Author: Brenda Gunderson, Ph.D., 2015

Lab 5: Understanding Normal and Random Data

Objective: In this lab, you will use some additional graphical tools to summarize the distribution for a variable or response and check assumptions before performing a statistical test. Graphs you might need to examine include time plots for data collected over time and QQ plots for checking whether a normal (bell-curve) model is a reasonable distribution for a quantitative variable. These techniques can be very useful at the start of data analysis to get a feel for the data.

Application: Brad is the manager of the Detroit Tigers and to prepare for the next round of the playoffs, he would like to run a hypothesis test involving the mean number of runs his players have scored over the last month of play. Brad knows that one assumption required for performing this analysis about the mean is that his data must be considered a random sample (the observations can be viewed as coming from the same parent population). He can examine this assumption by collecting the number of runs scored over the last month and creating a time plot.

Overview: Data on a quantitative variable should be examined graphically. If the data has been collected over time, the first graph to examine is a **time plot**. If the resulting time plot appears to be stable, or if the data was not collected over time, then graphs that can be used to summarize the distribution for a single quantitative variable or response are a **histogram,** a **boxplot**, and perhaps a **QQ plot**. Each graph provides different information about the distribution. The overall shape of the distribution and existence of outliers can generally be used to assess if the data appear to be coming from a relatively homogenous population. If so, then various **numerical summaries** may be used to characterize the center of the distribution and the spread of the distribution.

Note that some graphical tools are introduced solely in lab, not in lecture, so it will benefit you to read this overview thoroughly

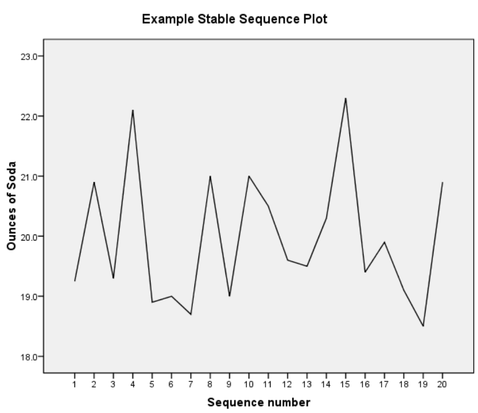
Sequence (Time) Plots: Data might be gathered over time. Employment rate, stock prices, and sales figures are just a few examples. When data is gathered over time, such as the number of runs scored over a one month by the Tigers, it is generally wise to examine the data plotted against time. Plots against time can reveal the main features of a time series, overall patterns and striking deviation from those patterns. Some overall patterns that may arise are:

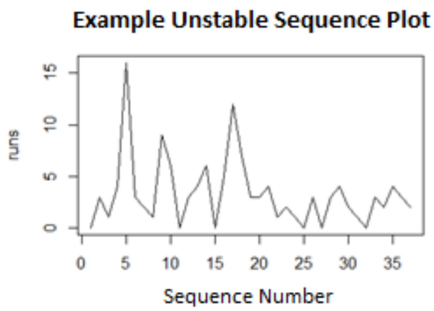
- A persistent, long-term rise or fall called a **trend** (either increasing or decreasing).
- A pattern that repeats itself at regular intervals of time called **seasonal variation**.
- A persistent, long-term increase or decrease in the **variation** of the observations called a **pattern in variation**.

If data is collected over time, a sequence plot can be used to check the assumption of a random sample, often needed for inference procedures. A random sample consists of *independent* and *identically distributed* (i.i.d.) observations. This means the observations can be considered as all coming from the same parent population (with the same or *identical* distribution) and are *independent* of one other.

With a sequence plot, you can check the *identically distributed* aspect of a random sample by looking for evidence of **stability** in the plot. Stability is supported when both the mean of the observations and the amount of variation among observations appear to be constant over time and there does not appear to be any *pattern* in the resulting plot.

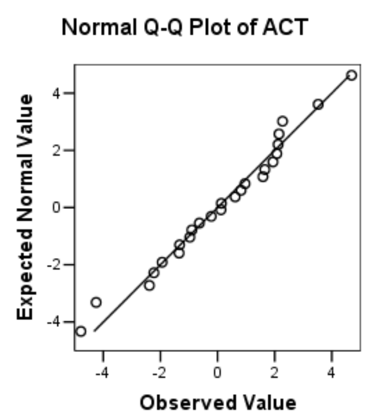
Below are two sequence plots; in the first plot the observations appear to support that the underlying process that generated the observations is stable, but that is not the case for the observations in the second plot on the right. In this case, there appears to be an increasing trend, thus the underlying process does not appear to be stable; the observations should not be considered a random sample.



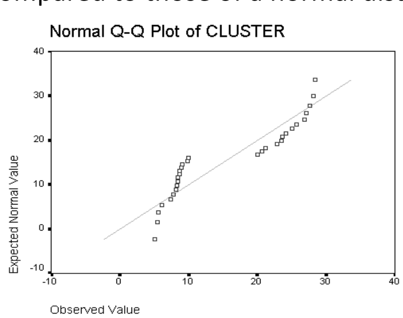


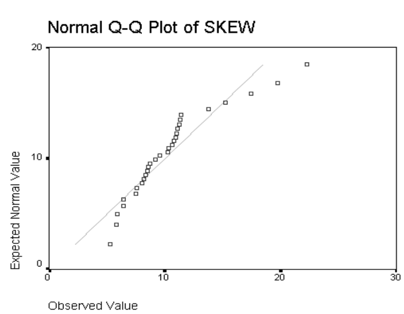
Going back to the earlier application, here is the time plot Brad created of the number of runs the Detroit Tigers scored between September 1, 2013 and October 19, 2013. This is another example of an unstable time plot. Brad notes that there is a large variation in the number of runs scored by the Tigers early in September but less variation in their run total around game 21. The underlying process for the number of runs scored does not appear to stable, so Brad should not consider this a random sample.

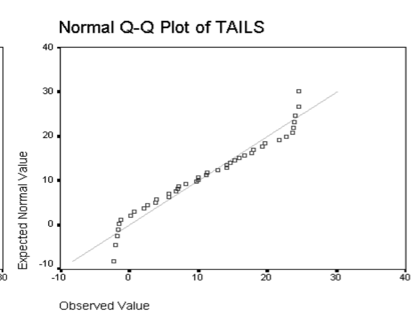
QQ Plots: Sometimes the **assumption of a normal model for a population of responses will be needed in order to perform certain inference procedures**. A histogram can be used to get an idea of the shape of a distribution. However, there are more sensitive tools for checking whether the shape is *close* to a normal (bell-curve) model. The *best* plot that can be used to check for normality is called a QQ Plot, which plots the percentiles (quantiles) of a standard normal distribution against those of the observed data. If the observations follow an approximately normal distribution, the resulting points should follow roughly a straight line with a positive slope. Strong deviations would indicate possible departures from a normal distribution. At the right is an example of a QQ Plot showing data that does seem to come from a population with an approximately normal distribution.



The three graphs below are examples for which a normal model for the response does not seem reasonable. The QQ plot on the far left indicates the existence of two clusters of observations. The QQ plot in the center shows an example where the shape of the distribution appears to be skewed right. The QQ plot on the far right shows evidence of an underlying distribution that has shorter tails compared to those of a normal distribution.

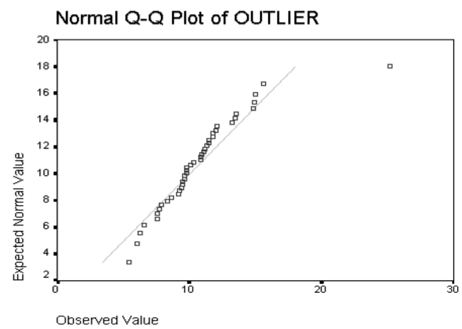






Note: Many inference procedures, including some you will use later in the semester, require the assumption of normally distributed population(s). Most of these procedures are robust which means we would need to see strong evidence of a departure from normality to conclude this assumption is not met. Some mild departures from normality would still allow us to conclude the underlying model for the response is reasonably normally distributed.

Finally, consider the QQ plot at the right. In this case, we would say the QQ plot shows evidence of an underlying distribution which is approximately normal except for one large outlier that should be further investigated. Outliers could appear in either the upper or lower tail.



Warm-Up: Matching

Match the graph or descriptive statistic to one of its primary uses.

____ 1. Histogram	A. Measure of center, not sensitive to outliers
____ 2. Bar Chart	B. Compare distributions (but not their shapes)
____ 3. Mean	C. Examine distribution of a categorical variable
____ 4. Median	D. Helps assess if can treat data as a random sample
____ 5. Side-by-side Boxplots	E. Measure of spread
____ 6. IQR	F. Examine distribution of a quantitative variable
____ 7. Time Plot	G. Helps assess if underlying distribution is bell-shaped.
____ 8. QQ Plot	H. Measure of center, sensitive to outliers

ILP: More Visualizing and Exploring Quantitative Data

In this In-Lab Project, you will create additional useful graphs and obtain more descriptive statistics for quantitative data using R.

Task 1: **QQ Plots.**

1. Recall the data set **employee.Rdata** contains information on employees at a company, and that a histogram of the current salary data indicated a strongly skewed to the right distribution. Let's examine a **QQ plot** (otherwise known as a Quantile-comparison plot) to see that it supports this non-normal feature. To create a QQ plot, go to **Graphs > Quantile-comparison plot** and select the **SALARY** variable.

   Provide a sketch and explain how the QQ plot supports that the underlying distribution of salary does not appear to be normal (or bell-shaped)

2. We can also consider making QQ plots for different populations. Recall that the distribution of salary for minorities was far less skewed than the whole distribution. Let's try making a QQ plot of salary for minorities. To do this, we will need to make a new dataset that only contains minorities in it. We do this by going to **Data > Active data set > Subset active data set**. The subset expression should be formatted like VARIABLE=="value" – here, the variable we are splitting by is MINORITY, and we want all observations with the value "Yes" for MINORITY; so type in **MINORITY=='Yes'**. Give the data set a new name and click OK – if done correctly, the new data set will appear in blue text at the top.

   Now, make a QQ plot of the **SALARY** variable as we did before with the new data set. How does this distribution of minorities compare to the entire distribution from the previous question? Would you consider this a normal distribution?

3. Let's try analyzing the normality of IQ scores given in the dataset **iq.Rdata,** which contains information from high school students collected for the purpose of examining the relationship between IQ scores and GPA. Before creating a QQ plot, we can get an idea of the distribution of IQ scores by simply making a histogram. Make a sketch of this histogram below. Do you think the population of IQ scores might follow a normal distribution?

4. Let's see if our suspicions from the previous problem are correct, and make a QQ plot of the IQ scores. You will need code similar to what we used in problem 1 – recall that variable names can be viewed by clicking on **View data set**.

   Draw the resulting QQ plot below. Is a normal distribution a reasonable model for IQ scores in the population based on this QQ plot?

Task 2: **Time Plots.**

1. The data set **oldfaithful.Rdata** contains the date and duration of eruptions (in minutes) of the Old Faithful geyser. The data was collected several times per day over 23 consecutive days. To make a time plot, we will have to use some custom code, as there is no R Commander function for a time plot. To make a time plot of the **DURATION** variable from the **oldfaithful** data set, type the following code into the **R Script** box at the top of your R Commander window (note the type is the letter L but must be lowercase):

   ```
   plot(oldfaithful$DURATION, type="l", main="Time Plot of variable
   by name")
   ```

   Highlight this code, and then click the **Submit** button. Make a sketch of this time plot.

2. Does the plot appear to show any trends or changing variation? Would you consider it to be stable?

3. Would it be reasonable to conclude these data are a random sample of eruptions?

4. Would it make sense to go on to make a histogram of these data? A QQ plot of these data? Descriptive summaries? Why or why not?

5. Let's examine another time plot – download the data set **chemical.Rdata**, which contains the concentrations of a certain chemical in 15 consecutively produced batches of solution. Create a time plot of the **concentration** variable and make a quick sketch here, using code similar to problem 1.

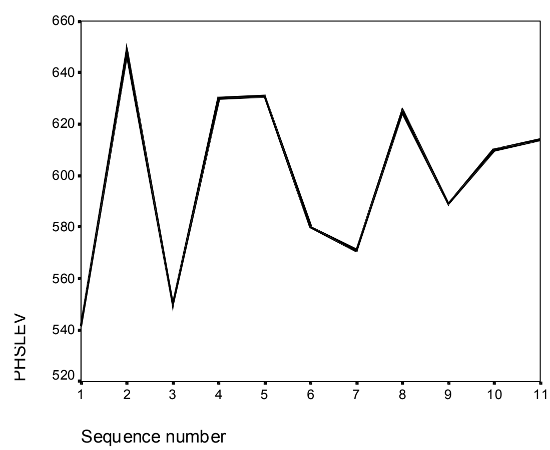
6. Does the plot appear to show any trends or changing variation? Would you consider it to be stable?

7. Would it be reasonable to conclude these data are a random sample of eruptions?

8. Would it make sense to go on to make a histogram of these data? A QQ plot of these data? Descriptive summaries? Why or why not?

Cool-Down: Check Your Understanding About Time Plots

A new method of measuring phosphorus levels in soil is under consideration. A sample of 11 soil specimens is analyzed using the new method. The time series (sequence plot) for the 11 observations is presented.



Comment on the overall stability of these data based on this plot.

Would it be appropriate to go on to make a histogram of these data? Explain.

Optional Review of Time Plot and QQ Plot Examples

If you would like to look at more examples of QQ plots (which help assess whether the model for the underlying population of responses seems to be normal) and examples of time plots (which help assess whether our underlying process appears to be stable and if we can consider the data to be a random sample) … try out these tasks using some simulations with R (R scripts).

Task 1: Go to the **Extra Review** link on your course site and click on the **QQ Plots in R** link in the list. Review the background about the simulator and download the **qqplot** script file (which will open up the R program).

1. Begin the program by entering the following command. **qqplot ()**
2. Select your sample size by entering a number between 1 and 10000.
3. Select if you want a QQ plot from a normal or non-normal distribution.
4. If a non-normal distribution, select the type of distribution you would like to see.
5. Once your QQ plots and the corresponding histograms have been created, you will be asked if you wish to save the plot. You will then be prompted to select a sample size. Create QQ plots for many different samples.
6. Sketch the QQ plot and Histogram for a sample from a skewed right distribution.
   QQ plot: Histogram:

Task 2: Go to the **Extra Review** link on your course site and click on the **Time Plots in R** link in the list. Review the background about the simulator and download the **timeseries** script file (which will open up the R program).

1. Begin the program by entering the following command: **timeseries ()**
2. Select your sample size by entering a number between 1 and 10000.
3. Select if you would like an example of a stable or unstable time plot.
4. If unstable time plots are selected, select the type of pattern you want to see.
5. Once your time plots have been created, you will be asked if you want to save the plot. You will then be prompted to select a sample size. Try out the various options and explore the various patterns of time plots.
6. Create a time plot for a sample of 1000 data points taken from a process with an increasing mean and decreasing variance. Sketch the time plot below.