Southern African Association for Institutional Research

Foundations of Institutional Research 2015

Introduction to statistics for institutional research

Presented by Dion van Zyl

1

Purpose of this session

The purpose of this session is to **offer a perspective on the use and value of statistics within the context of institutional research**

Focus is not to cover statistics comprehensively or to provide a crash course in statistics but rather to provide a basic introduction to statistics, in particular in the context of Institutional Research

A basic understanding of numbers are assumed

Favouring more traditional research approaches

Content

- The role of statistics in the research process
- What are stats used for?
- Data sources for IR
- The ‘science’ behind stats
- Key building blocks of (useful) statistics
- Conclusion

The role of statistics in the research process

Institutional researchers have a critical role to play in the provision of appropriate information to support decision-making and planning

SOUTHERN AFRICAN ASSOCIATION FOR INSTITUTIONAL RESEARCH

What are stats used for?

- Inform critical assumptions & institutional knowledge
- To see how we are doing
- To support decision-making (management & academic)

Evidence based…but with this comes great responsibility

SOUTHERN AFRICAN ASSOCIATION FOR INSTITUTIONAL RESEARCH

Some data sources for IR

- Data sourced from the institutional database
  - Student data
  - Staff data
  - Financial data
  - Asset data
  - Traditional research data
  - Learning analytics & big data
- Data sourced from external databases
  - Data about potential students
  - Data about the Higher Education system
- Data sourced from surveys

Southern African Association for Institutional Research

The ‘science’ behind stats

The (scientific) research process

SOUTHERN AFRICAN ASSOCIATION FOR INSTITUTIONAL RESEARCH

Scientific

- Latin: ‘scientificus’ -> **producing knowledge**
- Characterised by the methods we employ
- Synonyms: systematic, methodical, organised, well organised, rigorous, precise, accurate
- Knowledge contribution and the processes we followed

Research

A systematic process of collecting, analysing, and interpreting information [data] to increase understanding of a phenomenon about which we are interested

(Leedy & Ormrod, 2010)

Statistics

Provide a way of dealing with numbers

Statistics aims to **explain variation in the data** (similarities, differences & relationships)

Statistics is NOT a shopping list!

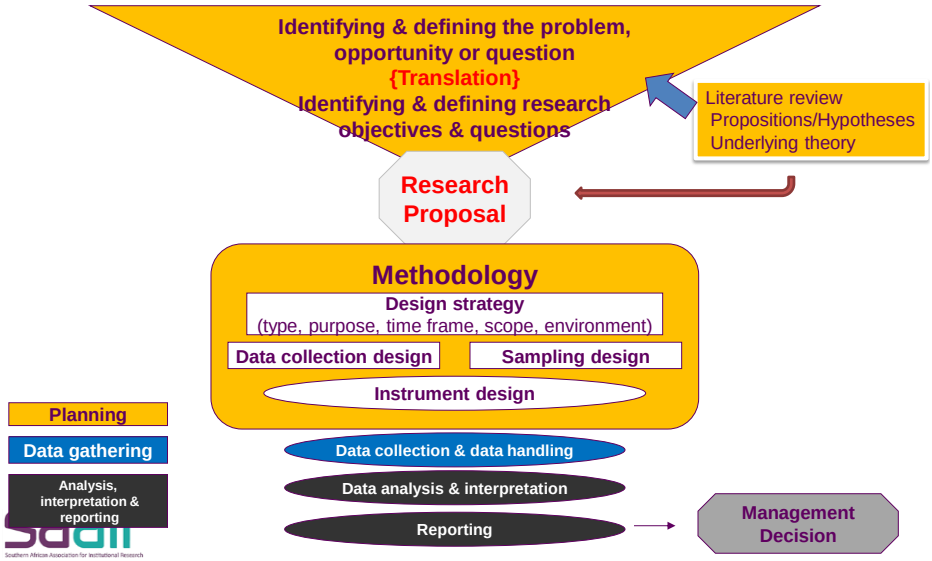
Key building blocks – where it all starts

- It all starts with the **research problem** (translating into research questions, objectives & hypotheses)
- Choosing the most **relevant research design** to address the research questions
- **Measurement** (and the scales we use)
- **Questionnaire design**
- **Sampling**

Key building blocks – where it all starts

- **Data collection & fieldwork**
- **Data handling**
- **Data analysis (variation & model building)**
- **Moving from results to findings to interventions**

The research problem



Avoid setting yourself up for failure…

What if we ask the wrong questions to formulate the problem statement?

Symptoms

Problem

This will impact on what to measure and the statistics we produce

saair
Southern African Association for Institutional Research

Measurement

Name the tallest mountain in the world

Kilimanjaro

Everest

Drakensberg

Mauna Kea

saair
Southern African Association for Institutional Research

What it takes to ‘measure’

Conceptualisation vs Operationalisation

Scales (and the assumptions we make)

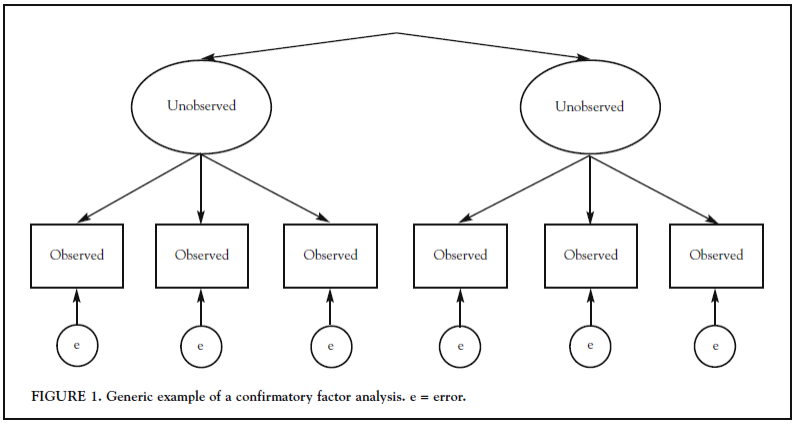


FIGURE 1. Generic example of a confirmatory factor analysis. e = error.

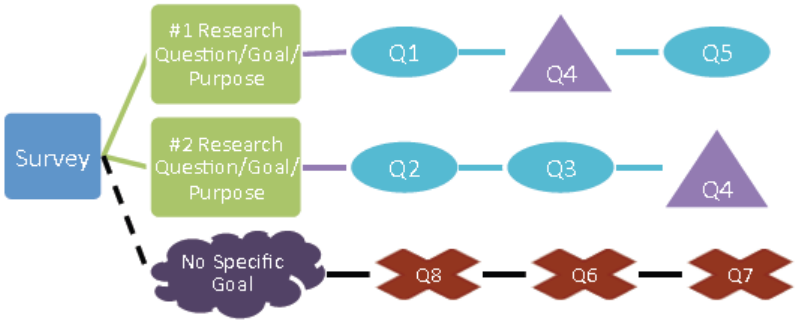
Schreiber (2006)

Questionnaire design

- In preparing a questionnaire, you must **consider how you intend to use the information**; you must **know what statistics you intend to use**.
- Depending on the statistical technique you have in mind, you may need to **ask the question in a particular way**, or provide different response formats.

Leedy & Ormrod (2009)

Survey scaffolding



Practical problems I have seen

- Favouring certain measurements!!
- Operational definition does not correspond to the conceptual definition
- Trying to cover too much with your measurements
- Doing all the correct measurements but presenting the wrong statistics
- Doing all the correct measurements and presenting the correct statistics, but drawing the wrong conclusions

COMMON FAULTS IN QUESTIONNAIRE DESIGN

Leading questions
'It is well know that most people want free education to be made compulsory. How do you feel about it?'

Threatening questions
'Have you ever participated in student riots?'

Double-barrelled questions
'Rate the following: Professionalism and level of service of the academic staff'

- Questions must be definitive
  - E.g. Interpretation of ‘small increase’ in fuel prices (1%, 5%, 10%, 50%)
- Questions must be simple
  - Even those less educated should understand them
  - Avoid jargons!
- Avoid prestige-loaded questions
- Avoid questions that may embarrass the respondent
  - E.g. Have you ever cheated in your exam?

Tustin *et al.* (2005)

Faults on pre-coding

How old are you?

11−20 ☐
20−30 ☐
30−40 ☐

- Not including all possible responses.
- Someone aged 30 can tick two boxes.

Over-complex questions

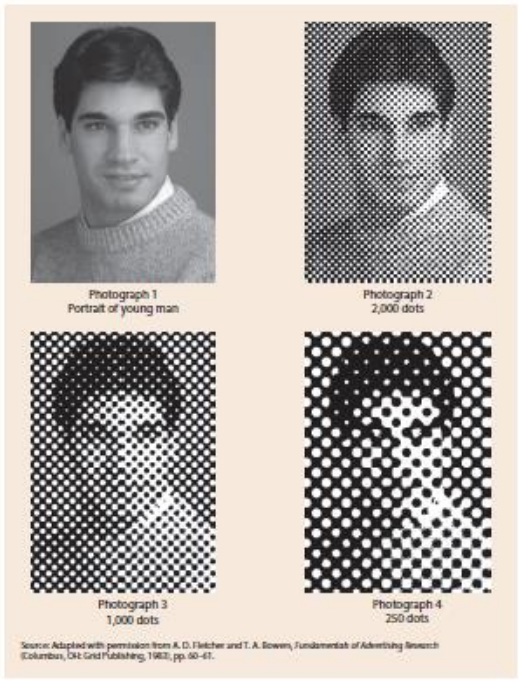
How well do you think Chelladurai’s model of leadership explains your own attitudes?

Validity & reliability

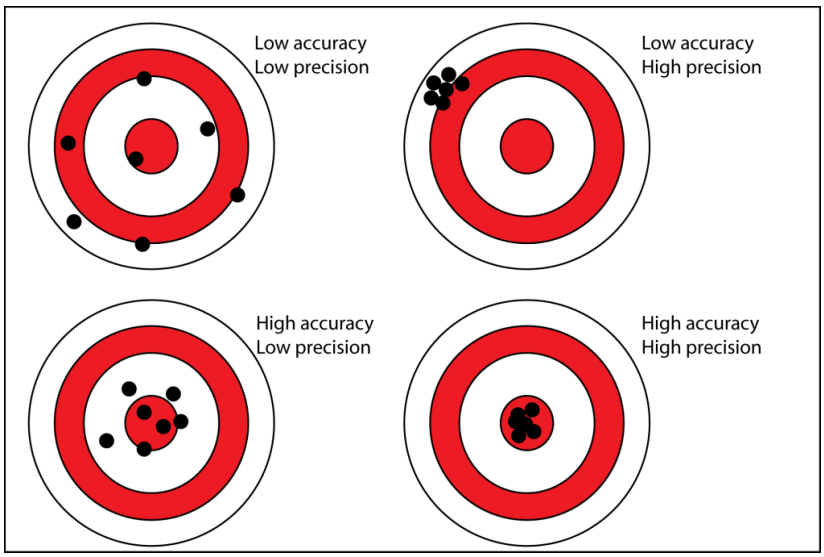
- **Validity** = Extent to which the measurment accurately measures what it were intended to measure. I.e. the extent to which research findings are really about what they profess to be about.
- **Reliability** = Extent to which data collection techniques will yield consistent findings, similar observations made or conclusions reached.

Sampling

- **Balance some element of risk/uncertainty against available subjects and/or resources**
- **How accurate and precise are our measurements in relation to our population?**



Accuracy
Precision
Representativeness
Bias/Error
Generalisation



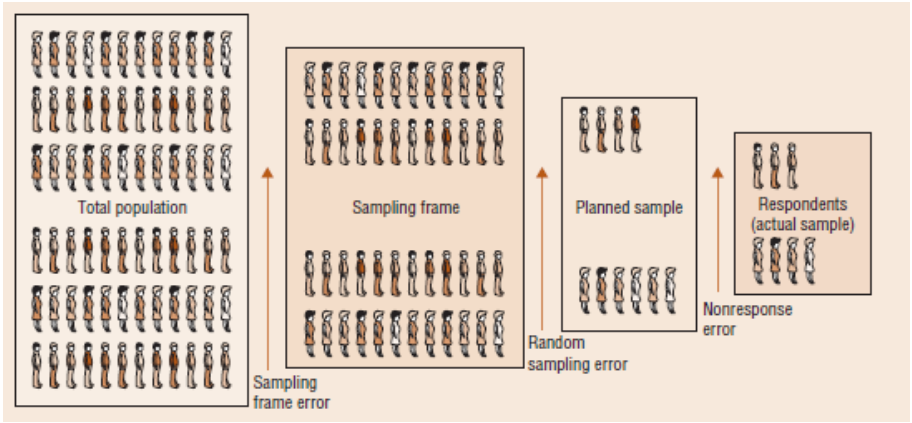
Some key areas of consideration

- Accurate definition of **population** (theoretical vs. target vs. observed)
- Identification of relevant **sampling units** (unit of analysis vs. unit of observation)
- Specification of **sampling frame**/description from which the sampling units are to be selected
- **Sample size** – Power of analysis
- **Sampling techniques**

Sampling error

- **Random**
  - Function of sample size
  - As sample size increases, random sampling error decreases
- **Systematic**
  - Result from non-sampling factors, primarily the nature of a study's design and the correctness of execution
  - Errors are not due to chance fluctuations

Sampling error



DATA PREPARATION

Data coding

Data base design

Data capturing

Checking the data

Dealing with missing values

Data transformation

Data analysis

Descriptive statistics:

- Describe large amounts of data in an summarised fashion by means of tables and figures
- Describe important characteristics of your data by means of descriptive measures

Inferential statistics:

- Apply probability theory
- Use sample data and descriptive measures to draw conclusions and make inferences about population

saair SOUTHERN AFRICAN ASSOCIATION FOR INSTITUTIONAL RESEARCH

Types of data

- **Categorical**
  - Data whose values cannot be measured numerically but can either be classified into sets (categories) or placed in rank order
- **Continuous**
  - Data whose values can theoretically take any value (sometimes within a restricted range) provided they can be measured with sufficient accuracy

saair

Data types

- **Categorical**
  - Nominal
  - Ordinal
- **Continuous**
  - Interval
  - Ratio

Categorical - Nominal

- E.g.
  - Gender: 1=Male; 2=Female
  - Province: 1=GT; 2=WC, 3=FS; etc.
  - Current working status: 1=Working; 2=Unemployed; 3= Student; 4=Housewife; etc.
- Measures data by assigning names to them
- No order in the data and cannot be ranked
- Coding facilitates measurement
- Simplistic

Continuous - Interval

- E.g.
  - Temperature in degrees Celcius/Fahrenheit
- Data whose values can theoretically take any value
- Has equal units of measurement
- Difference or 'interval' between any two data values for a particular variable can be stated, but for which the relative difference cannot be stated
  - The difference between 15 and 20 degrees is the same magnitude as the difference between 25 and 30
  - However, 30 degrees is not twice as hot as 15 degree
- Zero point established at random
- Sometimes within a restricted range

Continuous - Ratio

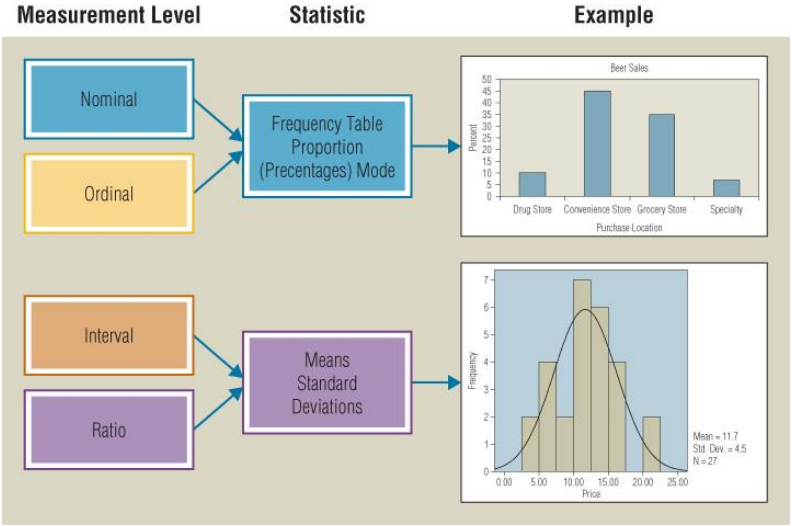
- E.g.
  - Grades, height, weight, age, length, income,
- Data whose values can theoretically take any value
- Has equal units of measurement
- Difference or 'interval' and relative difference between any two data values for a particular variable can be stated
  - The difference between R20 and R30 is the same magnitude as the difference between R40 and R50
  - And, R40 is twice as much as R20
- Has an absolute zero point (0 = total absence of the quality being measured)
- Can express values in terms of multiples and fractional parts

Data types

1 = Definitely disagree	1 = Definitely disagree
2 = Disagree	2 = Disagree
3 = Uncertain	**3 = Neither agree nor disagree**
4 = Agree	4 = Agree
5 = Definitely agree	5 = Definitely agree

Fuzzy
Distance between 1 and 2 the same as between 2 and 3?

Basic descriptive stats



These stats **describe** data − they do not necessarily **explain** the 'why'

SOUTHERN AFRICAN ASSOCIATION FOR INSTITUTIONAL RESEARCH

Frequency tables

	n	%
Male (1)	26	65.0
Female (2)	14	35.0
Total	**40**	**100.0**

saair
Southern African Association for Institutional Research

Frequency tables

Tutor class	n	%
10:00 (1)	11	27.5
12:00 (2)	0	0.0
14:30 (3)	10	25.0
17:30 (4)	8	20.0
20:00 (5)	5	12.5
22:15 (6)	6	15.0
Total	**40**	**100.0**

(**n** = number of responses)

saair
Southern African Association for Institutional Research

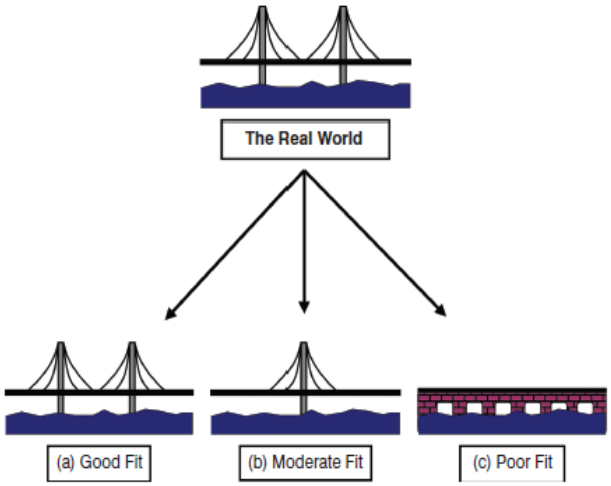
Measures of central tendency

- **Mean** − average sum of scores/number of scores
- **Mode** − most common value − 'typical' value
- **Median** − middle value

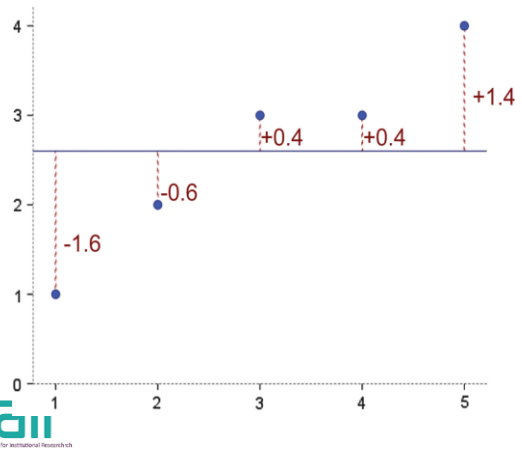
The mean

- A **hypothetical value** that doesn't have to be a value that is actually observed in the data
- Model created **to summarise our data**
- With any statistical model – **have to assess the fit** (**good, moderate, poor**)
- To determine **accuracy of model** – **look at how different our real data are from the model that we have specified**

Assessing the fit



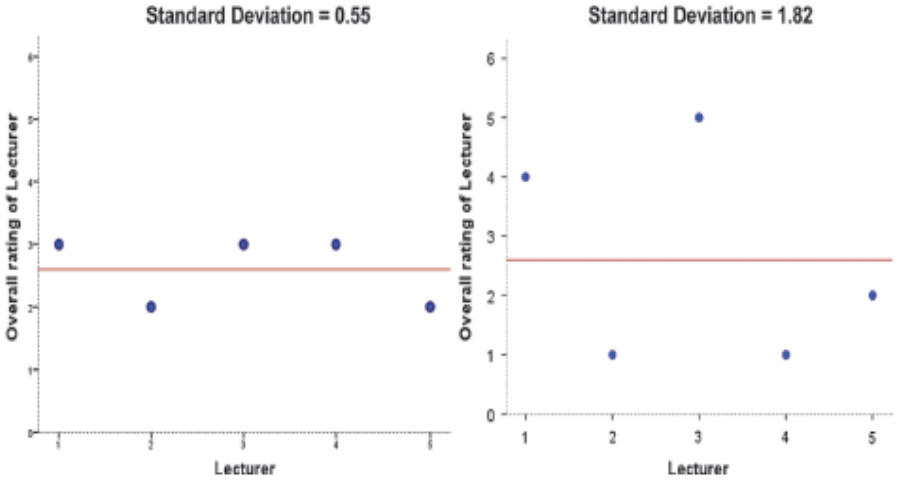
- Easiest way to do this - **look at difference between observed data and the model fitted**



Deviance
Error

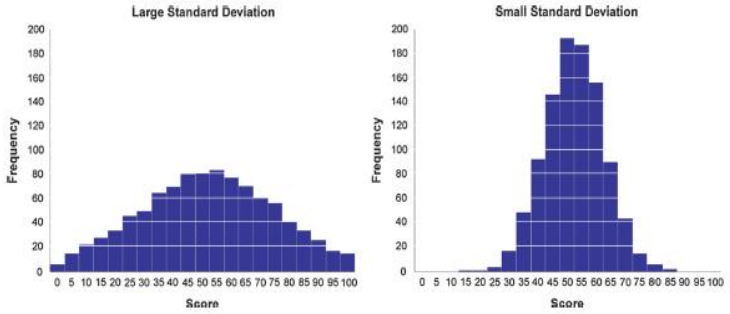
Over (+)
Under (-)

Assessing the fit



Assessing the fit

- Small SD => High precision
- Large SD => Low precision



SD tells us 'something' about shape of the distribution
(kurtosis, skewness)

Expressing the mean as a model

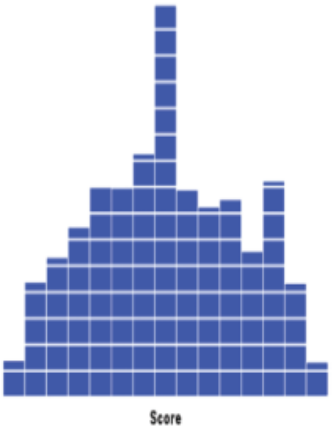
In statistics essentially **boils down to one equation**

$$\text{Outcome}_i = (\text{model}) + \text{error}_i$$

Other measures of central tendency

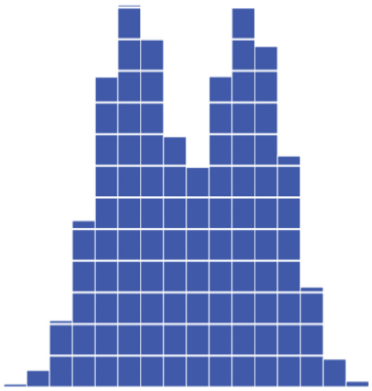
The mode

- Score that occurs **most frequently** in the data set



The mode

- One problem with the mode is that it can often take on several values (bimodal, multimodal) – this distribution also impacts on mean and std dev.



The median

- **Middle score** when scores are ranked in order of magnitude

  108, 103, 252, 121, 93, 57, 40, 53, 22, 116, 98

  Median = 98

- Mean = average score
- Median = score of the average sample unit

The median

- **Relatively unaffected by extreme scores** at either end of the distribution (whereas mean can be effected)
- Also **relatively unaffected by skewed distributions**
- Can be used with ordinal, interval and ratio data (but not nominal data)

Skewness

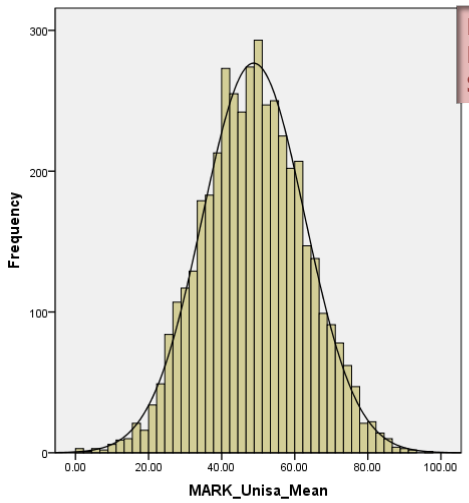
- **Skewness** is a measure of assymetry and refers to the tail of the distribution
- The following are measures of skewness:
  - A **normal distribution** is symmetrical and the mean, mode and median are the same or very close together (bell-shape curve)
  - A **positively skewed distribution** is when the mean is pulled to the right (much greater than the median)
  - A **negatively skewed distribution** is when the mean is pulled to the left (much less than the median)

Kurtosis

- **Kurtosis** is a measure of the presence of extreme values
- The following needs to be noted:
  - If the **distribution is normal**, kurtosis will be **0**
  - If the **distribution is relatively peaked** in the middle, kurtosis will be **> 0**
  - If the **distribution is relatively flat**, kurtosis will be **< 0**

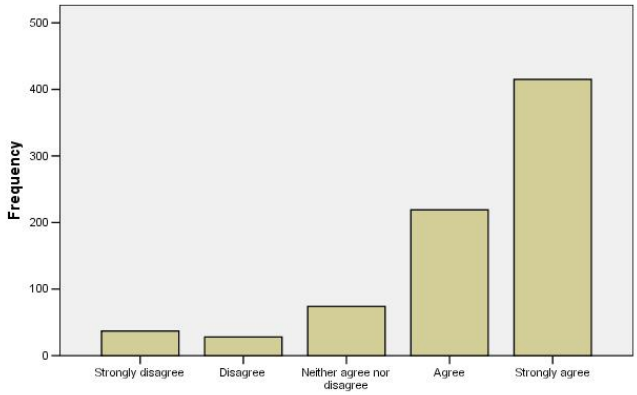
Choosing the correct statistical model is critical

Variation in our data



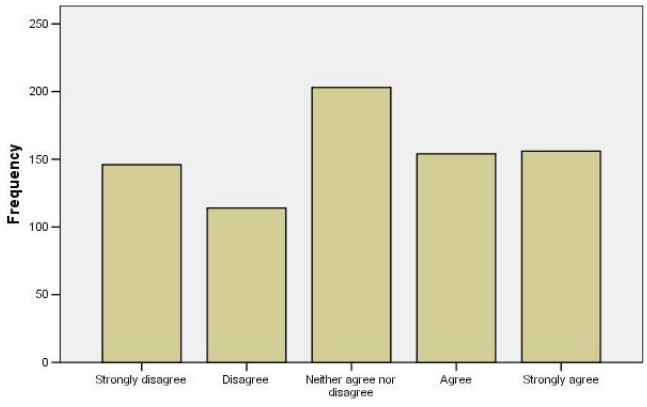
N = 4 373
Mean Unisa mark = 48,76
Standard deviation = 14,01

B3.2 Participating in school sport improves my health



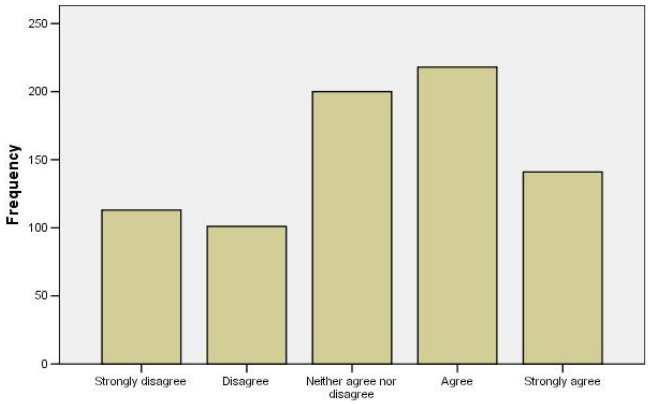
B3.2 Participating in school sport improves my health

B3.14 It is compulsory at my school to participate in sport

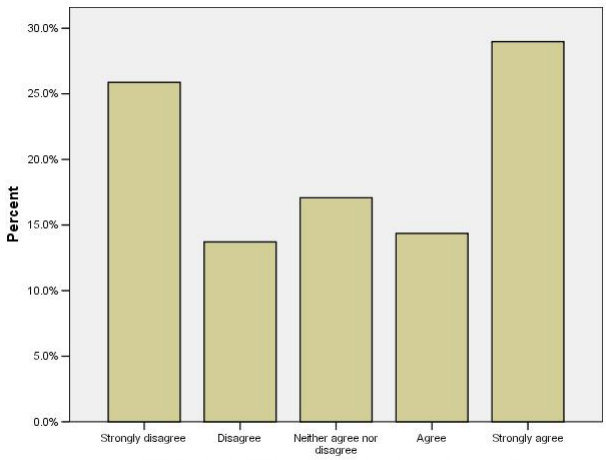


B3.14 It is compulsory at my school to participate in sport

B3.9 A variety of sport facilities is available at my school, therefore I take part



B3.9 A variety of sport facilities is available at my school, therefore I take part



B3.6 I enjoy individual sport (eg tennis, boxing, karate)

What if we draw the wrong conclusions from our analysis?

Results show that amongst students there is a positive correlation between sleeping with your shoes on and developing a headache

Can we deduce that if a student sleeps with his shoes it increases the probability of developing a headache (and can therefore not write exam)?

saair
Southern African Association for Institutional Research

Beware!!

You can't save poor research design (and dodgy data) with 'good' statistical analyses

saair
Southern African Association for Institutional Research