Topic 1: Defining and Collecting Data

Dr. Lei Pan

School of Accounting, Economics and Finance
Curtin University

What is Business Data Analytics?

- Business and economic statistics is about getting the best information from data.
- When you have good information, you can use it to make good decisions.
- The 'business' part means
  - data is primarily economic or business data
  - decision is primarily an economic or business decision

How business data can be used? (1 of 2)

Consider some decisions that business and government may need to make:

Business decisions

e.g. A bank is deciding whether to give a home loan to a prospective customer. What is the risk this customer will not be able to afford to repay the loan? What is the risk that house prices may fall significantly leading to the home being worth less than the mortgage?

e.g. A company is thinking of setting up a new production plant. They need to understand what the demand for their product is likely to be out into the future. How would demand would be forecast?

Economic policy decisions

e.g. The government is thinking of reducing the income tax rate. What will be the impact on the economy? Will workers tend to work longer hours, as they get to keep more of their wages, or will this effect be very small?

How do we answer these questions? (1 of 2)

Economists and policy makers need data:

- They need **evidence** that can help them understand
  - the ways things are
  - weigh up options
  - identify problems
  - look at consequences
  - evaluate of performance
- It is NOT enough to go on intuition, or on what is commonly accepted wisdom.
  - often this can be wrong or misleading

How do we answer these questions? (2 of 2)

Many important issues and decisions revolve around the magnitude of an effect rather than whether it exists or otherwise.

e.g. The issue of climate change is a significant one. While most people accept climate change and humanity's role in it, a key question is the extent to which the climate is likely to change and the adaptations we will have to make to it. Estimating the magnitude of expected temperature rise, $+1°$C, $+2°$C or more matters a great deal.

e.g. As a retailer I may be quite sure that cutting the price of a product will generate greater sales of that product. However, the question of the magnitude of this change – whether sales increase by 1% or 100% – is pivotal to my decision whether to cut the price or not.

Applying statistics to business decision making

1. **Define** the data that you want to study in order to solve a problem or meet an objective.
2. **Collect** the data from appropriate sources.
3. **Organise** the data collected by developing tables.
4. **Visualise** the data collected by developing charts.
5. **Analyse** the data to reach conclusions and present those results.

What can you expect from taking this unit?

- Learn how to present, summarise and visualise data in a meaningful way.
- Learn how to use data to make informed decisions.
- Develop critical and analytical thinking about data and decisions made from it.

Data is everywhere (1 of 3)

- Data analysis skills are essential as the proliferation of data increases.
- The following figure shows
  - some indexes of terms related to the study of business and economic statistics such as ‘big data’, ‘machine learning’ and ‘data analyst’
  - these have become more popular relative to more established terms such as ‘economist’

Data is everywhere (2 of 3)

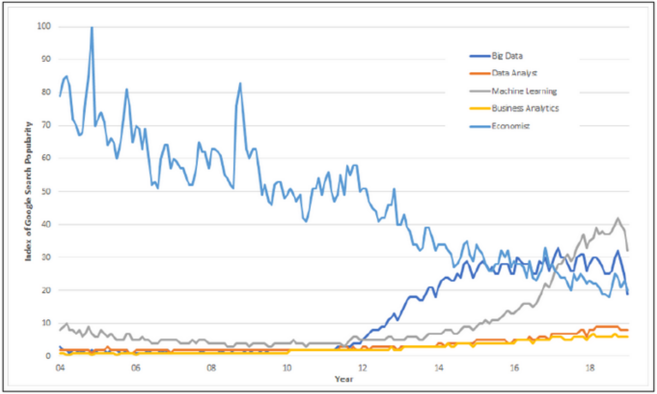
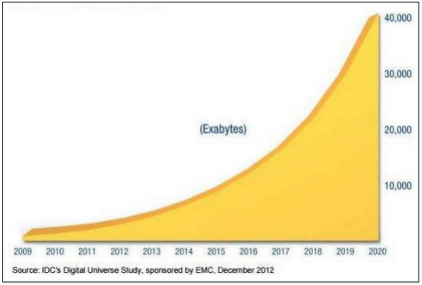


Figure: Google Index of Search Term Popularity (Worldwide)

Data is everywhere (3 of 3)

The figure below illustrates the explosion in data in recent years. In case you do not know what 'Exabyte' is then the table below might help.



The Digital Universe – 50-fold Growth from 2010 to 2020

1 Bit	=	Binary Digit
8 Bits	=	1 Byte
1000 Bytes	=	1 Kilobyte
1000 Kilobytes	=	1 Megabyte
1000 Megabytes	=	1 Gigabyte
1000 Gigabytes	=	1 Terabyte
1000 Terabytes	=	1 Petabyte
1000 Petabytes	=	1 Exabyte
1000 Exabytes	=	1 Zettabyte
1000 Zettabytes	=	1 Yottabyte
1000 Yottabytes	=	1 Brontobyte
1000 Brontobytes	=	1 Geopbyte

Table: What is an Exabyte

Curtin University

The Scientific Method

The scientific method is an iterative process. In general, we **reject a theory** if the data were *unlikely* to occur if the theory were in fact *true*.

Tools of Business Statistics

- Descriptive statistics
  - collect, describe, summarise, visualise and present data
  - examples are surveys, tables and graphs
- Inferential statistics
  - use *sample* data to make generalizations about a larger data set (*population*)
  - examples are estimation and hypothesis testing

Basic Concepts of Statistics

- **Variables** are characteristics of items of individuals.
- **Data** are the observed values of variables.
- **Population** is the set of all items or individuals of interest that you want to draw a conclusion.
- **Sample** is a subset of the population under study.
- **Parameter** is a numerical measure that describes a characteristic of a *population*.
- **Statistic** is a numerical measure that describes a characteristic of a *sample*.

Types of Data Analytics

- **Descriptive Analytics**: describing characteristics of a set of data.
- **Diagnostic Analytics**: discovering patterns or correlations in data.
- **Predictive Analytics**: building models that allow us to predict outcomes.
- **Prescriptive Analytics**: using models to design actions that will influence outcomes.

Types of Data (1 of 3)

- **Numercial or Quantitative Data**
  - This is data that takes a numeric value (e.g. income, age, GDP).
  - It can be of two types:
    1. **Discrete**

       e.g. number of people in a household, number of children in school, etc.
    2. **Continuous**

       e.g. your height, unemployment rate, etc.
- **Categorical or Qualitative Data**
  - This is data that does NOT take on numeric values.
  - But can be classified into distinct categories.

    e.g. country of birth, gender, day of the week, etc.

Types of Data (2 of 3)

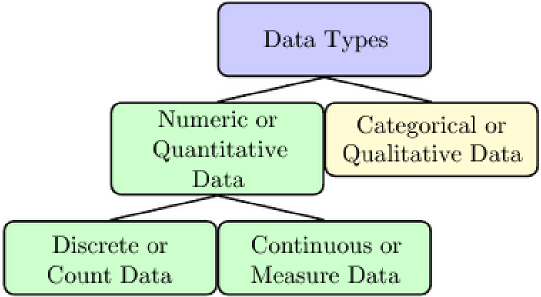


Figure: Data Types

Types of Data (3 of 3)

- When undertaking data analysis it is vital to understand the type of data you are dealing with and its properties.
- In many cases it may be necessary to make some transformation to the data in order to meaningfully interpret it.

You Do It

Identify whether the following variables are numerical or categorical. If numerical, state whether the variable is discrete or continuous.

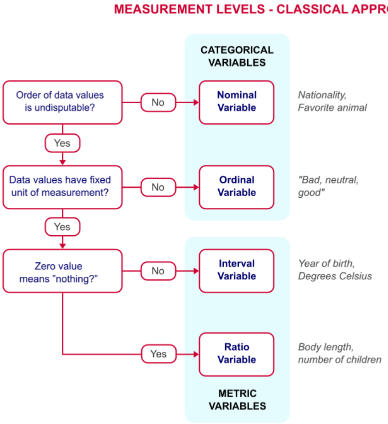
(a) your age

(b) income

Level of Measurement (1 of 2)

- The way a set of data is measured is called its *level of measurement*.
- Data can be classified into FOUR levels of measurement.
- They are (from lowest to highest level):
  - **Nominal scale**
    - ▷ categories (no ordering or direction)
    - ▷ examples: smartphone companies (Apple, Samsung, Sony)
  - **Ordinal scale**
    - ▷ ordered categories (ranking, order or scaling)
    - ▷ example: top five national parks in Australia
  - **Interval scale**
    - ▷ differences between measurements but NO true zero
    - ▷ example: daily temperature in Perth
  - **Ratio scale**
    - ▷ differences between measurements, true zero exists
    - ▷ example: height of 100 newborn babies

Level of Measurement (2 of 2)



Collecting Data (1 of 2)

- Important sources
  - data distributed by organisation or individual
  - designed experiment
  - survey
  - observational study
  - focus group
- Data sources can be classified as
  - primary sources
  - secondary sources

Collecting Data (2 of 2)

- **Structured data** follow some organising principle or plan, typically in a repeating pattern.
- **Unstructured data** follow no repeating pattern.
- **Outliers** and **missing values** may lead to the need for data cleaning.
- **Recorded variables** is a variable that has been assigned new values that replace the orginal values.

Normalization of Data

- An important issue in exploring data is to make sure that data are comparable.
- This may require scaling of the data in some particular way.

Example

A good example of this is carbon dioxide (CO_2) emissions.

Country	Total (kt)	Country	Per Capita (tons)
China	10,249,463	Qatar	37.78
United States	5,186,168	Trinidad and Tobago	34.52
India	2,034,752	Curacao	34.16
Russian Federation	1,789,074	Kuwait	27.22
Japan	1,243,384	Bahrain	24.29
Germany	757,313	Sint Maarten (Dutch part)	20.84
Iran, Islamic Rep.	616,976	Brunei Darussalam	19.19
Korea, Rep.	592,499	United Arab Emirates	18.78
Saudi Arabia	541,429	Luxembourg	18.7
Brazil	503,677	Saudi Arabia	18.08
Mexico	488,602	Oman	16.49
Indonesia	479,365	United States	16.4
Canada	475,735	Australia	16.35
South Africa	471,239	North America	16.11
United Kingdom	457,473	Kazakhstan	15.43
Australia	377,906	Estonia	15.11
Italy	344,768	New Caledonia	14.68
France	333,191	Gibraltar	14.52
Turkey	323,451	Mongolia	14.5
Thailand	303,118	Canada	13.53

Source: Climate change indicators from the World Bank available at: http://data.worldbank.org/indicator.

CO_2 emissins in 2013

Example (Cont'd)

- The first two columns are the total CO2 emissions for the 20 countries which have the highest emissions.
- China is at the top of the list followed by the U.S., India and so forth.
- Australia comes in at 16th place.

Example (Cont'd)

- While it may be of interest to examine total emissions, depending upon the precise question that is being addressed.
- It will often be more meaningful to normalise the data by population.
- This makes the quantity of emissions more comparable across countries.
- China has the largest emissions in main part because it has a very large population.
- However, each Chinese person's emissions are actually relatively low.

Example (Cont'd)

- The two right hand columns of the table are per capita CO2 emissions data.
- We can see that this tells quite a different story.
- China is no longer even on this list.
- Qatar heads the list along with some smaller countries – Trinidad and Tobago, Curacao and Kuwait.
- The U.S. and Australia rank highly on this table indicating that not only do they have high total emissions but their emissions per person are also high.

Ways to Normalize Data

- percentages
- proportions
- rates of return
- per person
- per dollar invested

Adjust for Inflation (1 of 8)

- A particular important way of scaling data is adjust for inflation.
- This is relevant for *nominal* data.
- That is data which is measured as a *monetary* value.
- We often need to correct for the effects of inflation.

Adjust for Inflation (2 of 8)

For example, suppose you were asked to choose between the following two options.

A. being gifted \$100 now

B. being gifted \$100 in 10 years time

Most people would choose the option A. Because prices tend to rise over time, \$100 will be worth less than it is now.

Adjust for Inflation (3 of 8)

- This is a really important observation because a lot of series are measured in dollar values.
- One of the most important is Gross Domestic Product (GDP).
- This measures the total value of all goods and services produced in the economy in a given period.

Adjust for Inflation (4 of 8)

The figure below shows Australia's GDP per capita.

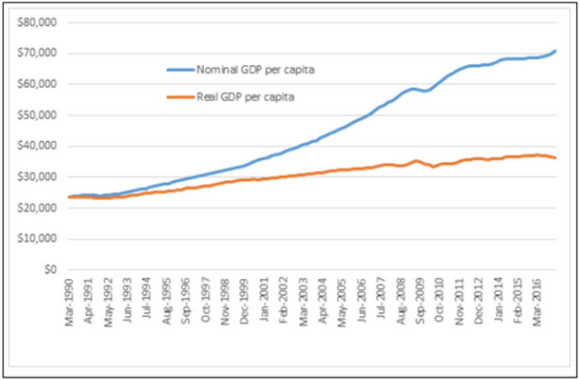


Figure: Australia's nominal and real GDP per capita

Adjust for Inflation (5 of 8)

- The figure shows a massive rise in nominal GDP per capita.
- It has risen from around $25,000 in 1990 to $70,000 in 2016.
- This is a rise of around 3 times.
- But does this mean we are 3 times wealthier than in 1990?
- No, because prices also rose over this period.

Adjust for Inflation (6 of 8)

- To correct for this we divide by a price index.
- The reason for doing this note that it is always the case that the total value of spending can be decomposed into a price and quantity component: Value $=$ Price $\times$ Quantity.
- What we are trying to do usually is examine how the *quantity* part changed.
- Thus we can calculate: Quantity $=$ Value/Price.

Adjust for Inflation (7 of 8)

- This is just what we would like to do for GDP. We will use the formula:

$$\text{Real GDP per capita} = \frac{\text{Nominal GDP per capita}}{\text{IPD}/100}$$

- IPD is the Implicit Price Deflator – this is a price index that is constructed especially to measure the change in price of all the goods and services produced in the economy.
- An alternative is the Consumer Price Index (CPI).
- This index is slightly narrower as it measures the change in consumption/consumer goods.
  - i.e. excludes prices of exports and imports, investment goods among others

Adjust for Inflation (8 of 8)

- Real GDP per capita has risen from $25,000 in 1990 to a bit more than $35,000 in 2016.
- This is a much less spectacular rise but more meaningful.
- It tells us:
  - much of the rise in Australia's nominal GDP reflected changes in prices rather than just increases in the quantity of output
  - our production of the quantity of goods and services per person has improved by around one third

Index Numbers (1 of 4)

- Let us also consider index numbers in some detail as these are often misinterpreted.
- Indexes are numeric and continuous data but measures relativities rather than level.

Index Numbers (2 of 4)

To see what this means consider the CPI data for Australia's capital cities in the table below.

Date	Sydney	Melbourne	Brisbane	Adelaide	Perth	Hobart	Canberra
March 2016	108.7	108.2	108.5	107.0	107.9	106.4	106.2
June 2016	109.3	108.6	109.0	107.5	108.2	106.4	106.4
September 2016	110.4	109.1	109.7	108.4	108.6	107.1	107.3
December 2016	110.9	109.9	110.2	108.7	109.0	108.0	107.9
March 2017	111.3	110.9	110.5	109.1	109.0	108.9	108.6

CPI for Australia's state capital cities

Curtin University

YEARS 50 OF INNOVATION

Index Numbers (3 of 4)

- These indexes measure the price change for each city relative to the base period.
- The base period is 2011-12 when the index is set equal to 100.

Index Numbers (4 of 4)

A couple of things are worth noting.

1. The fact that these indexes are based on 100 is somewhat arbitrary.
    - This is a choice that the Australian Bureau of Statistics (ABS) has made.
    - They could just easily have chosen 1 or 1000.
    - In fact, in New Zealand the CPI base value is 1000.
2. The index values are only comparable for a given city across time, NOT for a given time period across cities.
    - correct interpretation: prices in Sydney have risen by 11.3% between March 2017 and the base period
    - incorrect interpretation: prices are higher in Melbourne than Canberra in March 2017 because Melbourne's index number is 110.9 compared with 108.6 for Canberra
3. Given this, a common way of presenting the CPI is to look at how it has changed over time.

Why Sample?

- A *census* is a sample of the entire population.
- A sample is less time-consuming than a census.
- A sample is less costly to administer than a census.
- A sample is less cumbersome and more practical to administer than a census.

Good Data?

- A sampling method is *biased* if it produces results that systematically differ from the truth about the population.
- *Selection bias*: when all observations do NOT have an equal chance of being selected or excluded from the survey
- *Non-response bias*: distortion that can arise because a large number of units selected for the sample do not respond
- *Sampling bias*: significant differences across different samples of the same size
- *Response bias*: distortion that arises because of the wording of a question or the behavior of the interviewer

You Do It

A simple random sample of $n = 300$ full-time employees is selected from a company list containing the names of all $N = 5,000$ full-time employees in order to evaluate job satisfaction.

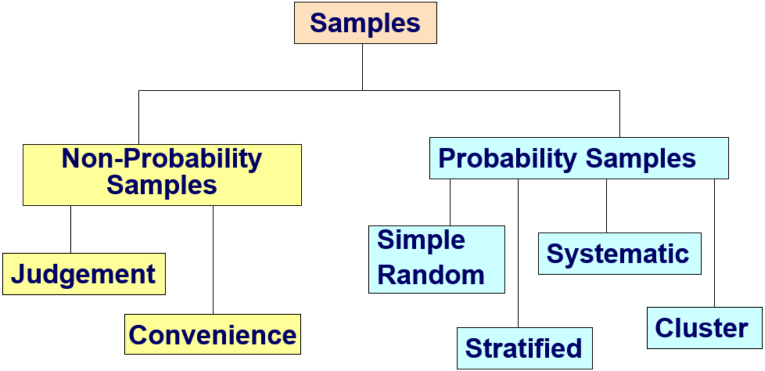
(a) Give an example of possible selection bias.

(b) Give an example of possible non-response bias.

(c) Give an example of possible sampling bias.

(d) Give an example of possible response bias.

Sampling Techniques

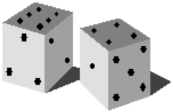


Statistical Sampling

- Items of the sample are chosen based on known or calculable probabilities.
- A sampling method that gives each unit in the population a known, non-zero chance of being selected is called a **probability sampling method**.

Simple Random Samples

- Every individual or item from the population has an equal chance of being selected.



Stratified Samples

- A **stratified random sample** is selected by:
  1. dividing the population into **mutually exclusive** subgroups
  2. then taking a simple random sample from each subgroup
  3. simple random samples are then combined to give the full sample
- It allows us to obtain information about each subgroup.
- It can be more efficient than simple random sampling.

Systematic Samples

- For a **1-in-k systematic sample**,
  1. order the units of the population in some way
  2. randomly select one of the first k units in the ordered list
  3. this selected unit is the first unit to be included in the sample
  4. continue through the list selecting every kth unit from then on
- It is *convenient*, *fast* but could be *biased*.

Cluster Samples

- In **cluster sampling**,
  1. units of the population are grouped into clusters
  2. one or more clusters are then selected at random
  3. if a cluster is selected, all units of that cluster are part of the sample