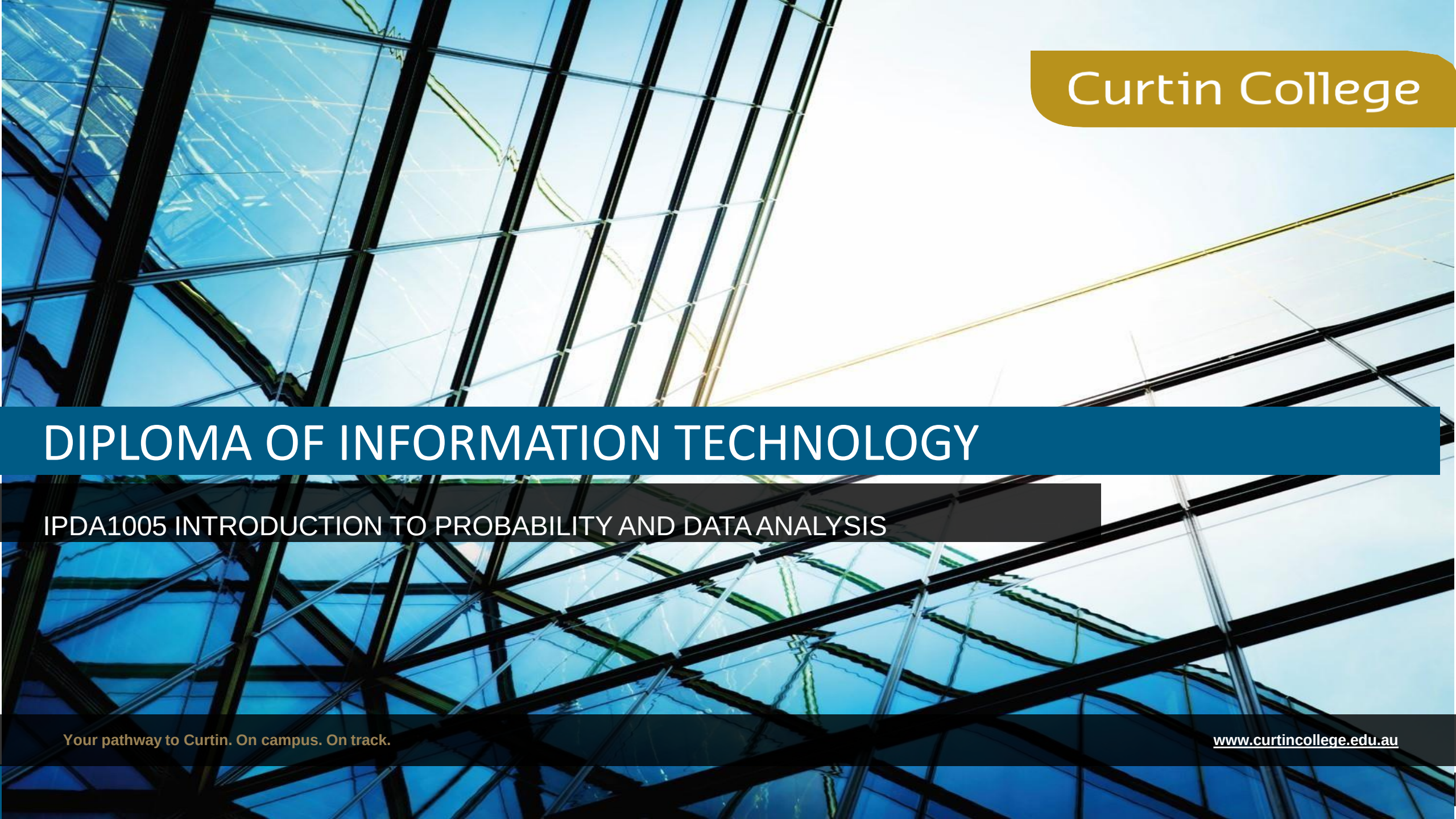


DIPLOMA OF INFORMATION TECHNOLOGY

IPDA1005 INTRODUCTION TO PROBABILITY AND DATA ANALYSIS

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

Outline

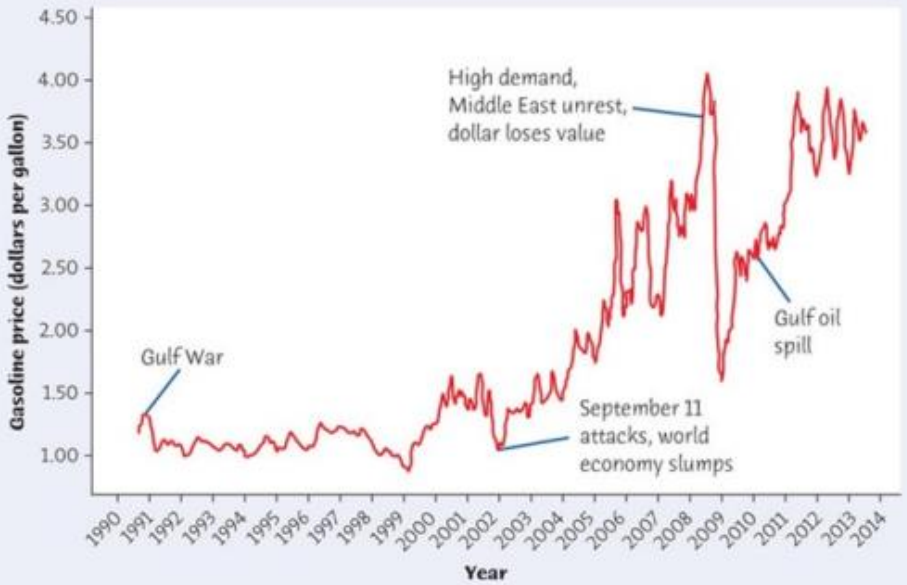
1. Statistical investigations
2. Populations, samples and sampling
3. Types of Variables
4. Summarizing Data
   - Numerical Summaries
   - Graphical Summaries
   - Describing and comparing distributions

A Broader View of Statistics

- Statistics is not just about finding the right method to test or analyze data.
- It is
  - the art and science of collecting, displaying, analyzing, and drawing appropriate conclusions from data
  - a way of rationally assessing claims
  - a framework for understanding and modelling variation and uncertainty.
- Data is everywhere, and so too is variation.
  - Transactional data – who's buying what and when
  - Your Facebook likes, dislikes, what you comment on, your listed interests, the stories or advertisements you click on, ...
  - Temperature measurements across Australia
  - Clinical trials on the effect of cholesterol-reducing medication

Variation is Everywhere

- Lots of things vary over time.
- Variation occurs over different scales, e.g., time.
  - Overall trend in petrol prices
  - Short-term variation
  - Unpredictable shocks
- Because of variability, our conclusions are always uncertain – what is the price of petrol going to be next year? Next week?



Study combined with ... (messaging, websurfing, tweeting

Computers in Human Behavior

Volume 53, December 2015, Pages 63–70

Make it our time: In class multitaskers have lower academic performance

Saraswathi Bellur[a], Kristine L. Nowak[a], Kyle S. Hull[b]

Show more

doi:10.1016/j.chb.2015.06.027

Get rights and content

Highlights

- Data highlight the prevalence of multitasking both within and outside classroom.
- In-class multitasking was found to be negatively predictive of current college GPA.
- Multitasking during homework increases time spent studying outside class.
- Texting emerged as a dominant multitasking activity within and outside classroom.
- Implications for technology use, practices and policies in academia are discussed.

Multitasking Increases Study Time, Lowers Grades

July 23, 2015 By: Colin Poitras Category: Nation & World

How Many Energy Drinks is Too Many?

Teenagers who consume energy drinks 'at risk of heart attack'

18:19, 2 APRIL 2013 BY MARK WAGHORN

Researchers said the drinks popular among 10 to 19-year-olds can trigger sudden heart attacks in young, apparently healthy people

Share | Share | Tweet | +1 | Pinterest

Enter your e-mail for our daily newsletter | Subscribe

Danger: Such drinks can be harmful drunk alone as well as with alcohol

★ Recommended In News

MENTAL HEALTH
Young man who saved stranger's life with three simple words is repaid the most special way

AIRBUS SA
Airbus unveils 'Son of Concorde' supersonic jet capable of travelling from London to New York in ONE HOUR

SPACE
Mystery surrounds bizarre Mars 'space crab' spotted in NASA photo of the Red Planet

FEEL-GOOD NEWS
Teenager's 'letter to heaven' balloon release at her father's grave turns up at family home 25 MILES away

BABIES
Posting photos of your kids on Instagram? People could be doing

---

ELSEVIER

Canadian Journal of Cardiology

Volume 31, Issue 5, May 2015, Pages 572–575

Viewpoint

Energy Drink Overconsumption in Adolescents: Implications for Arrhythmias and Other Cardiovascular Events

Fabian Sanchis-Gomar, PhD, MD[a], Helios Pareja-Galeano, PhD[a, b], Gianfranco Cervellin, MD[c], Giuseppe Lippi, MD[d], Conrad P. Earnest, PhD[e]

Show more

doi:10.1016/j.cjca.2014.12.019

Get rights and content

Energy drinks (EDs) have increased in popularity and are now consumed by 30%-50% of adolescents (ie, aged 10-19 years) and young adults.[1] It is now estimated that 31% of 12- to 19-year old adolescents regularly consume EDs.[2 and 3] Alcohol mixed with energy drinks has also become increasingly popular among adolescents and college students.[3 and 4] EDs mainly differ from other soft and sports drinks regarding their high caffeine content and their promotion as a means to relieve fatigue and improve physical and cognitive performance. The problem with ED consumption is that these beverages often contain high amounts of labeled and even masked caffeine, as well as other substances such as guarana, ginseng, and taurine in variable quantities, which may generate uncertain interactions.[5,6 and 7] Guarana (*Paullinia cupana* of the Sapindaceae family) is a Brazilian plant containing "guaranine," which is nothing more than caffeine, in

Different Types of Studies

- Main distinction is between
  - observational studies, including surveys
  - experimental studies.
- Observational study:
  - Observes individuals or individual experimental units and measures variables of interests, but does not attempt to influence responses.
- Experimental study:
  - Imposes controlled treatment on individuals or experimental units to measure responses.
  - Asks whether the treatment changes the response.
- When studying cause and effect, experiments are the only source of convincing data.

Population, Sample and Inference

- In a statistical study, the *population* is the entire group of individuals or experimental units about which we want information or about which we want to draw conclusions.
- A *sample* is the part of the population from which we actually collect information – observations.
  - How we collect or identify a sample is a really important, and strongly affects the nature and strength of the conclusions.
- *Statistical inference* is the process of using information from a sample to draw conclusions about the population.

What Australians Really Think

- 1200 individuals were asked questions related to foreign policy, e.g., "Are you personally in favour or against this reduction in the budget to Australia's overseas aid?"
- Categories: Strongly in favour, Somewhat in favour, Somewhat against, Strongly against, Neither/don't know/no view.

- What is the population? What is the sample? How are they related?

Bad Sampling

We want the sample to be representative of the entire population. Sampling is *biased* if it systematically favours some part(s) of the population.

- Standing outside a bar as patrons are leaving and surveying them about their opinions about alcohol consumption.
- Facebook polls
  - A *voluntary response* sample has an element of self-selection. Such samples are biased towards those whose strong opinions lead them to respond. Also, the web site may use an algorithm that targets users' interests.
- How might we take a sample of iron ore from a large stockpile to determine the iron ore content?
  - A sample consisting of only very large rocks at the base of the stockpile will likely be biased.

How Do You Sample This?

Iron Ore Plunges Most Since August 2009

The idea is to sample the particulate stream before the stockpile is constructed. See https://www.youtube.com/watch?v=vCa9Ui2oSyQ

Observing the entire population

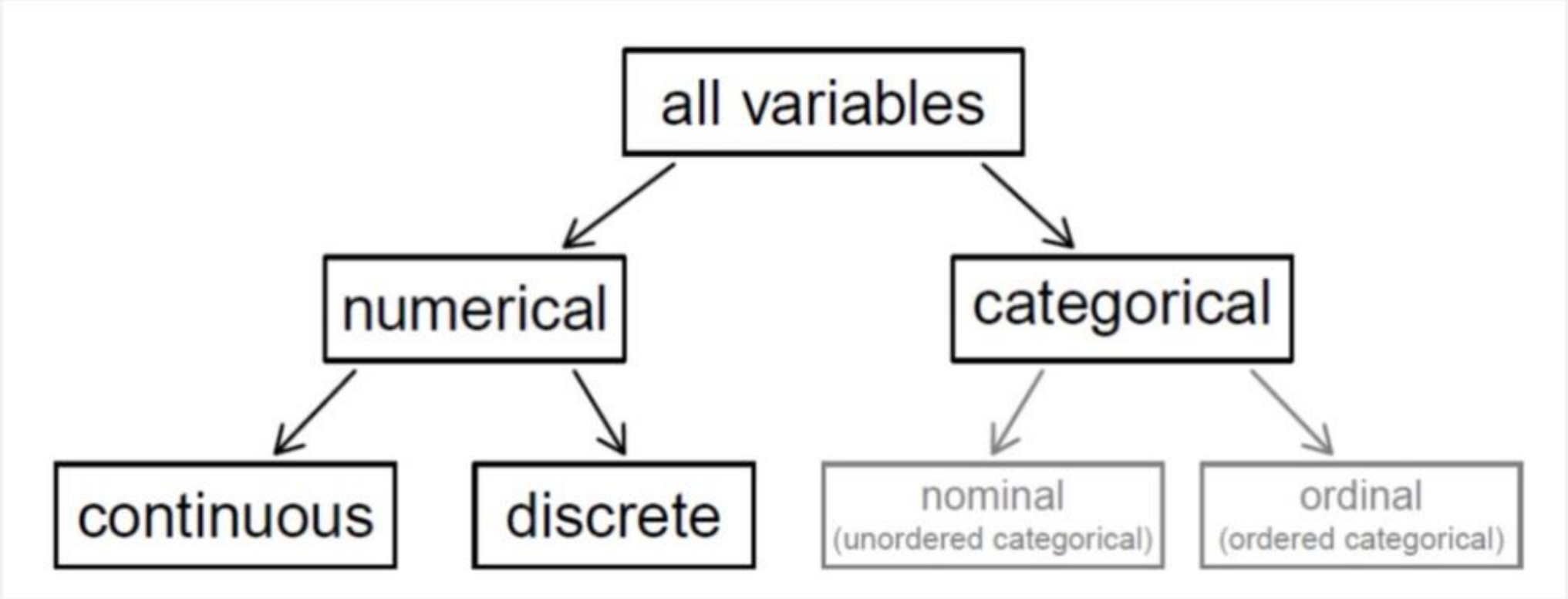
- When do we observe the entire population?
- That is called a census.

Simple Random Sampling - the GOLD standard

Definition:
A Simple Random Sample (SRS) is chosen in such a way that every possible set of individuals of the same size is equally likely to be selected.

- Equivalent definition: every individual in the population has the same chance of being included in the sample.
- Why use random sampling?
  - By definition, it avoids bias in selecting samples.
  - It creates independence between observations, so probability models can be applied.
  - Results from random samples have *calculable* margins of error that let us work with the unavoidable uncertainty.
  - If sampling is non-random, bias is likely, and uncertainty cannot be quantified.
- **Note:** SRS does not *guarantee* good representation, since weird stuff can happen just by chance.

Types of Variables



From Diez, D.M., Barr, C.D., and Cetinkaya-Rundel, M. (2015) OpenIntro Statistics, 3rd ed.

Types of Variables

- Quantitative or numerical variables have numerical values that indicate some characteristic of each unit.
  1. Continuous: Hourly temperature at Perth Airport, wind speed, height and weight of people (measurements, asking "how much?")
  2. Discrete: Number of tropical cyclones making landfall in North West Australia per cyclone season, score in a basketball game, number of failures in a repeated success/failure experiment (counts, asking "how many?")
- Categorical variables place units into a category.
  1. Nominal (unordered categories): State of residence – WA, SA, NSW, ...; colour of flowers; names of people.
  2. Ordinal (ordered categories): Strongly agree, Somewhat agree, Neutral,...

An Example with Different Types of Variables

- Suppose we surveyed students in a statistics unit.
- Four variables were recorded for each student:
  1. Number of units begin taken at the same time
  2. ATAR score
  3. Whether the student had previously taken a statistics unit
  4. attitude to maths selected from
     {Love it, OK (I guess), Necessary evil, What's maths?}
- Give the type of each variable:
  1. Number of units: numerical discrete
  2. ATAR score: numerical continuous
  3. Previous stats: categorical nominal
  4. Attitude to maths: categorical ordinal

Exploratory data analysis - Summarizing Data

- In *exploratory data analysis* we summarise data both graphically and numerically.
- We look for overall patterns, trends, unusual observations, what might be "happening".
- "Exploratory", because it doesn't lead to formal conclusions. Rather, it indicates questions that we might want to ask about the data.
- Numerical summaries: Descriptive statistics such as mean, median, standard deviation, range, etc.
- Graphical summaries: Histograms, scatterplots, boxplots, etc.
- We can use these to describe and compare data.

Numerical Summaries

A simple numerical summary of a sample data may include

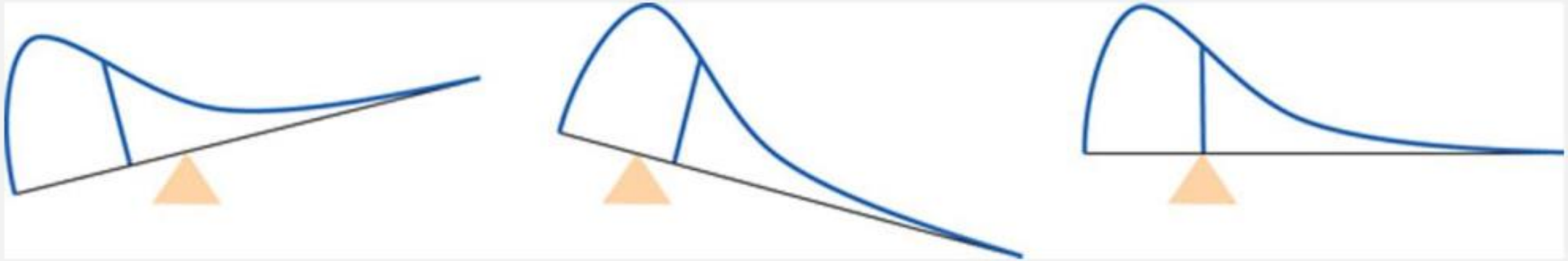
- Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right)$
- Variance: $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

- Standard deviation: $s = \sqrt{s^2}$
- Median
- Minimum and maximum
- Quartiles 1 and 3, and Interquartile Range (IQR)

The Mean as the Centre of Gravity

- The mean of a density curve is the balance point, at which the curve would balance if made of solid material.



Example of Numerical Summaries

Numerical summaries in *R*:

```
> x <- c(44.0, 79.6, 70.6, 82.7, 61.8, 50.7, 47.1, 60.9,
 70.5, 67.9, 47.2, 62.3, 70.1, 58.5, 43.7)
> mean(x)
[1] 61.17333
> sd(x)
[1] 12.58074
> summary(x)
   Min.  1st Qu.  Median    Mean   3rd Qu.    Max.
  43.70   48.95   61.80    61.17   70.30     82.70
> IQR(x)
[1] 21.35
```

The function `describe()` in the package `psych` provides another compact summary.

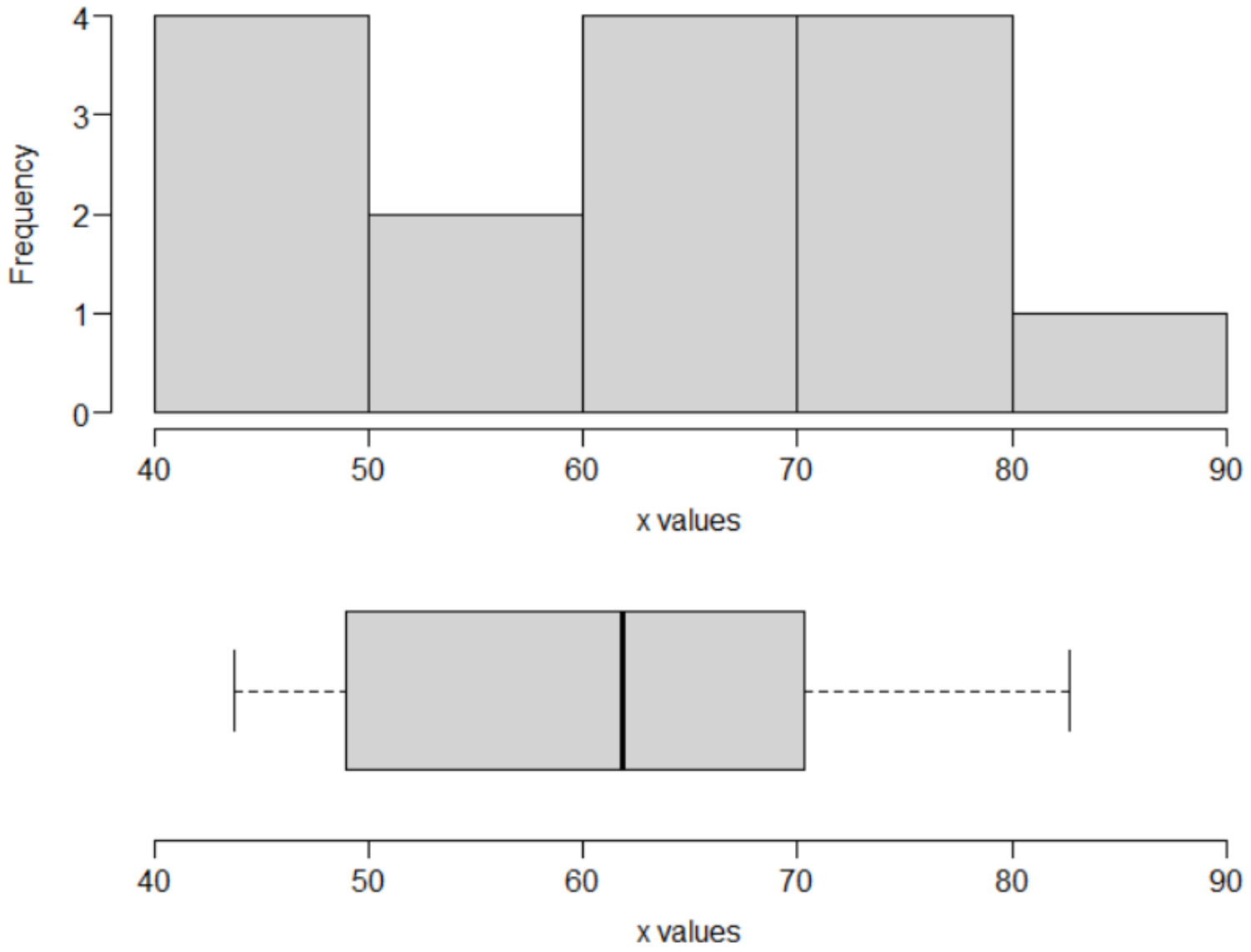
Calculation of Order Statistics

- Order statistics (min, quartiles, median, max) are based on a sorted (ordered) list of the data.
- In the sorted list, the median is the middle value.
- The first quartile (Q_1) is the median of observations below the median. Q_3 is the median of observations above the median.

  $43.7 \;\; 44.0 \;\; 47.1 \;\; 47.2 \;\; 50.7 \;\; 58.5 \;\; 60.9 \;\; 61.8 \;\; 62.3 \;\; 67.9 \;\; 70.1$
  $70.5 \;\; 70.6 \;\; 79.6 \;\; 82.7$

- *R* computes Q_1 and Q_3 in a more sophisticated way (which you don't need to worry about).
- The interquartile range (IQR) is given by $Q_3 - Q_1 = 70.5 - 47.2 = 23.3$.
- The "5-number summary" comprising Min, Q_1, Median, Q_3, Max is visually represented in a boxplot.

Histogram and Boxplot



Outliers and Boxplots

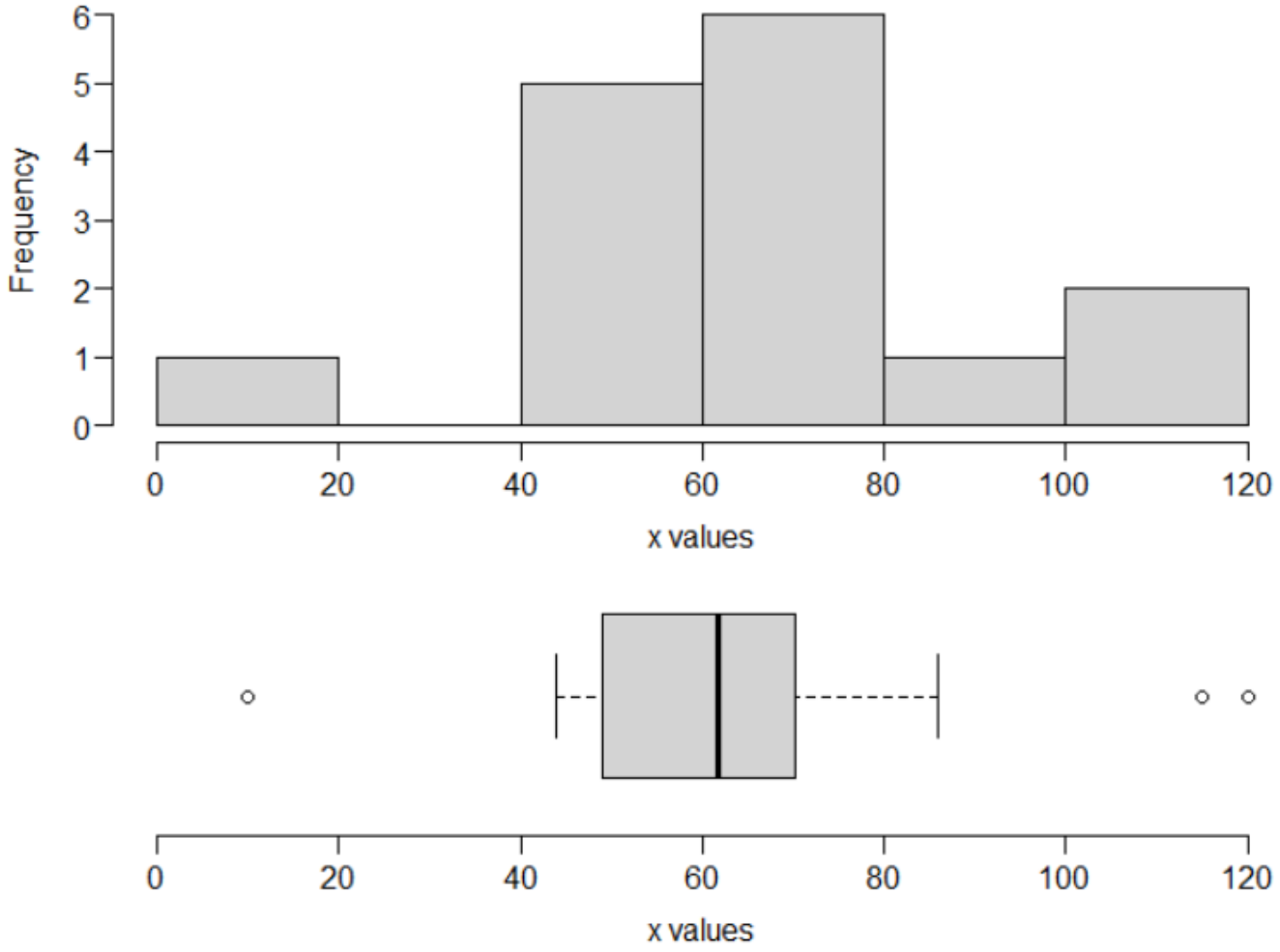
- An empirical convention for detecting outliers was developed by John Tukey (1915–2000).
- Observations that lie outside the following range are considered potential outliers:

  $$(Q_1 - 1.5 \times \text{IQR}, \quad Q_3 + 1.5 \times \text{IQR})$$

- When such observations exist, the whiskers are no longer extended to the largest and smallest observations. Instead they are extended to the largest and smallest observations *within* the above interval.
- The outlying observations are marked by asterisks or little circles.

Example of Boxplot with Outliers

The previous data is modified by decreasing the lowest value, and increasing the two highest values.



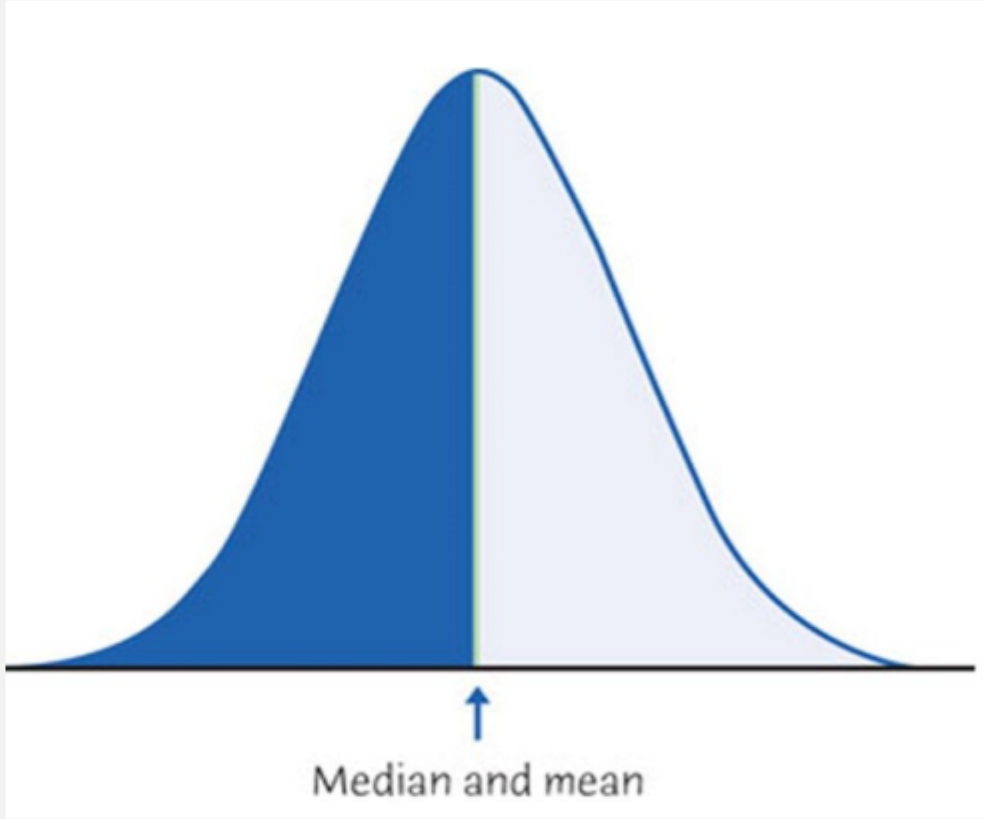
Describing distributions and data

Three main distribution concepts:

- Location
  - Also called 'central tendency'
  - measured by mean, median, mid-range
- Spread
  - Also called 'dispersion', 'variation', 'variability'
  - measured by standard deviation, variance, interquartile range (IQR), range
- Shape
  - Has numerical measures called skewness and kurtosis that we won't use much.
  - Better to stick to pictorial concepts and graphs (more to come...)
- Anything else that is interesting
  - Maybe outliers (if not mentioned already)

Distribution shapes - symmetric case

- The median of a distribution is the mid-point of the distribution, with half the values above the median and half the values below it. Similarly for the median of data.
- This helps in characterising a symmetric distribution (or symmetric data).
- The median and the mean are the same for a symmetric distribution/data. They both lie at the centre.



Distribution shapes - skewness

- A mean is sensitive to all numerical values, but a median is sensitive to just one or two values.
- In skewed distribution/data the mean is pulled away from the median in the direction of the long tail.
- The direction of the long tail is also the direction of the skewness: negative/left or positive/right.
- If distribution is right-skewed then mean $>$ median. If distribution is left-skewed then mean $<$ median. And vice-versa.