

Basic R and Exploratory Analysis

In the book *Introductory R*, read through the chapters entitled “What is R?”, “A first R session”, “Basics” and “Using the help files”. If you aren’t familiar with command line interfaces or the use of vectors and matrices, work through the exercises in the ‘Basics’ chapter. If you’re happy that you’re on top of this, try the following more advanced exercises.

The effect of sample size on frequency distributions

Use the function `rnorm()` to do the following (type `?rnorm()` to get the help file):

1. Create an object called “Norm1” with 25 normally distributed random numbers with mean 2 and standard deviation 2.
2. Create an object called “Norm2” with 50 normally distributed random numbers with mean 2 and standard deviation 2.
3. Create an object called “Norm3” with 100 normally distributed random numbers with mean 2 and standard deviation 2.
4. Create an object called “Norm4” with 500 normally distributed random numbers with mean 2 and standard deviation 2.
5. Set the parameter “mfrow” so that R will draw a 2x2 grid of graphs in its graphics window. You can do this by inputting the following code:

```
par(mfrow=c(2,2))
```

6. Use the `hist()` function to draw frequency histograms of Norm1, Norm2, Norm3 and Norm4 in the same window.
7. What do you see as the sample size increases?

The effect of increasing the mean on the shape and variance of Poisson distributed data

Use the function `rpois()` to do the following:

1. Create an object called “Pois1” with 200 Poisson distributed numbers with a mean of 2
2. Create an object called “Pois2” with 200 Poisson distributed numbers with a mean of 10
3. Create an object called “Pois3” with 200 Poisson distributed numbers with a mean of 50
4. Set the parameter “mfrow” so that R will draw a 3x1 grid of graphs in its graphics window. You can do this by inputting the following code:

```
par(mfrow=c(3,1))
```

5. Use the `hist()` function to draw frequency histograms of Pois1, Pois2 and Pois3 in the same window.
6. How does the shape of the Poisson distribution change as the mean increases?
7. Calculate the variance for Pois1, Pois2 and Pois3. How does the variance change as the mean increases?
8. Try to set the graphics window so that it will only show one plot at a time again.

Exploratory Analysis Problems

1. Download the spreadsheet entitled “Francis Black 1966 Measles Data.xlsx” from the module QM+ page. This is a dataset from a famous study which demonstrated the very strong link between population size and infectious disease epidemiology. For 19 islands from the Atlantic and Pacific Oceans we have data for the population size in 1956 and the percentage of months during the study period when measles was reported as being present on the islands.
2. Check through the spreadsheet and save it as a text file that you can read into R
3. Read the text file into R
4. To check whether the data have been read in properly and are in the format we want, you can use the `head()` function to give you the top 6 lines of the data frame, or you can use the `str()` function to give you details about each variable in the data frame. In particular, you will want to check whether the variables that should be factors have been recognised as such.

5. If all is well, we can start with some preliminary looks at our data. Using the `hist()` function, have a look at the frequency distribution of the population sizes and, annual susceptibles input and the percent of months with measles reported. If you don't like the number of bins used in a particular histogram, you can change it using the `breaks=` argument.
6. What do you notice about these variables? Are there any issues with these variables that might affect your analysis, and if so how might you deal with them?
7. It's possible that the region might be an important predictor of the patterns in our data. Using the `plot()` function, generate some boxplots showing how population size, susceptibles input and the percent of months with measles reported are related to region.
8. Do the same thing but with the order of the variables reversed in the `plot()` function to get the raw data plotted as a strip plot. This is often a useful way of visualising data with relatively small sample sizes
9. Draw scatter plots of the percent of months with measles reported against susceptibles input and population size (with any adjustments that you might have made earlier). What do you conclude?
10. You can change your plot parameters to change the look of the plot in the arguments for the `plot()` function. There's a good brief guide here <http://www.statmethods.net/advgraphs/parameters.html>. Try changing the plot symbol (`pch=16` for a filled circle for example), the colour (`col="red"` or `col="steelblue"`: see <http://bc.bojanorama.pl/wp-content/uploads/2013/04/rcolorsheet-0.png> for the full set). `xlab="Some text"` will set the x-axis label and `ylab="Some more text"` the label for the y-axis, and `main="Title"` will give our graph an overall title.
11. You can use the `points()` function to overwrite the points for a particular region with a different plot symbol or a different colour. You need to specify the points using a subscript (e.g. `[Region=="Pacific"]`) for both the variables used in the plot function.