A system for statistical analysis

04-08 July 2016

Data coding, manipulation and management

Graeme.Hutcheson@manchester.ac.uk

Manchester Institute of Education, University of Manchester

This workshop deals with...

- Measurement theory and data classification
- Data-frame structure
- Information coding
- Data management
Data are an integral part of the ‘Research Cycle’

- Research conclusions are based entirely on analyses of the DATA - if results are to apply to the population, the data MUST provide an accurate representation of the attribute.
- The overall aim of this session is to provide a considered method for coding and dealing with data.
- We will not be dealing with issues of sampling, or with issues relating to specific analyses.
- We will restrict ourselves to the coding and management of data. Although this area is probably the easiest to deal with and ‘get right’, it continues to cause problems for analysts.
Types of data

Although a number of different schemes have been proposed that utilise a variety of categories and sub-divisions, we will only distinguish between three distinct scales of measurement — unordered categorical, ordered categorical and numeric, based on a simplified version of Stevens' 1946 classification of measurement scales.

It is important to identify these three measurement scales as they are qualitatively distinct and represent what is, perhaps, the minimum required for any course that aims to provide a general introduction to data analysis.
1: Unordered categorical scales

An unordered categorical categorical scale of measurement is achieved when the data are recorded as categories which have no meaningful order. The only information provided is the category identifier. This scale is also known as a **classificatory** scale and a **labelling system**.

### Examples of unordered categorical data

<table>
<thead>
<tr>
<th>Car Manufacturer</th>
<th>Gender</th>
<th>Treatment Group</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>ford</td>
<td>male</td>
<td>a</td>
<td>subject01</td>
</tr>
<tr>
<td>citroen</td>
<td>male</td>
<td>a</td>
<td>subject02</td>
</tr>
<tr>
<td>volvo</td>
<td>female</td>
<td>a</td>
<td>subject03</td>
</tr>
<tr>
<td>volvo</td>
<td>male</td>
<td>b</td>
<td>subject04</td>
</tr>
<tr>
<td>renault</td>
<td>female</td>
<td>b</td>
<td>subject05</td>
</tr>
<tr>
<td>land Rover</td>
<td>female</td>
<td>b</td>
<td>subject06</td>
</tr>
<tr>
<td>toyota</td>
<td>female</td>
<td>c</td>
<td>subject07</td>
</tr>
<tr>
<td>toyota</td>
<td>male</td>
<td>c</td>
<td>subject08</td>
</tr>
<tr>
<td>volvo</td>
<td>male</td>
<td>c</td>
<td>subject09</td>
</tr>
<tr>
<td>chrysler</td>
<td>female</td>
<td>c</td>
<td>subject10</td>
</tr>
</tbody>
</table>

*note: Subject is a categorical variable and is not recorded as a number.*
Operations that can be applied to data recorded on an unordered categorical scale

**Unordered operations:**

- most frequent car is Volvo (n=3) ✓
- same numbers of male and females (n=5) ✓

Applying statistics to data recorded on an unordered categorical scale

Unordered scales only contain information about category membership. These scales can, therefore, only be analysed using techniques that apply to category membership (for example, chi-square and multinomial logistic regression).

Techniques designed for ordered or numeric data **cannot** be used to model these data (as there are no numbers in the data an order cannot even be established).
2: Ordered categorical scales

An ordered categorical scale of measurement is achieved when the data are recorded as categories that can be arranged in order according to some criteria. The only information provided is the category identifier from which an order can be established.

Examples of ordered categorical data

<table>
<thead>
<tr>
<th>Highest Qualification(^1)</th>
<th>Agreement Rating(^2)</th>
<th>Exam Grade(^3)</th>
<th>Mental Health Rating(^4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>strongly agree</td>
<td>B</td>
<td>no symptoms</td>
</tr>
<tr>
<td>2</td>
<td>disagree</td>
<td>A</td>
<td>impaired functioning</td>
</tr>
<tr>
<td>3</td>
<td>neither</td>
<td>A</td>
<td>no symptoms</td>
</tr>
<tr>
<td>4</td>
<td>disagree</td>
<td>C</td>
<td>mild symptoms</td>
</tr>
<tr>
<td>5</td>
<td>agree</td>
<td>D</td>
<td>moderate symptoms</td>
</tr>
<tr>
<td>6</td>
<td>strongly disagree</td>
<td>E</td>
<td>mild symptoms</td>
</tr>
<tr>
<td>7</td>
<td>agree</td>
<td>D</td>
<td>moderate symptoms</td>
</tr>
<tr>
<td>8</td>
<td>neither</td>
<td>B</td>
<td>mild symptoms</td>
</tr>
<tr>
<td>9</td>
<td>disagree</td>
<td>A</td>
<td>mild symptoms</td>
</tr>
<tr>
<td>10</td>
<td>agree</td>
<td>D</td>
<td>impaired functioning</td>
</tr>
</tbody>
</table>

Variable Order:
\(^1\)Highest Qualification: No qualification, O-level, A-level, Degree, Masters, Doctorate.
\(^2\)Agreement Rating: strongly agree, agree, neither, disagree, strongly disagree.
\(^3\)Exam Grade: A, B, C, D, E.
\(^4\)Mental Health: no symptoms, mild symptoms, moderate symptoms, impaired functioning.
Operations that can be applied to data recorded on an ordered categorical scale

**Ordered operations:**

- agreement: str. agree > neither 
- educational achievement: degree < doctorate

**Unordered operations:**

- mental health: mild symptoms is the most frequent
- Examinations: e and c are the least frequent

Applying statistics to data recorded on an ordered categorical scale

Ordered scales only contain information about category membership and the order of these categories.

These scales can be analysed using techniques that apply to ordered categories (for example, proportional-odds logit models, Kruskal-Wallis, Friedman, Spearman’s correlation).

They may also be analysed using techniques designed for unordered categorical data (eg., multinomial logistic regression and chi-square), although the information about order will be lost.

Techniques designed for numeric data **CANNOT** be used to model these scales (as there are no numbers in the data).
3: Numeric scales

A numeric scale of measurement is achieved when the recorded data can be considered to have a direct relationship with the variable being measured. At a very basic level, the data provides information about the actual magnitude (size, quantity, distance, etc.,) of the information. In other words, the numbers representing the variable have substantive meaning.

Examples of numeric data

<table>
<thead>
<tr>
<th>Temperature difference (deg C)</th>
<th>Daily gas consumption (m³)</th>
<th>Examination result (%)</th>
<th>Hourly rate of pay (£ sterling)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.3</td>
<td>69</td>
<td>62</td>
</tr>
<tr>
<td>2</td>
<td>9.8</td>
<td>74</td>
<td>62</td>
</tr>
<tr>
<td>3</td>
<td>11.4</td>
<td>82</td>
<td>58</td>
</tr>
<tr>
<td>4</td>
<td>11.5</td>
<td>81</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>17.8</td>
<td>67</td>
<td>71</td>
</tr>
<tr>
<td>6</td>
<td>13.1</td>
<td>85</td>
<td>59</td>
</tr>
<tr>
<td>7</td>
<td>13.4</td>
<td>75</td>
<td>63</td>
</tr>
<tr>
<td>8</td>
<td>13.6</td>
<td>69</td>
<td>63</td>
</tr>
</tbody>
</table>
Operations that can be applied to data recorded on a numeric scale

**Numeric operations:**

\[ 74m^3 + 69m^3 = 143m^3 \checkmark \]

\[ 73\% - 61\% = 12\% \checkmark \]

**Ordered operations:**

\[ £7.98 > £5.80 \checkmark \]

\[ 69m^3 < 74m^3 \checkmark \]

**Unordered operations:**

most frequent hourly pay = £5.80 (n=4) \checkmark

most frequent examination result = 63\% (n=2) \checkmark

Applying statistics to data recorded on a numeric scale

Numeric scales contain information about quantities and can be analysed using techniques that apply to numeric scales (for example, OLS regression models, \( t \)-tests, ANOVA, Pearson's correlation).

Numeric data may also be analysed using techniques designed for ordered and unordered categorical data (eg., chi-square, multinomial and proportional-odds models), although information about order and quantity will be lost.
Data Frames and Data Structures

Data-frames

Data may be represented using vectors, matrices, lists and data-frames. The system proposed here uses data-frames as they are, perhaps, the easiest to deal with and also provide a structure for data that will be familiar to anyone who has used a spreadsheet or a statistics package such as Excel, Gnumeric, SPSS, STATA, S-Plus or SAS.

Data-frames are simply matrices where each column can contain a different scale of data (for example, numeric or categorical). Data-frames are rectangular in shape as they have the same number of observations, or cases, recorded for each variable. Data-frames are particularly useful as they can be used to represent entire data sets and provide a format for easily dealing with data.
An important distinction when representing information is that between a **data-table**, which aims to illustrate the data concisely and clearly and a **data-frame**, which is a representation of the data in a form that can be used by an analysis package.

A common cause of confusion is that data-tables are often structured differently to data-frames. For example, although the data-table shown below illustrates the number of customers entering three stores over five different one-hour periods clearly and consisely, this structure may not be appropriate for computing statistics or drawing graphics. When the data are represented in this format, they cannot be easily manipulated by a statistics program (see ‘Data Set Structure’ in the SAGE Dictionary of Quantitative Management Research - available for download on the course web site).

<table>
<thead>
<tr>
<th>Number of customers entering store</th>
<th>store A</th>
<th>store B</th>
<th>store C</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>
A more useful format for data represents each variable in its own column which enables individual variables to be explicitly identified for analytical purposes. For example, if we wish to model the number of people entering individual stores (using, for example, a regression model for count data), we need to identify the ‘number entering’ and the ‘store’ as single variables. A model investigating the relationship between the ‘number entering’ and the ‘store’ is given by the equation...

\[ \text{Number} \sim \text{Store} \]

The data, as it is structured in the table above, does not allow us to do this easily as each column of data provides information about more than one variable; the store and the number of customers (the data set does not contain 2 individual variables to put into the model).

The data need to be re-structured so that each column only provides information about a single variable. Although the resulting table is larger (it contains nearly twice the number of cells) it is often much easier to work with the data in this format.

These two different formats are also known as ‘long’ and ‘short’ formats. The data table above is an example of a short format, whereas the one below shows the same information in a long format.
A data-frame showing the number of customers entering three stores over five different one-hour periods

<table>
<thead>
<tr>
<th>Store</th>
<th>customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>storeA</td>
<td>13</td>
</tr>
<tr>
<td>storeA</td>
<td>14</td>
</tr>
<tr>
<td>storeA</td>
<td>11</td>
</tr>
<tr>
<td>storeA</td>
<td>12</td>
</tr>
<tr>
<td>storeA</td>
<td>9</td>
</tr>
<tr>
<td>storeB</td>
<td>5</td>
</tr>
<tr>
<td>storeB</td>
<td>8</td>
</tr>
<tr>
<td>storeB</td>
<td>3</td>
</tr>
<tr>
<td>storeB</td>
<td>9</td>
</tr>
<tr>
<td>storeB</td>
<td>4</td>
</tr>
<tr>
<td>storeC</td>
<td>9</td>
</tr>
<tr>
<td>storeC</td>
<td>12</td>
</tr>
<tr>
<td>storeC</td>
<td>17</td>
</tr>
<tr>
<td>storeC</td>
<td>8</td>
</tr>
<tr>
<td>storeC</td>
<td>12</td>
</tr>
</tbody>
</table>

It is easy to change between long and short formats and most users will probably do this by simply cutting and pasting blocks of data within a spreadsheet package. When manual manipulation of data becomes impractical, automated tools are available within the R program, Rcmdr and a number of add-on packages (see, for example, the 'reshape' and 'memisc' packages). These techniques will, however, be of most interest to those users who deal with data regularly or who have to restructure large amounts of data.

Note: Researchers should consider carefully whether to invest time in learning these functions. Although useful, it is often much quicker (and safer) to restructure individual data sets within a spreadsheet package using cut-and-paste commands in a familiar spreadsheet.
Frequency tables

A popular and very economical format for presenting categorical data is in the form of contingency tables that identifies each cell and also provides a cell count.

For example, the ‘HairEyeColor’ data set that can be loaded directly from the ‘datasets’ library which is accessed via the Data, Data in packages, Read dataset from an attached package... menu options in the Rcmdr.

The R code to load these data as a data-frame is...

```r
data(HairEyeColor, package="datasets")
HairEyeColor <- as.data.frame(HairEyeColor)
```

which can be copied directly from the Rcmdr script window into your R-script or R-markdown file.

HairEyeColor data in frequency-table format (32 rows)

These data can be viewed using the ‘view data set’ button in the Rcmdr.
Frequency tables

Although this format is very economical, it does not always allow easy analysis as we often need to represent each of our observations on a single row. The variable Hair, for example, only contains 32 entries, 8 for each hair colour even though there are data from 592 people. A bar chart of Hair only shows the number of distinct groups, not the number of individual cases. A bar plot of hair colour results in the graph...

![Frequency Chart](image)

Frequency Tables - Transforming

Although some analytical techniques can deal with frequency-tables directly, it is often useful to restructure these into long-format data-frames with each row representing a single case.

Frequency tables can be easily re-structured into data-frames using a spreadsheet package, or R.

If you are interested... the R code to re-format the 32 row HairEyeColor dataset into a 592 row long-format data set (called HairEyeColorLong) is...

```r
HairEyeColorLong <- as.data.frame(lapply(HairEyeColor, function(x) rep(x, HairEyeColor$Freq)))
```

(Not a particularly ‘friendly’ function, but it is one that is rarely used.)
HairEyeColor data in Long-format (592 rows)

Viewing these data using the Rcmdr ‘view data set’ button...

Now when we plot a bar chart of Hair, we get information about the number of individual cases...

We are also able to run models of each of the variables in the dataset. For example, a model predicting eye colour using information about hair colour and sex...

\[ \text{Eye} \sim \text{Hair} + \text{sex} \]
Coding Data

Data can be coded using many different packages. Here we will describe a simple method using...

- a spreadsheet (Gnumeric, Libre Office or Excel) to input and manipulate data.
- a simple data-frame format where variables are represented in columns and individual cases represented in rows.

A simple spreadsheet with three variables (named Variable1 to Variable3) and 4 cases (data from 4 individuals) looks like...
A simple spreadsheet data-frame

To construct our data set, we simply add variables and cases to the spreadsheet following a number of conventions...

Coding aims

There are many rules and conventions that can be applied to coding data. The following are a few that can be applied as general rules for data coding. The main principles are that data should...

▶ accurately represent the measurement scales (in particular - code categories as categories and not as numbers).
▶ be able to code information without the use of any ‘hidden’ codes or labels.
▶ be coded clearly and unambiguously.
▶ be of a form that can be easily imported into different software packages (the coded data should be transportable).
General Coding Conventions

- Variable names should be included in the first row.
- There should be NO EMPTY ROWS OR COLUMNS in the data file (spreadsheet).
- Avoid spaces, commas, underscores, quotation marks or mathematical signs and other 'strange' characters whenever possible...
  
  $%&*?/\ | "!~#+-_

- Avoid using highlights, colours, lines or anything else in the data files.
- No formulas.

Coding measurement scales

- Explicitly code the 3 basic measurement scales
  
  - **Unordered categorical**
    Code data using text
  
  - **Ordered categorical**
    Code data using text preceded by a number to explicitly indicate order (so that it orders the data appropriately for graphics and analyses)
  
  - **Numeric**
    Code data using numbers that best represents the information.
Coding missing data

- Problems may arise for any coding system which uses numeric, ‘hidden’ or ‘substitution codes’ to indicate missingness.
- Information about missing data is qualitatively different to the information about the variable.
- Missing data should be indicated using a unique categorical indicator code such as “NA” which can be ignored by the analysis (SPSS, however, insists on a numeric code for missing data which is identified using a ‘hidden’ label).
- If there are multiple missing codes (e.g., not applicable, unanswered, spoiled) this information needs to be coded separately using an additional variable.

An example data set

<table>
<thead>
<tr>
<th>Subject</th>
<th>Age</th>
<th>Nationality</th>
<th>EconStatus</th>
<th>FactorSocial</th>
<th>ManGrade</th>
<th>ManGradeMiss</th>
</tr>
</thead>
<tbody>
<tr>
<td>subject01</td>
<td>23</td>
<td>Russian</td>
<td>2ses</td>
<td>NA</td>
<td>1junior</td>
<td>answered</td>
</tr>
<tr>
<td>subject02</td>
<td>24</td>
<td>English</td>
<td>4ses</td>
<td>-1.231</td>
<td>NA</td>
<td>spoiled</td>
</tr>
<tr>
<td>subject03</td>
<td>31</td>
<td>Welsh</td>
<td>1ses</td>
<td>0.821</td>
<td>1junior</td>
<td>answered</td>
</tr>
<tr>
<td>subject04</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>3upper</td>
<td>answered</td>
</tr>
<tr>
<td>subject05</td>
<td>43</td>
<td>Irish</td>
<td>2ses</td>
<td>0.076</td>
<td>2middle</td>
<td>answered</td>
</tr>
<tr>
<td>subject06</td>
<td>41</td>
<td>German</td>
<td>2ses</td>
<td>NA</td>
<td>2middle</td>
<td>answered</td>
</tr>
<tr>
<td>subject07</td>
<td>19</td>
<td>German</td>
<td>3ses</td>
<td>2.652</td>
<td>3upper</td>
<td>answered</td>
</tr>
<tr>
<td>subject08</td>
<td>38</td>
<td>Portuguese</td>
<td>3ses</td>
<td>1.611</td>
<td>2middle</td>
<td>answered</td>
</tr>
<tr>
<td>subject09</td>
<td>59</td>
<td>Spanish</td>
<td>NA</td>
<td>-0.812</td>
<td>NA</td>
<td>notApplic</td>
</tr>
<tr>
<td>subject10</td>
<td>24</td>
<td>Scottish</td>
<td>2ses</td>
<td>3.769</td>
<td>NA</td>
<td>notApplic</td>
</tr>
<tr>
<td>subject11</td>
<td>39</td>
<td>Irish</td>
<td>4ses</td>
<td>0.023</td>
<td>1junior</td>
<td>answered</td>
</tr>
<tr>
<td>subject12</td>
<td>22</td>
<td>Irish</td>
<td>3ses</td>
<td>-0.182</td>
<td>NA</td>
<td>spoiled</td>
</tr>
<tr>
<td>subject13</td>
<td>64</td>
<td>Japanese</td>
<td>5ses</td>
<td>0.034</td>
<td>3upper</td>
<td>answered</td>
</tr>
</tbody>
</table>
The data structure above enables the management grade (ManGrade) and the missing data (ManGradeMiss) to be modelled and graphed. Below are shown the relationships between these variables and the FactorSocial variable.

Note: the R commands for these graphics can be copied from the Rcmdr and edited, if required.
Saving data-frames

- There are many software programs that can be used to save data - R, SPSS, Excel, Gnumeric, Stata, S-Plus, Dbase. Although these all have their own ‘specialised formats’, they are all able to load data saved in certain text-based formats.
- For true data transparency and data portability, it is worth using a text-based method for saving your data files.
- When saving as a text file you will need to select an appropriate character to distinguish between variables (the ‘delimiter’).

Saving data-frames: comma-separated variables (.csv)

- tabs and commas are the most common ‘delimiters’ as they do not tend to be coded as part of the data (as opposed to blank spaces which often appear in multiple word categories and post codes). I use comma delimited datasets as these have the advantage of an explicit character (tabs are represented as a hidden character in data files and can be difficult to identify and deal with) and also a special file-designator (.csv) so that data files are obvious.
- Many packages use comma separated variable files (.csv files) as a default data type and these are also easy to read and save in spreadsheet packages.
The problem of data proliferation

Data are routinely manipulated and changed by, for example, recoding variables, combining and renaming categories, transforming observations, changing reference categories, dummy-coding variables and changing measurement scales (e.g., recoding ordered categorical data into numeric so that a factor analysis can be computed). If these changes are written to the data-frame, it quickly increases in size and complexity with individual variables represented in multiple columns.

Data may also be recoded and saved into ‘new’ data sets; For example, ‘Results.csv’, ‘ResultsAll.csv’, ‘ResultsFinal.csv’, ‘ResultsPublish.csv’, ‘ResultsFinal2.csv’, ‘ResultsVeryFinal.csv’, ‘ResultsVeryFinalAbsolutelyNoMessing.csv’ etc.

This proliferation of data sets can cause problems.

The problem of data proliferation

Multiple variables and data files are problematic for a number of reasons...

▶ It is confusing as multiple variables may represent a single attribute.

▶ If data are amended or corrected, all data files and recoded copies of the data need amending along with the ‘original’ variable.

▶ Analyses may apply to different versions of your data.
A solution to data proliferation

Use a ‘master’ data-frame that contains the most accurate and complete representation of the information available. This data-frame is the only one that is saved to disk and is the data file that is accessed at the start of all analysis sessions.

Any changes to the data (recodes, transformations, renaming etc.,) should be made on a temporary basis and not saved to the master data file (unless absolutely necessary).

The master data file should only include the most complete coding of the information with each variable accurately coded according to its measurement scale.

Data manipulation

Even when the data are coded in an appropriate format, they may still require manipulating to run certain analyses and graphics. For example, category labels may need to be changed or categories collapsed; numeric variables may need to be transformed; ordered categorical data may need to be re-coded as numeric so that factor analyses can be run and the order of categories may need to be changed to optimise graphical plots.

These can all be done easily in the Rcmdr as part of the analysis process. There is usually no need to save data manipulations to the data set as these can be easily saved to command files.
Some of the common data transformations that can be applied are:

- Recoding categories (e.g., 5 SES categories to 3)
- Sub-setting data (e.g., only include males)
- Re-labelling (change category names)
- Data transformation (log, sqrt, etc.)
- Changing ordered categorical data into numeric (for example, to allow a factor analysis).
- Changing the reference category

These may all be achieved at the analysis stage. For information about specific commands, consult the software manuals (see help in the R-studio, or read manuals for packages **memisc** and **reshape**).

This method of working with data has a number of advantages...

- Data sets and variables do not proliferate.
- You may only load the data you actually need for each analysis.
- The command files document all data and analysis operations.
- It is simple to exchange command files securely.
Importing data from SPSS (an example of importing data)

SPSS data can be imported using the Rcmdr.

In general, the data import for SPSS files is quite accurate, with missing data appropriately re-coded as \texttt{NA} and variable names identified accurately.

You will, however, need to decide if you want to import your categorical data as numbers (numeric variables) or as words (categorical variables)...
Importing data from SPSS preserving factor names

Importing data from SPSS preserving numeric codes
For detailed information about importing SPSS data, refer to the documentation from the package `foreign` which can be read by clicking on the package name in the packages window of the R-studio.

You should always carefully check the accuracy of any imported data.

Conclusion
Data coding and data management is very important, but is often neglected.

Using correct coding reduces errors in analysis.

Using a master data file and real-time software data manipulation minimises data proliferation, confusion and errors.

Correctly coded data and good data management will almost certainly save you time in the long run.

EXERCISE:
Data manipulation
During this exercise you will load a csv data file, run some descriptive statistics, manipulate and change the data and save your analyses to a script or markdown file. An example markdown file is available on the course CD (session03exercise.Rmd).

**Load the ‘Arrests’ data file from the course CD...**
(Use the **Import Data set, From Text File...** tab in the Rstudio.)

**Use the Rcmdr menu Statistics, Summaries... to run descriptive statistics on the Arrests data.**

Look at the output to see how R treats the variables - some are treated as continuous and some as categorical, purely on the basis of how they have been coded...

---

**Recode continuous to categorical...**

In the ‘Arrests’ data set, ‘year’ is considered as a continuous variable - this is not ideal for a number of analyses or graphics...

**Recode ‘year’ as a categorical variable.**

Use the Rcmdr menu ‘Data, manage variables in active dataset’ and save the new variable as ‘yearCAT’. Either supply names for each category, or use the existing numbers.

**Re-run the descriptive statistics and see the way yearCAT is described compared to year.**

‘yearCAT’ allows a variety of graphics to be produced...

- Use the Rcmdr graphs menu to produce a ‘plot of means’ for the number of checks per year.
- Use the Rcmdr graphs menu to produce a ‘Barplot’ showing the number of cases for each year.
Recode categories...

Using the Arrests data set, draw a bar chart of ‘released’...

This graphic may look better if the x-axis title is removed and the
category labels are changed from ‘yes’ and ‘no’ to ‘detained’ and
‘released’.

**Change the category labels for released**

Use the Rcmdr menu ‘Data, Manage variables..., Recode
variables...’; save the variable as releasedRECODE using the
recode directives...

"No" = "detained"
"Yes" = "released"

Re-draw the barchart for released (use ‘options’ to remove the
x-axis label)

---

It is useful to treat ‘checks’ as a categorical variable and change
the number of categories (we will be using this variable later in the
course).

**Re-code checks changing it to ordered categorical with 4 groups**

Use ‘Data, Manage variables..., Recode variables...’; make
each new variable a factor, name the new variable checksORD
and use the following recode directives...

0 = "0 checks"
1 = "1 check"
2 = "2 checks"
else = "3 or more checks"

**Draw a barchart for the variable checksORD**

You are now able to analyse checksORD as ordered
categorical, which may be more appropriate for your analysis.