Applied Research Methods

Workshop 4 - Crime Survey Data & Assessing Normality

In the workshops so far, we have been working with a combination of crime incident records and census data. This week, we will switch to examining another valuable form of data when studying crime, which is survey data. Surveys are a classic means of data collection, particularly concerning individuals' experiences or opinions with respect to certain topics. In the context of crime research, they can range from large, national-scale surveys which seek to examine macro-level trends to small-scale surveys examining particular topics (such as those that might be carried out in the course of a dissertation).

This week, we will work with a survey of the former type - the Crime Survey for England & Wales. This is a high-quality survey which gives excellent insight into the experience of crime and related issues in the UK, and it forms the basis for a large volume of research. We will take a first look at this dataset this week, before looking at it in more detail in future Workshops.

In addition to this, we will pick up an important topic from the lectures, which is the normal distribution. We will discuss how normal distributions can be observed in real data, and outline a number of methods for assessing whether a given distribution corresponds to a normal form.

Crime Survey for England & Wales

The <u>Crime Survey for England & Wales (CSEW)</u> is a large-scale survey carried out on an annual basis in the UK. In each annual cycle, over 30,000 respondents are asked about their experience of, and attitudes towards, crime. Because of its large sample size, and because of its direct questioning, it has come to be considered as a very reliable measure of crime; one that is not subject to the issue of under-reporting in the same way as recorded crime data, for example. The CSEW is often cited as an indicator of whether levels of crime in England and Wales are truly rising or falling. Further information about the survey and its history can be found at the <u>Office for National Statistics</u> or indeed on the survey's <u>Wikipedia entry</u>.

Data from the CSEW can be obtained via the UK Data Service (after registering, if necessary). The UK Data Service hosts CSEW data in several forms, which are useful for different research purposes:

- A 'secure' version, which represents the most detailed form of the data. This dataset contains all variables, but includes some which may be sensitive, such as lower-level geographical information relating to respondents' residences. For this reason, access requires approval and training, and must be done in a secure computing environment.
- A 'safeguarded' version, which contains the same data but with sensitive variables removed. This is available to all registered users of the UK Data Service (e.g. anyone affiliated with a university), though the purpose of use must be registered.
- A 'teaching' version, which contains a reduced set of variables in order to simplify the dataset. This is openly available to anyone.

We will use the 'teaching' version here. You could access via the UK Data Service, but to save time, the data you need is in the file "CSEW1314teachingopen.csv".

Data User Guide and Codebook

Many published datasets - and especially those that have been provided for research purposes - are accompanied by contextual documentation explaining the nature of the data. This is good practice, and this information provides vital information when analysing the data - as you know, understanding the nature of data (e.g. how it was collected) has important implications for how it is interpreted. This kind of the information is not typically contained in the dataset itself (e.g. in the spreadsheet file) and so needs to be provided separately. Two kinds of documentation are common:

- User Guide: The user guide for a dataset typically explains the provenance of the data and outlines how it was collected and constructed. For surveys in particular, this should explain the survey methods used in data collection.
- **Codebook**: Data is often stored in such a way that it is not necessarily clear what it represents for example, if a dataset contains a 'sex_at_birth' column containing 0s and 1s, it is far from clear which value represents 'male' and which represents 'female'. In some cases, the name of a column may not make it clear what it records such as 'ethgrp_type3'. A codebook explains this for each variable, it explains what it represents and what the values correspond to.

For the CSEW dataset, both of these are contained in the same file. We have provided this as "CSEW_user_guide_2013-14_teaching_dataset.pdf". Read the first 5 pages of this document, which explain the nature of the dataset we are working with here. Pay particular attention to the 'Derived variables' and 'Missing values within the dataset' sections, since these describe distinctive aspects of the present dataset.

Now take a look at the list of variables on page 6, and the Codebook, which starts on page 7. There is no need to go through this in detail, but try to get a feel for what it decsribes. Although the lack of formatting makes it appear monotonous, it is designed to act as a reference - for any variable you might be interested in, you can look it up and see what the data contains.

For example, take a look at the 'work2' variable at the bottom of page 8. If you simply opened the data and saw a variable called 'work2', without any further information, it would not be clear what this represented... The codebook indicates that this variable stores respondents' answers to the question of whether they had any paid work in the last week. Furthermore, by looking at the 'Numeric' and 'Label' columns, you can see how it is stored in the dataset. This tells us that the response 'yes' is represented by the value 1 in the data, while 'No' is represented by 2 and a refusal to answer is represented by 8. Have a quick browse through some of the other variables.

```
work2

Any paid work in last week?

type: numeric (double)

label: work2

range: [1,8]

unique values: 3

tabulation: Freq. Numeric Label

4,703 1 Yes

4,138 2 No

2 8 Refusal
```

If you were to download the data from the website it would be provided in a STATA file. We have converted this to a .csv file for simplicity. We won't cover this here, but R can read data from other

software, including STATA. To read the data in use the "read.csv" command and use the "head" command to check everything looks sensisble:

>csew<-read.csv("CSEW1314teachingopen.csv")</pre>

>head(csew)

Use the "summary" command for a little more detail. You should see that the dataset contains 32 variables and 8,843 observations. Each of these observations (i.e. rows) corresponds to a single respondent to the CSEW.

>summary(csew)

Missing data

A key thing to note is that many of the cells in the dataset contain just a dot '.'. This represents missing data - values which, for some reason, are not present in the data. There are many reasons why data might be missing, including errors made in the recording process. In the CSEW data, one reason why data points are missing are because not all questions were asked to all respondents – as explained in the user guide, respondents were randomly allocated to 4 groups (A, B, C & D - recorded in the 'split' variable), and each group was asked a particular sub-set of the questions. If you look at the 'walkdark' column using the "table" command, for example, you'll see that only about a quarter of respondents were asked this question. That is, data are missing for 6769 (of 8843) observations.

>table(csew\$walkdark)

It is important to be aware of missing data and to deal with it appropriately. We will discuss how to do this later.

Exploring data

In Workshop 3, we outlined a number of ways to gain insight into data via exploratory data analysis - either by calculating summary statistics or visualising various aspects of the data. Given our new CSEW dataset, this is a natural first step in the analytical process.

The CSEW dataset contains a number of variables, which relate to either the personal and sociodemographic characteristics of respondents, or to their experience of or attitude towards crime. Both of these are aspects for which it would be helpful to understand the nature of the data before proceeding any further with analysis, and we can do this by calculating descriptive statistics or plotting graphs.

There are many variables of interest in the dataset, but we'll just choose one to visualise here. We'll examine the variable 'walkdark', which contains responses to the question 'How safe do you feel walking alone after dark?'. This is a variable that we might expect to reflect the fear of crime in general, and might also reflect individuals' confidence in the police to promote a safe environment. These issues are of both practical and theoretical interest, so might be of interest for a range of research questions.

The variable 'walkdark' is categorical (specifically, it is an ordinal variable, since the responses can meaningfully be ordered), and so a natural way to visualise its distribution is via a bar chart. To do this, we will create a table and then use this to create a barplot (as we did last week):

>csewtab<-as.data.frame(sort(table(csew\$walkdark)))</pre>

>barplot(csewtab\$Freq, names.arg=csewtab\$Var1, las=2, horiz=T, border=F, cex.axis=0.8, cex.names = 0.8)

This should produce the graph in the left panel below. The problem with this plot is that it includes the missing data for this variable, which creates a "blank" category for about 75% of the observations (as the data are missing for about 75% of observations). We can address this in two ways. Here, let's just adjust the plot so that we only show the data for the first five rows of the table, which produces the figure in the right panel below:

>barplot(csewtab\$Freq<mark>[1:5]</mark>, names.arg=csewtab\$Var1<mark>[1:5]</mark>, las=2, horiz=T, border=F, cex.axis=0.8, cex.names = 0.8, xlab="Frequency")



Recall that we can select values form a variable, or a range of them, using square brackets. This is what we have done in the text highlighted yellow above. This tells R that we only want to plot the data for the first 5 rows of the table. From this, we can see that most people feel 'Fairly safe' or 'Very Safe'.

Normal distribution

In the lectures, we introduced the normal distribution, and examined a number of its key properties. The normal distribution is a key concept in statistics - not only does it commonly arise in data, but, via the Central Limit Theorem, it represents the foundation for many of the statistical methods that we will learn on this module.

In the lecture, we talked through the normal distribution in theoretical terms, but how do we recognise one in real-world data? Or, in other words, how do we determine whether a given distribution is normal or not?

The normal distribution is a generic 'bell curve' - a central peak with symmetric 'tails' on either side. There are two important ways in which a distribution can depart from this.

The **skewness** of a distribution relates to the extent to which the data is asymmetric - i.e. that it is biased to one side. This usually manifests itself as a 'leaning' distribution, or - equivalently - one in which the 'tail' stretches further on one side of the peak than on the other. **Positive** (or 'right') skew refers to a situation where there is more data on the right of the peak than would be expected, while **negative** (or 'left') skew refers to the opposite.



The **kurtosis** of a distribution relates to the extent that the distribution is 'peaked'. Again, this refers to peaks that are either more or less pronounced than would be expected according to the standard 'bell' shape. **Positive** kurtosis refers to a situation where the curve is more 'peaked', while **negative** kurtosis refers to a situation where the curve is 'flatter'.



Both skewness and kurtosis have corresponding statistics, and are calculated as standard by most statistical packages. They can be used to assess the normality of a given distribution - **for normal data, both skewness and kurtosis will have value 0**, and departures from this in either direction correspond to the 'positive' or 'negative' forms above.

Assessing normality

When given a dataset, there are a number of ways in which its normality can be assessed. The first of these is by examining the skewness and kurtosis statistics. The R base package does not include functions to do this, though they can be coded. As such, we will illustrate a different approach. The second approach is a graphical one, and one that is already familiar to us - to plot a histogram. We have already seen that histograms are an effective way to visualise the 'shape' of numerical distributions, which is the task that we are trying to accomplish here. By plotting a histogram, we can make a visual assessment of how closely the distribution corresponds to a normal curve, which may depend on the volume of data and the width of bins.

The third, and slightly more sophisticated, method is another graphical one, and is called a normal probability plot (also known as a normal Q-Q plot). This is a plot in which the data values are displayed in a scatterplot against those that would be expected if the distribution was truly normal. For a normal distribution, these values would be expected to lie along a straight line. An example of this is shown below - the plot on the right is a normal probability plot of the data that is shown in the histogram to the left. Since the data lies along the straight line, this data appears to be approximately normal.



If you want to reproduce these plots, type:

```
>data<-rnorm(500, 120,20)
>par(mfrow=c(1,2))
>hist(data, col="green", border=FALSE, seq(0,200,2),main="Sample=500",
freq=TRUE, xlim=c(50,200), xlab="Value", ylab="", yaxt='n')
```

```
>qqnorm(data)
>qqline(data, lty=2)
```

The rnorm command generates data for a variable (that I labelled data) that is normally distributed. It will generate different data each time you run it. It has 500 observations, a mean of 120 and a standard deviation of 20. The qqnorm command generates the Q-Q plot. In the final line of code, the "Ity" parameter is used to tell R what type of line we want to plot. In this case, the straight line shown is a dotted line (Ity=2).

When data is not normal - because it displays skewness or kurtosis, or both - it will be manifested in the probability plot. In addition to visualising the distribution, we can use the **Shapiro Wilks test**. For this test, the null hypothesis is that the data tested come from a normal distribution and a significant p-value would indicate departure from a normal distribution. To do this, type:

>shapiro.test(data)

This should generate output like that below (the result may not be identical as the "rnorm" command generates different data each time). In this case, the p-value is greater than .05, which means that we cannot reject the null hypothesis – the evidence suggests the data are from a normal distribution.

Shapiro-Wilk normality test

data: data W = 0.99558, p-value = 0.1705

Assessing normality in the CSEW Assessing normality in the CSEW

To finish off, we'll assess the normality of one of the variables in our dataset. Our 'teaching' dataset contains a small number of numerical variables which summarise responses from the larger dataset - the User Guide briefly explains how these were derived.

One of these variables is 'worryx', which reflects the respondent's worry about being a victim of crime (for which high scores corresponding to high levels of worry). This is the kind of variable that we might wish to investigate when examining fear of crime, and in particular when exploring whether this fear is influenced by other socio-demographic variables.

For now, let's simply take a look at the variable to see whether it appears to follow a normal distribution using the methods outlined above. First, let's plot the data. If you do this, you will run into problems due to the fact that many of the observations contain "NA" values. To deal with this, we will create a new dataframe in which we include those observations for which our variable of interest contains numerical values. To do this, type:

>newdata<-csew[which(worryx!="NA"),]</pre>

This creates a new dataframe called "newdata". This contains only 2,074 rows of data (check this using the summary command). We can now generate a histogram:

```
>hist(newdata$worryx, col="green", border=FALSE, freq=FALSE,
xlab="Value", main="")
```

This should produce the plot in the left panel shown below. Let's also generate a Q-Q plot:

>qqnorm(newdata\$worryx)

>qqline(newdata\$worryx, lty=2)



The plot suggests that skewness and kurtosis are evident here. In the centre of the plot, where the majority of the data lies, the trend is clearly curving upwards, reflecting the positive skewness in the data. At either end, the distinctive deviations from the main trend (which give the overall curve a sort of 'S' shape) reflect negative kurtosis, as per the table above. To test this with the Shapiro Wilk test, type:

>shapiro.test(newdata\$worryx)

Shapiro-Wilk normality test

data: newdata\$worryx W = 0.9298, p-value < 2.2e-16

The statistically significant p-value, which is much less than 0.05, indicates that we can reject the null hypothesis – the data are not drawn from a normal distribution.

We have assessed the normality of the 'worryx' variable from a number of perspectives, and a clear picture has appeared that the variable does not follow a normal distribution. This will have implications for how we treat it when we begin more formal inferential analysis in future Workshops.