches
4. Width of head in inches
5. Girth of the neck in inches
6. Body length in inches
7. Girth of the chest in inches
8. Weight in pounds
9. Name

Apply the following graphical techniques to the multivariate random variable composed of variables 3 to 7. For each technique, describe its usefulness in analyzing this data set. (Think about finding maximum values, minimum values, grouping the data, correlations, etc.)

* 1. Stars

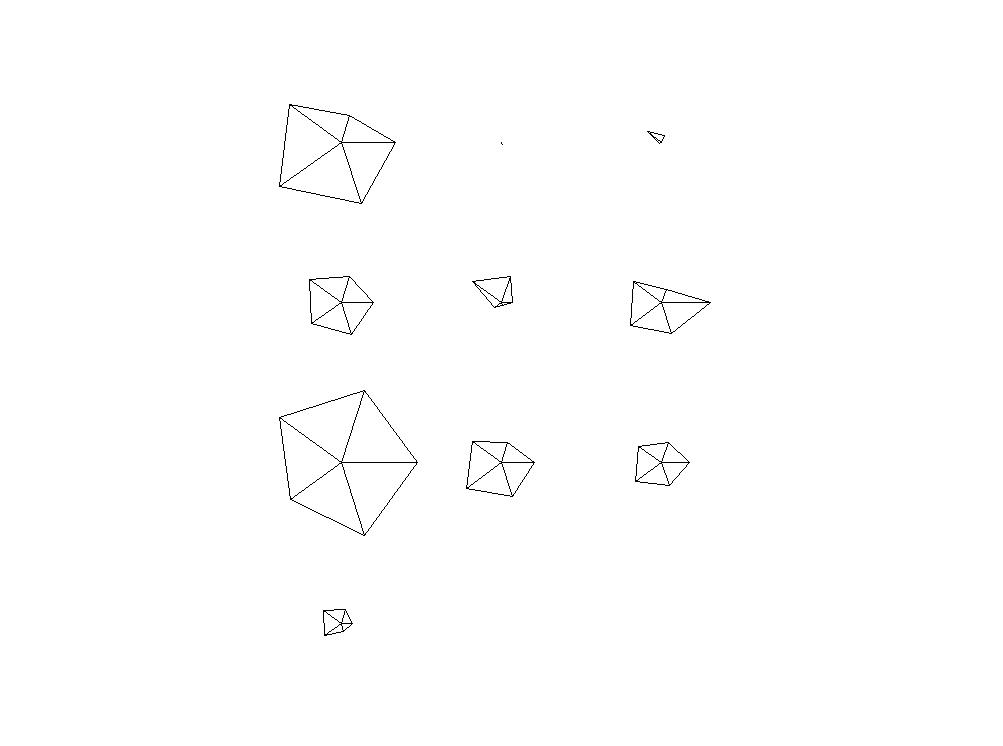


Figure 8

* 1. Chernoff faces

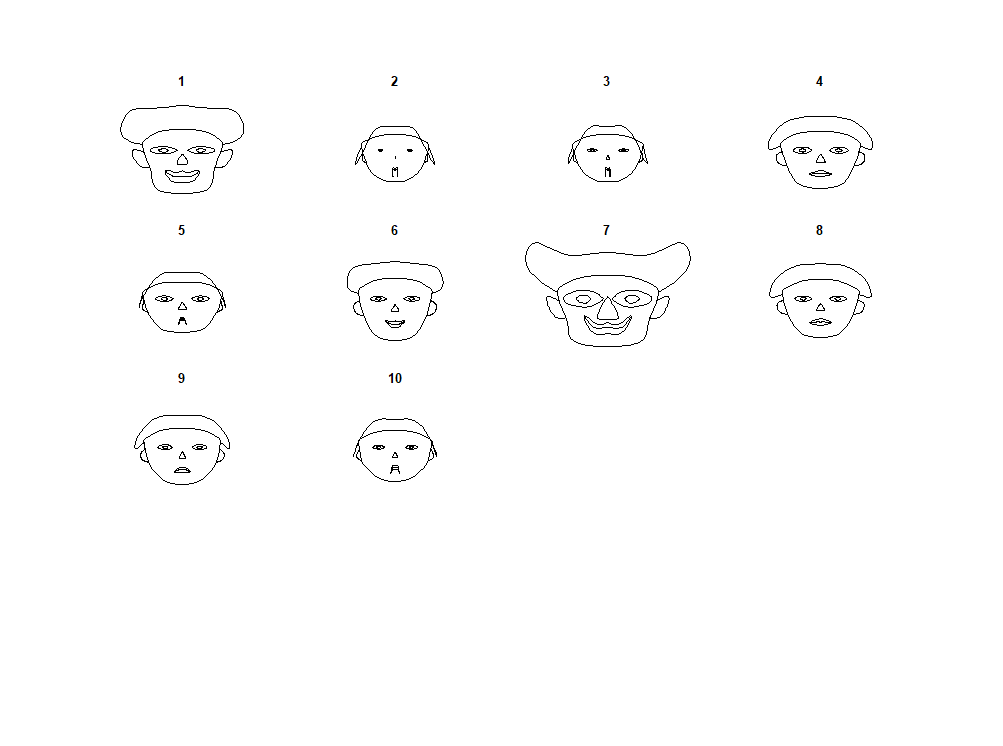


Figure 9

* 1. Parallel Coordinates

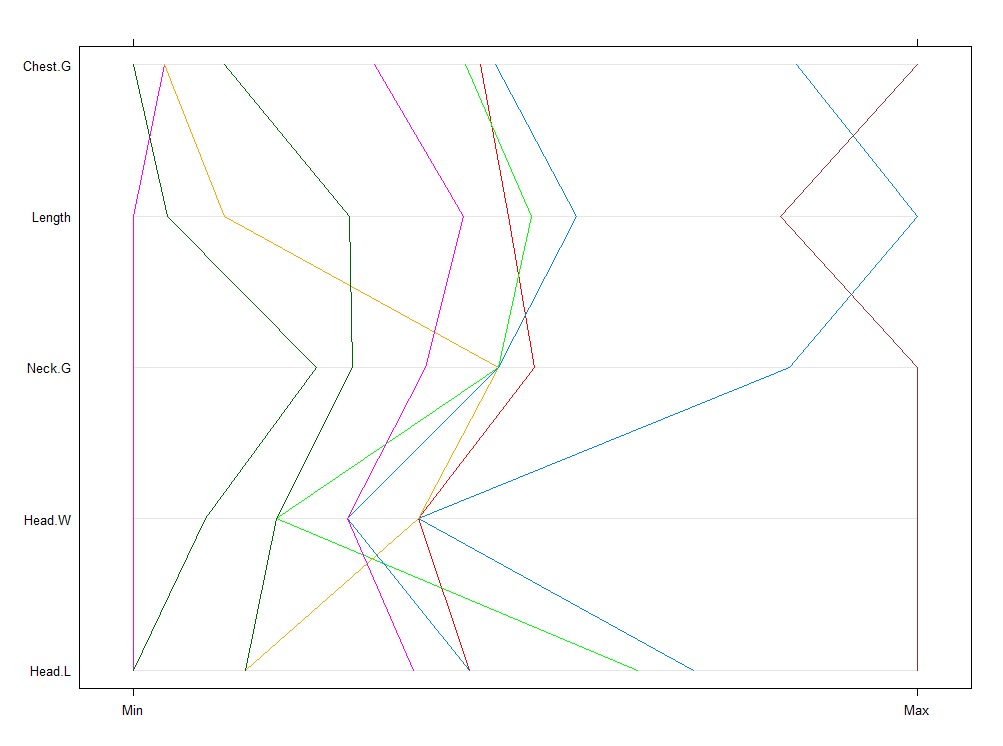


Figure 10