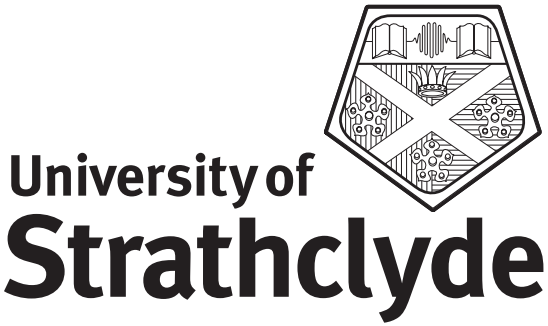


MM104 / MM107
Statistics and Data Presentation Semester 2

Project 1: Visualising Categorical Data

Ainsley Miller
ainsley.miller@strath.ac.uk

Lecture Overview

In this lecture we will

- discuss what categorical data is.
- introduce sampling, in particular, random sampling.
- discuss how to visualise categorical data.
- learn how to use Minitab to create graphs and tables.

Types of Data

There are two types of data

1. Qualitative data (also called categorical data)
2. Quantitative data (also called numerical data)

Qualitative/Categorical data

Qualitative data is also called categorical data. This data type answers the question "What type" or "Which category".

Common examples of qualitative data are:

- Hair colour
- Eye colour
- Martial status

Qualitative/Categorical Data

Qualitative data can be split into two sub-categories

- Nominal data
- Ordinal data

Nominal data is data that does not follow any particular natural pattern e.g. hair colour, eye colour

Ordinal data is data that can be put into an order e.g. pain level - low/moderate/high

Population

All individuals (people, animals, plants objects etc), in a group that we are interested in and for whom many observations can be made.

Sampling

Most statistical work is concerned with using samples to draw conclusions about some larger population.

> A sample is defined as being a group of individuals taken from a larger population and used to find out something about that population.

Representative Samples

A sample should be representative of the population and should have all the characteristics in terms of the proportion of individuals with particular qualities as has the whole population.

In a sample from a human population, for example, the sample might have, for example, the

- same proportion of men and women as in the population.
- same proportion in different age groups as in the population.
- same proportion in occupational groups as in the population.
- same proportion with different diseases as in the population.

Random Sampling

In order to get a sample which is representative of the population a way of choosing members of the sample which does not depend on their own characteristics is needed.

The only way to do this is to select them at random, so that whether or not each member of the population is chosen for the sample is purely a matter of chance.

Physical methods of randomising are often not suitable for statistical work. Random numbers generated by a computer are generally used.

Random Sampling Cont...

Random sampling ensures that the only ways in which the sample differs from the population will be those due to chance.

Why Illustrate Data ?

A picture tells a thousand words.

Data visualisation allows us to curate data into a form that's easier to understand. These visualisations allow us to highlight the trends and outliers.

Contingency Tables

Contingency tables (two-way tables) are used in statistics to summarise the relationship between several categorical variables.

One Way Tables in Minitab

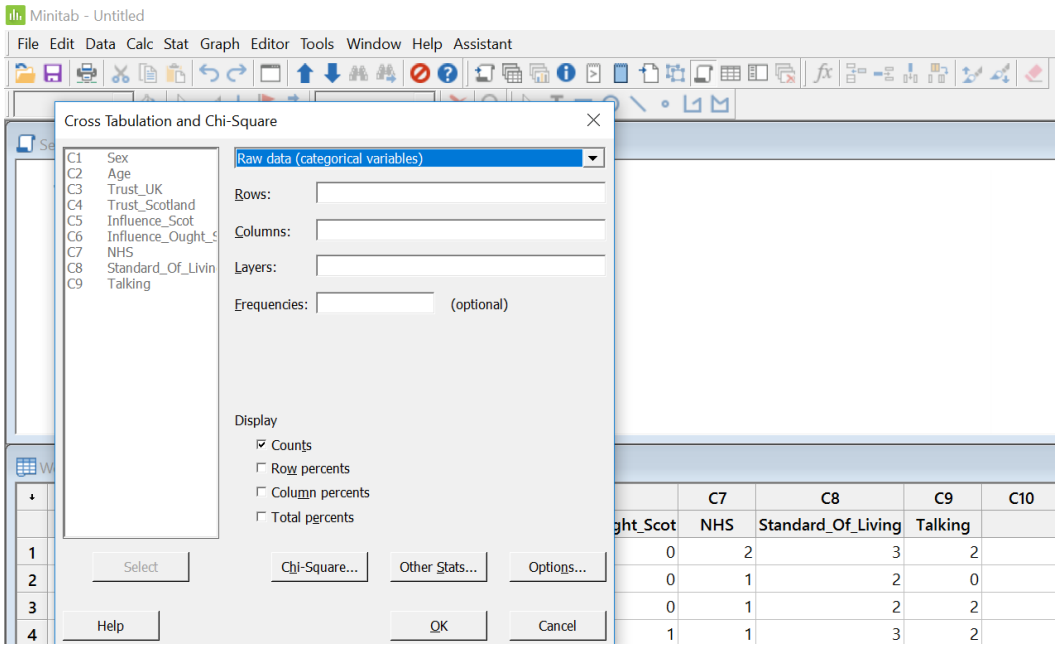
You can firstly get a feel for your data by creating what is called a One-Way table this gives you the number of observations recorded for each category. A one way table displays categorical data in the form of frequency counts and/or relative frequencies .

Stat > Tables > Tally Individual Variables

Contingency Tables in Minitab

You create a Contingency Table using the following steps:

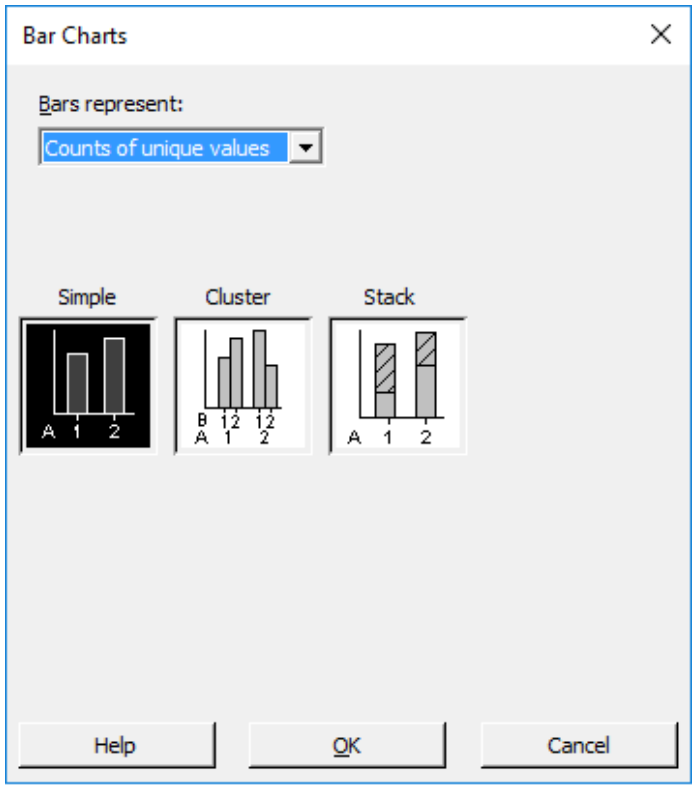
> Stat > Tables > Cross Tabulation and Chi Square



One Way Bar Chart in Minitab

A one way bar chart is the chart equivalent of a one-way table.

Graph > Bar Chart > Simple > OK



One Way Bar Chart in Minitab cont...

You will need to add additional information to your chart.

Click on *Labels* to add a title and axes labels.

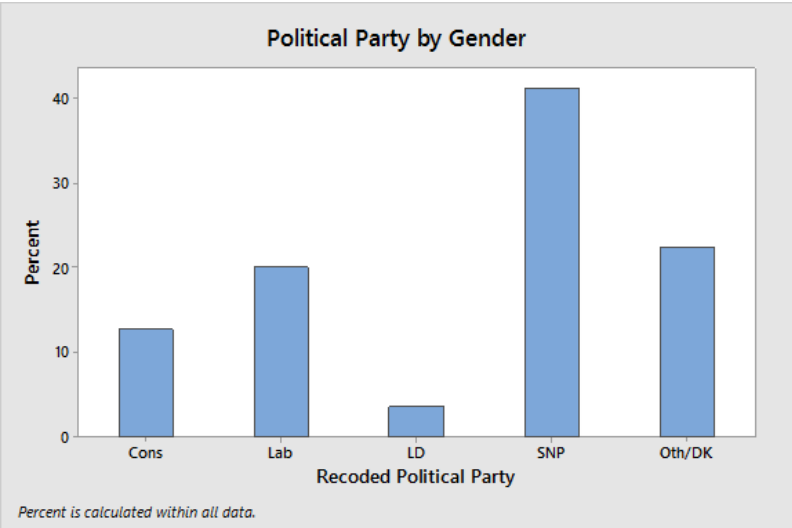
If appropriate you may wish to only plot subsets of your data. To do this click on *Data Options* to specify subsets of the data to plot, Click on OK after setting a condition.

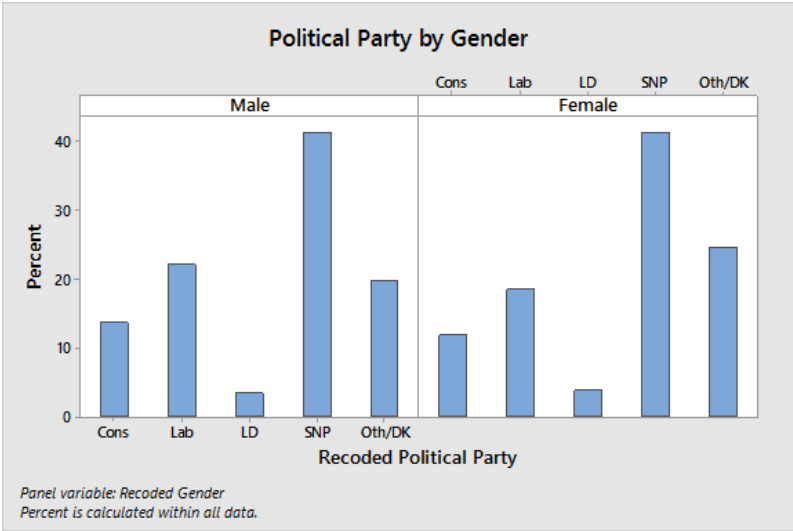
One Way Bar Chart in Minitab cont...

You may also want to compare bar charts across values of a second variable. Click on *Multiple Graphs* to compare bar charts across values of a second variable It is normally a good idea to check the *Y Scale* over all graphs. Click on the *By Variable* tab to select the variable to split the chart by and click OK.

Once you have made all the selections for the chart click on OK

Examples of One-Way Charts

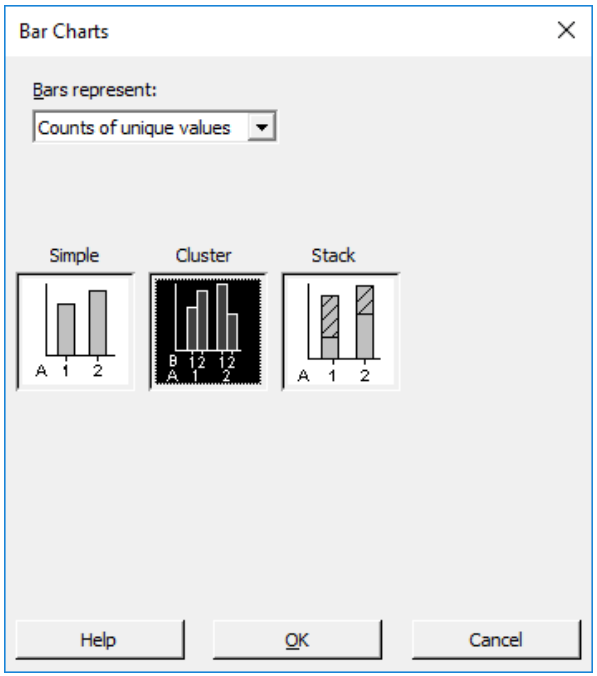




Two Way Clustered Bar Chart

A clustered bar chart displays more than one data series in clustered horizontal columns.

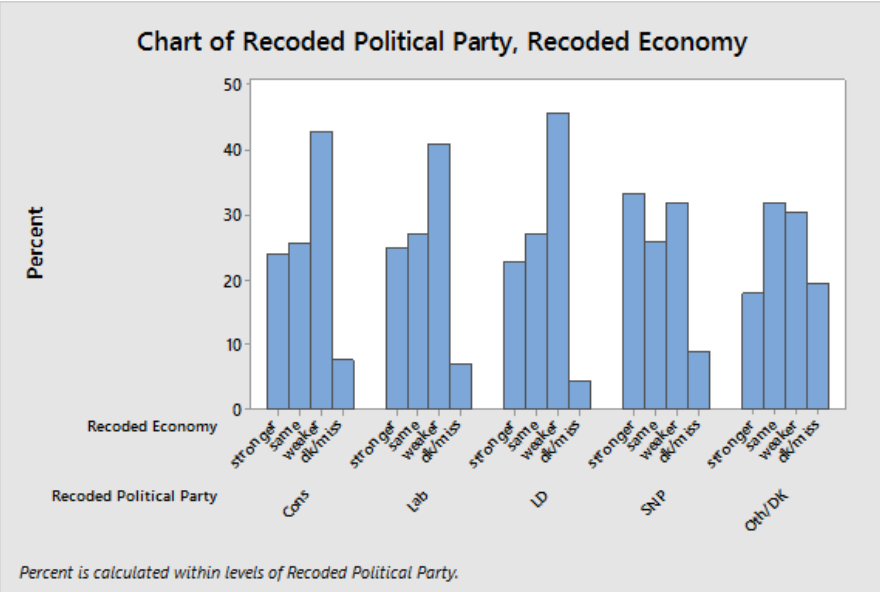
Graph > Bar Chart > Select Cluster > OK. Then select the two variables in the clustered bar chart

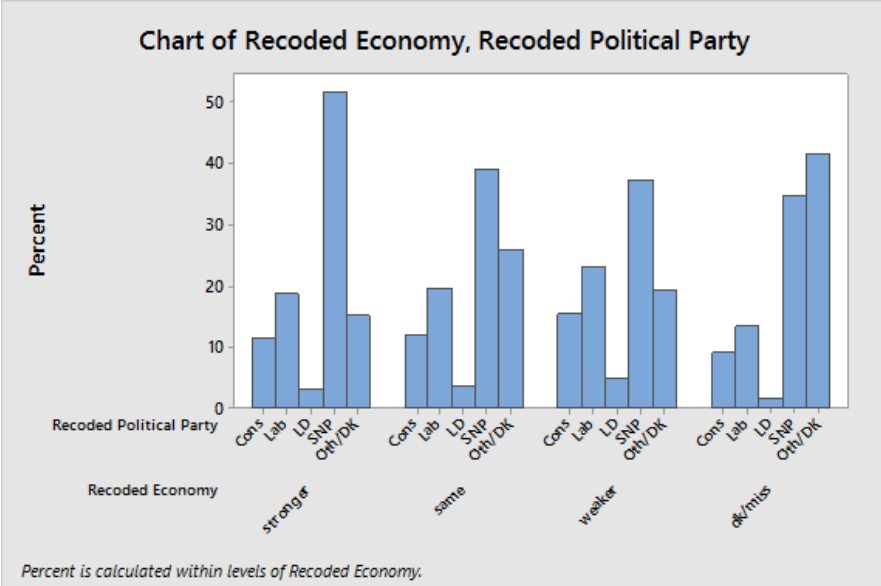


Two Way Clustered Bar Chart cont...

Click *Chart Options* to get the correct percentages. Select *Show Y as Percent*. Click OK

Examples of Clustered Bar Charts





Tips on Creating Effective Figures

1. Avoid cluttered charts and graphs.
2. Label everything to avoid ambiguity i.e. label axis with units (if appropriate).
3. Some of you may be tempted to use the stacked bar chart option as a chart; we wouldn't usually recommend this as it can be quite difficult to interpret!
4. Remember when visualising data, you want to make the reader/viewer's reading experience as easy as possible.

Problems with Pie Charts

Some of you may be tempted to use pie charts as well; generally we **do not use** pie charts as these can be quite deceiving when representing data!

- They force us to compare areas (or angles), which is pretty hard.
- People can manipulate these charts pretty easily in order to put a point across.