MA26620: Practical 4

Investigating Normality, Housing Data and Assignment One

1 Investigating Normality

We've seen in the lectures that the Normal distribution is incredibly important and frequently occurring, in part due to the Central Limit Theorem. Often in our models we make the assumption that something is Normally distributed. But how do we know whether this is a sensible assumption or not?

Well, we've said in the lectures that the best way to do this is a quantile-quantile (or Q-Q) plot, which is more sophisticated than looking at a histogram. At the start of this practical, we're going to look at generating random samples in R from different distributions, and then we're going to see what the resulting histograms and Q-Q plots look like.

1.1 Generating some random data

There are many commands in R for generating random samples (strictly speaking pseudo-random) from different distributions. The following produces a random sample of size 100 from Normal, $N(\mu, \sigma^2)$, draws a histogram and adds a Normal density curve. Try, for example,

```
x <- rnorm(100,3,2)
```

This creates a sample of 100 pseudo-random numbers drawn from a $N(3, 2^2)$ distribution. Note that `rnorm` takes three inputs: the sample size, the mean μ and the **standard deviation** σ, NOT σ^2. You can type `x` to see all the numbers you've generated and `summary(x)` will show you the mean (which is hopefully not a million miles away from 3) and the various quartiles.

It's hard to see how our generated data is distributed from this list of numbers. Let's make a quick histogram. Last week we saw how to make various nice looking histograms in ggplot. For now though, we'll use the standard non-fancy built in R plotting tools.

```
hist(x, probability=TRUE, main="N(3,4)")
```

Does this look Normally distributed to you? Each of you will have generated a different sample so some people's data might look nice and symmetrical and not dissimilar to a bell-shaped Normal probability curve, with. Others' may look less so.

This is why a Q-Q plot is a more refined tool. See Blackboard (specifically Normal random samples > Is my data Normally distributed?) or your lecture notes for a reminder of what a Q-Q plot does.

We can make a simple Normal Q-Q plot by typing

```
qqnorm(x)
qqline(x)
```

The points will not stray far from the line, and any deviations don't have too strong a pattern to them. We'd naturally expect a bit more variation in smaller samples, so try

```
x1 <- rnorm(30);qqnorm(x1);qqline(x1)
```

which will run 3 commands from one line (that's what the semi-colons do) which generates 30 values from a $N(0, 1)$ distribution), then makes the qqplot.

You can easily repeat this a few times (up arrow, enter), generating a fresh sample `x1` each time, in order to get a feel for the variation you might expect. The straight line added by `qqline()` passes through the upper and lower quartiles (ie 25th and 75th quantiles) and it indicates the expected line if the distribution is truly Normal. Departures from the line show whether the data have more extreme values than expected (eg Exponential) or are more constrained (eg Uniform).

You can easily modify the above commands to explore some different distributions comparing with the Normal in each case. For example, generate 30 values from an Exponential distribution and make a Normal Q-Q plot:

```
x2 <- rexp(30,0.25);qqnorm(x2)
qqline(x2)
```

You'll see from your plot that any assumption of normality here would be a foolish one.

For a more extreme example of non-normality, let's use a data set called `faithful` that's supplied with R. (Remember `?faithful` to find out more; `head()` to see what the data looks like.)

```
attach(faithful)
hist(eruptions,breaks=25,xlim=c(1,6), probability=TRUE, col=gray(.75),
           main="Eruption durations(min) of the Old Faithful geyser")
```

As you can see the distribution is clearly bimodal; there are quite distinct 'subpopulations' of short and long eruptions. How does this appear in the Q-Q plot?

```
qqnorm(eruptions,main="Eruption durations(min) of the Old Faithful geyser")
qqline(eruptions)
detach(faithful) # tidy up!
```

1.2 Related functions in R

Functions for generating random numbers include `runif()` for Uniform, `rbinom()` for Binomial, `rexp()` for Exponential, `rpois()` for Poisson. Similarly, for the Uniform distribution, `dunif()` returns the density, `punif()` the cumulative distribution function (cdf), `qunif()` the quantiles. See Help for details. You can change to other distributions in the obvious way.

2 Housing data

2.1 The Data

The dataset 'houses' that is on Blackboard contains a subset of data concerning sold houses collected by estate agents in the town of Douglasville, Ohio, USA. Below is a description of fields in the dataset:

home	A unique reference number for each property assigned by the estate agent
nbhd	Notes in which of Douglasville's three main neighbourhoods the house resides. *Key: N=Normsville, R=Realtown, S=Spectralia*
offers	Number of offers that were received on the properties before it sold
sqft	Floor area of the property (measured in square feet)
brick	Is the house constructed of brick? (Yes/No)
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
price	Final selling price ($)
ptype	Type of property. *Key: Semi=Semi-detached, Bung=Bungalow, Det=Detached, Apt=Apartment*

2.2 Error checking

Begin by loading the data into a dataset called `houses` and attach it using the usual method (be sure to indicate that the first row of data gives column headings). Data input errors are fairly common in real world data, with the likelihood of errors increasing with datasets size. It's therefore a good idea when given new data to inspect the data carefully and look for any anomalous values. A good approach for beginning this task is to use the `summary(houses)` command to see some summary statistics. For instance, the field '`brick`' has a meaningless value of 'Yed'. This is close enough to 'Yes' for us to be fairly confident that this was just a typo in data entry. See if you can find another error – there should be one more.

So how do we fix these errors in the data? It's possible but surprisingly long-winded to fix it in RStudio itself by running various commands, or we could instead open the csv file in some spreadsheet software (e.g. MS Excel) or a text editor (e.g. Notepad) and edit it there, then save as a csv.

2.2.1 Using Excel

Open Excel, then within Excel open `houses.csv`. Find and fix the two input errors, then Save As 'houses2.csv' in a suitably named folder in your M drive, being sure to have chosen 'CSV (Comma delimited)' in the 'Save as type' box. Now come back to RStudio, detach `houses` (you may even want to remove it with `rm(houses)`), import `houses2` and attach that. Run a `summary` command of your new dataset to check that everything is sensible now.

2.2.2 Save it!

Once it's been corrected save your workspace (Session > Save Workspace As...) somewhere on your M Drive as this will enable you to reopen the corrected data later when you're working on the assignment.

2.3 Tables and barcharts

Let's begin to analyse the data by using the tables and barcharts methods that we have learned in previous practicals. To begin, we should get an overview of our dataset. Make a table that shows how many houses in the dataset have 2, 3, 4 or 5 bedrooms.

Transfer your table into Word and format it in a visually appealing way as just copying and pasting the text with no modification does not look very professional.

Do different neighbourhoods have a different proportion of 2, 3, 4, 5 bedroom houses? Make a proportional table to answer this and consequently make a stacked barchart to display this graphically. (consult your notes from previous practicals if you can't remember how, or ask for help in the practical). As a reminder: in order to make a barplot in R, you must first make a table (a proportional table in this case); the values in the columns determine the height of the bars.

Label and comment on your barplot. Are the neighbourhoods similar or different? How?

Clearly many other tables could be produced by looking at different combinations of variables. In your assignment you will be given credit for conducting your own investigations, so you are encouraged to experiment with delving into the data to see what you can find!

2.4 Boxplots

Let's use the `ggplot2` package that we used in the last practical to create a boxplot of house prices with separate boxes for each neighbourhood. (Note: you will first have to tick ”ggplot2” in the ”Packages” tab. If it doesn't appear in the Packages list, repeat the installation instructions from the previous practical).

To get you started, a basic boxplot can be generated using the command:

```
ggplot(houses2,aes(x=factor(nbhd),y=price))+geom_boxplot()
```

Again, copy into Word after you've edited titles, axis labels etc. and comment on what you observe. Maybe altering box widths to represent sample sizes would be a good idea? If you've forgotten how to do any of these things, consult earlier practical notes or search the web for ggplot2 help – there's lots out there. In fact searching the web might show you how to make even fancier plots than we've met in the practicals – feel free to use any methods/commands you find so long as you think they're helping you explain features of the data.

How about a few other boxplots to show how house price is distributed among the other categorical variables (not just neighbourhood)? Remember the `factor` command if you encounter one big fat boxplot when you were expecting several.

2.5 Scatterplots

Make a plot of house price vs floor area. How strong is the correlation? Is the relationship linear? Using the colouring techniques we met last week, investigate whether neighbourhood has an effect on price. How about some of the other categorical variables? Copy any insightful plots into Word and comment about what they tell you.

According to the linear regression model, how much does each additional square foot add to the price? Does this change much if you consider each neighbourhood separately? Is there anything surprising about these results and if so can you explain it?

2.6 Further investigations

You're now free to investigate the data further in whatever way you wish. The plots we've made so far are a good start, but you should now investigate some other aspects of the data of your choosing. The `subset` command may be useful for investigating subsets of the data individually.

3 MA26620 Assignment 1

Your first assessed assignment (worth 20% of the module mark) consists of two parts. For the first part of the assignment (counting for 20% of the assignment's marks), you should hand in your solution (word processed) to Question P3.Q2 from Practical 3 on Regression.

For the second, larger, part of the assignment (making up the other 80% of the marks), you should write a report (word-processed) on what you have learned about the Douglasville houses data based on a selection of the material you have gathered in this practical and any further investigations into the data that you conduct. The report should be written for a reader who is a competent statistician but is unfamiliar with the dataset (so for instance, you should refer to house types as detached rather than Det).

The assignment must be handed in via Blackboard under the Assignments content item as a single file (Word or PDF).

Your Report must include:

(i) a title for the report;

(ii) an introductory paragraph explaining briefly the nature of the investigation, where the data came from i.e. the nature of the data and the aims of your analysis;

(iii) the results of your R analysis, in a sensible order (quite probably different from the order you did them in the practicals). Explain the purpose of each element and give your comments; use sub-headings where it helps to make the structure of your report clearer.

(iv) a conclusion where you summarise briefly what you have learned and any questions that remain unanswered.

So if you have not done so already, give the report a title and write an introductory paragraph.

You should include *at least* one of each of the following: table, barplot, boxplot, scatterplot. Marks will be awarded for choice, presentation, editing, and commenting on each of these components, as well as the general structure and flow of your project. You must also give some thought to the order and structure of the content. Consider whether you have used the same style (e.g. font, type size) when you typed the project, that you have organised the text into suitable paragraphs and that the page breaks occur in natural places.

Consider also whether the plots or statistics you have included do indeed illustrate the points you wish to make or whether there are any others which would do the job better. There may be other questions you wish to follow up about this data and some credit (20% of the overall mark for Assignment One) will be given for any relevant additional investigations. Make sure that all plots are well labelled and well presented.

Please note that unnecessary or uninformative graphics will receive no credit and, if excessive, may be penalised. Be selective and always have a reason for your choices. There is no word (or page) limit but in the past, most assignments have been fewer than 10 pages including plots.

Assignments must be handed in via Blackboard by **5pm on Wednesday 28th November**, giving you just less than two weeks to complete the assignment.

3.1 Notes on policies

You should also be aware of the university's unfair practice (plagiarism) policy. Assignments will be handed in via TurnItIn on Blackboard which very effectively makes it clear if your work is not substantively your own. Plagiarism is one of the easiest ways to be excluded from the university or at the very least score zero on the assignment. Moreover, it is unfair to your fellow students, so don't do it. Both the copied and copier will be awarded zero marks, so don't send your work to others.

You should also be aware of the university's extension request policy. Any extension requests must have supporting evidence and be made to your year tutor (Dr Adil Mughal) at least three working days prior to the deadline.

Good luck!