Open Polytechnic
KURATINI TUWHERA

72160 Statistical Analysis

Suggested data for projects

For Assignments 2 and 3 you may use data which has already been collected. However, it must be raw data, as opposed to data which has already been grouped into classes. If you do not have access to anything suitable here are some suggestions.

Websites

1. **SURF for schools**

   http://www.stats.govt.nz/methods_and_services/schools_corner/SURF%20for%20schools.aspx

   Although this is not strictly authentic data, it is based on findings from real data. Again, you need original questions and must not copy words from the site. Use one of these datasets:

   - *NZ Income Survey CART SURF* (2011 data). Use the sub-SURF sample of 500 individuals. Actual ages not given.
   - *NZ Income Survey Super SURF* (2003 data). Use the sub-SURF1 sample of 11,315. Actual ages not given.
   - *New Zealand Income Survey* (2003 data). Ages are given but sample only includes people to 45 years. Sample of 200.
   - *Household Savings Survey* (2001 data). Sample of 300.

2. **Census at school**

   http://www.censusatschool.org.nz/

   You can select a random sample from this site - you need to devise some questions of interest to you. Remember that as the children self-report there may be some unreliable data. You are expected to remove any data items which are obviously incorrect (for instance, a reaction time of 0 secs or 40 secs).

   Apart from the census data, there are another couple of datasets you can sample from on this site:

   a) http://new.censusatschool.org.nz/resource/kiwi-kapers-1/
   b) http://new.censusatschool.org.nz/resource/nz-incomes-surf/

   The NZ Incomes dataset is similar to the SURF data above and is 2003 data. Note that age is not a numeric variable here, as only age-groups are given. You can

sample from either of these datasets by going to http://www.censusatschool.org.nz/ and clicking on *Explore the data* then *Get a random sample*.

3. **Australian Institute of Sport data**
   http://www.statsci.org/data/oz/ais.html
   To copy data from the Statsci site above, we recommend you copy and paste to Excel then use *Data > Text to columns*. Click on *Delimited*, *Next*, *Space*. This should put each column of data in its own Excel column, instead of being separated by a space.

4. **NCEA Resource (NZ crash statistics)**
   http://ncea.tki.org.nz/Resources-for-aligned-standards/Mathematics-and-statistics/Level-3-Mathematics-and-statistics
   Scroll down and click on the resources for 3.10B. The Word document tells you about the sampling method and the variables – DO NOT cut and paste – use your own words and only describe the variables you plan to use.

Other sources

There's plenty of data about cars for sale, houses for sale, marathon times etc. on the web, which you can take a sample from. Data of times which is measured in hours and minutes will need to be converted to decimal numbers. You will need to frame questions to suit the sort of analysis we have covered.

There is plenty of data around the home:

- How long are the tracks on CDs?
- Has the size of novels increased in recent years?
- Does the running time of movies differ between comedies and dramas (see Sky magazine)?
- Does TV3 play more ads per hour than TV1?
- What is the sugar level in various cereals?

What your data must look like

Remember that you need to treat the data you found as your population. Then you need to take a random sample from this population using an appropriate sampling method. As mentioned elsewhere, time series data is **not** suitable to use.

You will need about 100 cases (people/objects) in your sample and a dataset with 3 or 4 variables (at least one numeric) is good. We prefer you to use data which is relevant to you. Although NZ data is best, overseas data could be used if it has some relevance to NZ.

If the data is somewhat dated, as in the NZ Incomes data in (1) or (2) above, you must note that and discuss the changes which may have occurred since the data was collected.

You may need to make some assumptions about how the data was collected (just saying "*I don't know how the data was collected*" is not satisfactory).

