THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

DATA to insight
An Introduction to Data Analysis

Datasets for Data to Insight

Chris Wild, University of Auckland

"Garbage in, Garbage out" is a standard catch-phrase in information-technology and it certainly applies to statistical data analysis today.

Useful data means more than just a set of answers from a questionnaire. Before we spend time analysing data in the real world, we need to answer these questions about its relevance and credibility:

- Who collected the data, where did they collect it and why?
- How did they measure or categorise the various factors the data addresses?
- How were the people (or other entities) selected?
- What did they do when people refused to answer?
- What did they do about missing values?
- If questions were asked over a period of years, how did they treat the data on people they lost contact with or who died?
- What allowances were made for periods when data was unable to be collected?

Our datasets

The goal of data analysis is to make connections between patterns in the data and what's happening in the real world. That is how insights and discoveries are made. But you can only make these connections if you have a background knowledge about what the data conveys and the particular types of information gathered (we call these variables). That's why on this course we restrict ourselves to data that we think you'll already know something about.

The main datasets for this course are NHANES and Gapminder. We'll use other datasets only when we want to show patterns and behaviours that are not well illustrated within these two.

NHANES

The National Health and Nutrition Examination Surveys (NHANES) represent the general state of health of the American population. The data has been gathered by the American National Center for Health Statistics since 1994 and it is freely available on the internet. They are the world's most analysed health surveys - Google Scholar indexes 67,000 research papers with NHANES in their abstracts.

NHANES gathers data on a huge number of factors, though we have prepared a master dataset for you from the 2009 - 2012 surveys of 10,000 observations on about 70 variables. We will also use small subsets of the data to enable us to explore questions about health, economic circumstances and behaviour using data on individual people.

Gapminder

The Gapminder international comparisons data is sourced from very reputable sources, like the World Bank. Gapminder is a non-profit foundation based in Stockholm. It aims "to replace devastating myths (about country-level demographics) with a fact-based worldview", by making data readily available and providing tools to make it understandable. The founder of Gapminder was the late, great Hans Rosling (see the next Step, 1.8) of the Karolinska Institute in Sweden. The dataset enables us to explore questions about health and socio-economic circumstances in several countries, and visualise changes in these demographics over time.

For your reference

We have also provided overviews on the **NHANES** and **Gapminder** datasets. If you are interested to learn more about the data, and how it relates to the exercises in the coming weeks, view the PDF documents below and refer to them throughout the course.

- NHANES 2009-2012 dataset overview
- Gapminder dataset overview

Links to all data sets used in the course

- in csv format
- in tab-delimited text format. Having the data in tab-delimited text format should make it easier for people in Europe or with European settings on their computers.

[You do not need to download these data sets. They are built into both iNZight and iNZight Lite. The links make the data sets available for download if you want to do that at some future time.]