Why Do We Need Statistics?

Prof. Donald Davendra

ANDY FIELD

Types of Data Analysis

- Quantitative Methods
  - – Testing theories using numbers
- Qualitative Methods
  - – Testing theories using language
    - Magazine articles/Interviews
    - Conversations
    - Newspapers
    - Media broadcasts

The Research Process

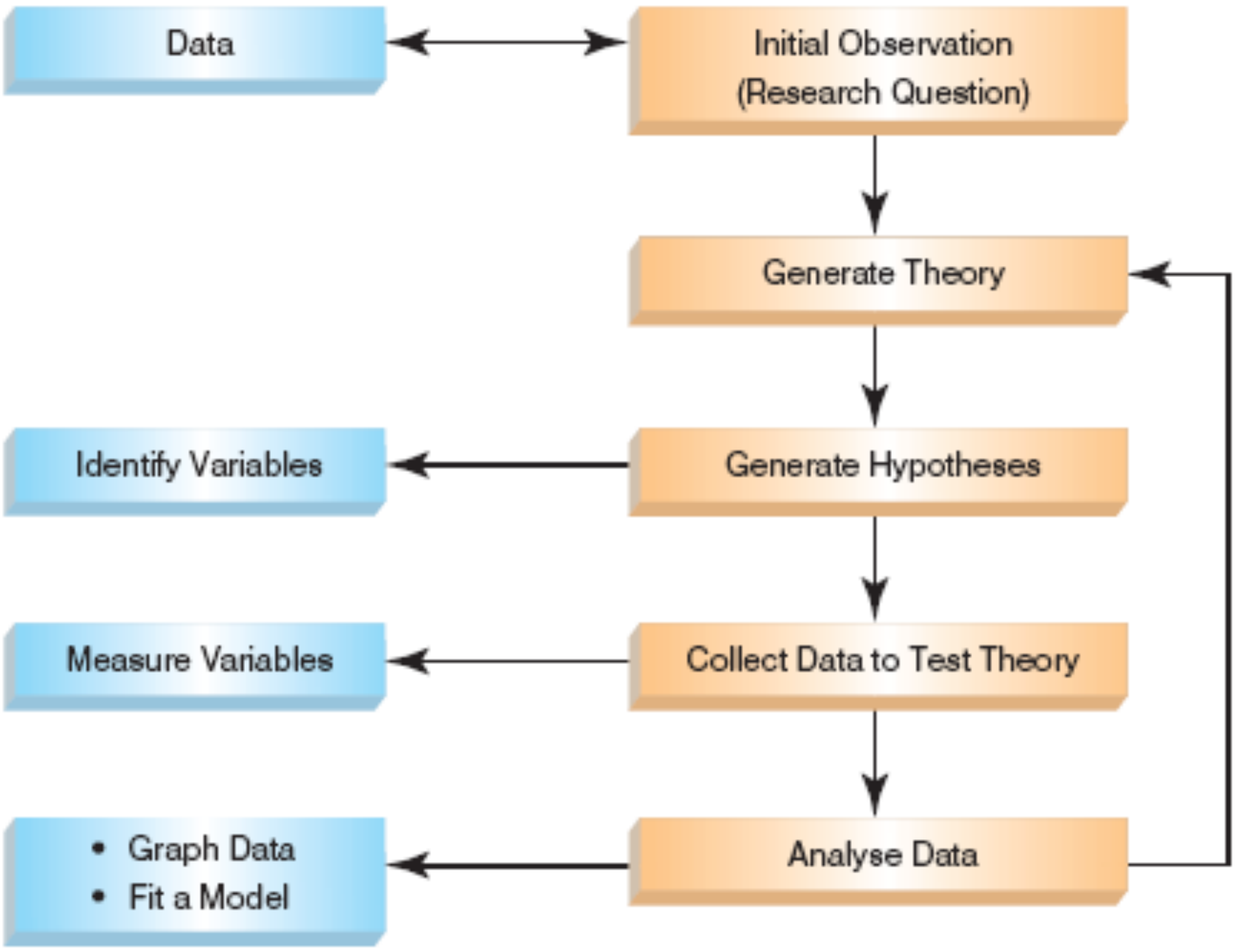


FIGURE 1.2 The research process

Initial Observation

- Find something that needs explaining
  - Observe the real world
  - Read other research
- Test the concept: collect data
  - Collect data to see whether your hunch is correct
  - To do this you need to define variables
    - Anything that can be measured and can differ across entities or time.

The Research Process

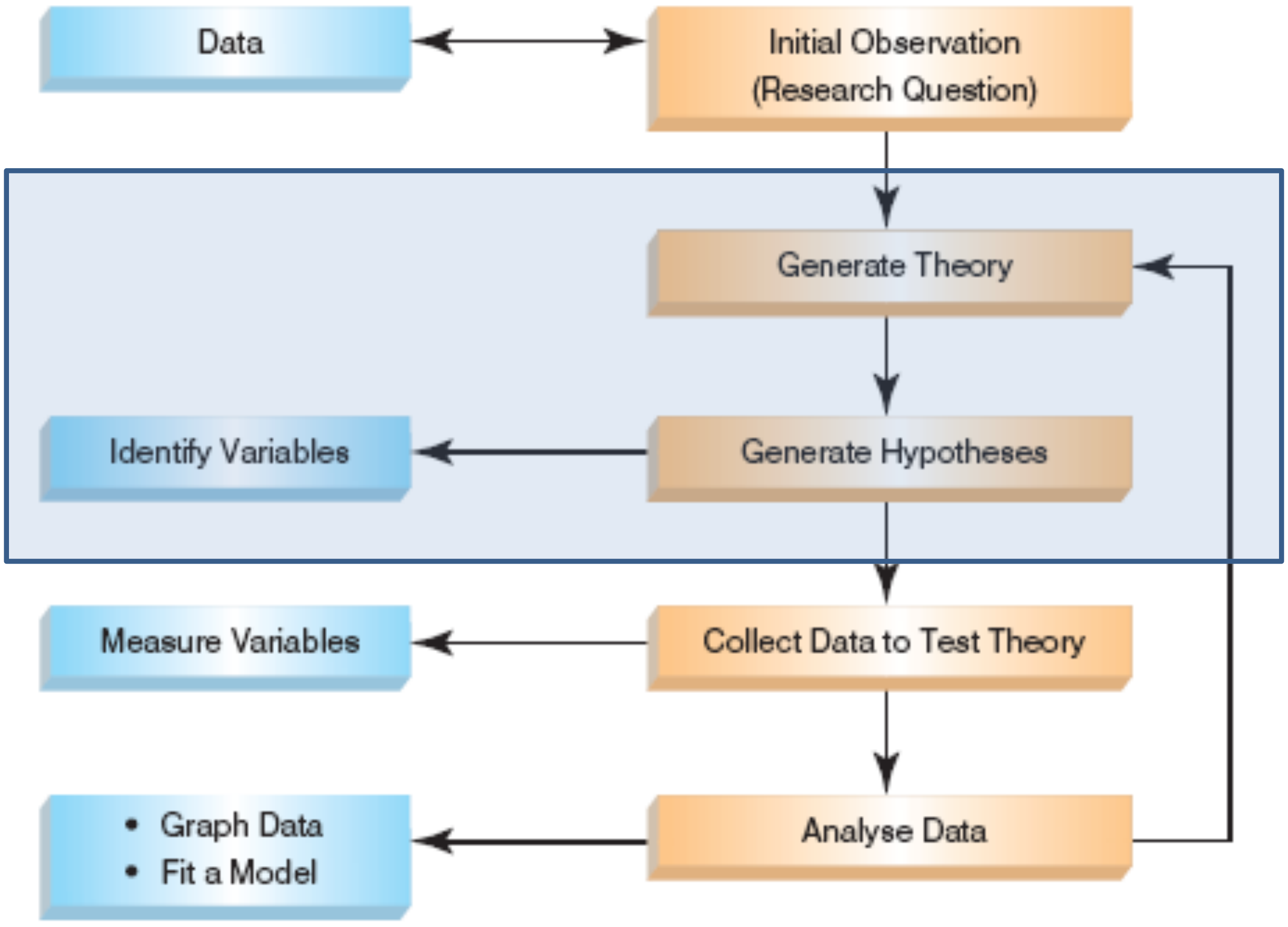


FIGURE 1.2 The research process

Generating and Testing Theories

- Theory
  - A hypothesized general principle or set of principles that explains known findings about a topic and from which new hypotheses can be generated.
- Hypothesis
  - A prediction from a theory.
  - E.g. the number of people turning up for a *Big Brother* audition that have narcissistic personality disorder will be higher than the general level (1%) in the population.
- Falsification
  - The act of disproving a theory or hypothesis.

TABLE 1.1 A table of the number of people at the *Big Brother* audition split by whether they had narcissistic personality disorder and whether they were selected as contestants by the producers

	No Disorder	Disorder	Total
Selected	3	9	12
Rejected	6805	845	7650
Total	6808	854	7662

DISCOVERING STATISTICS USING R

The Research Process

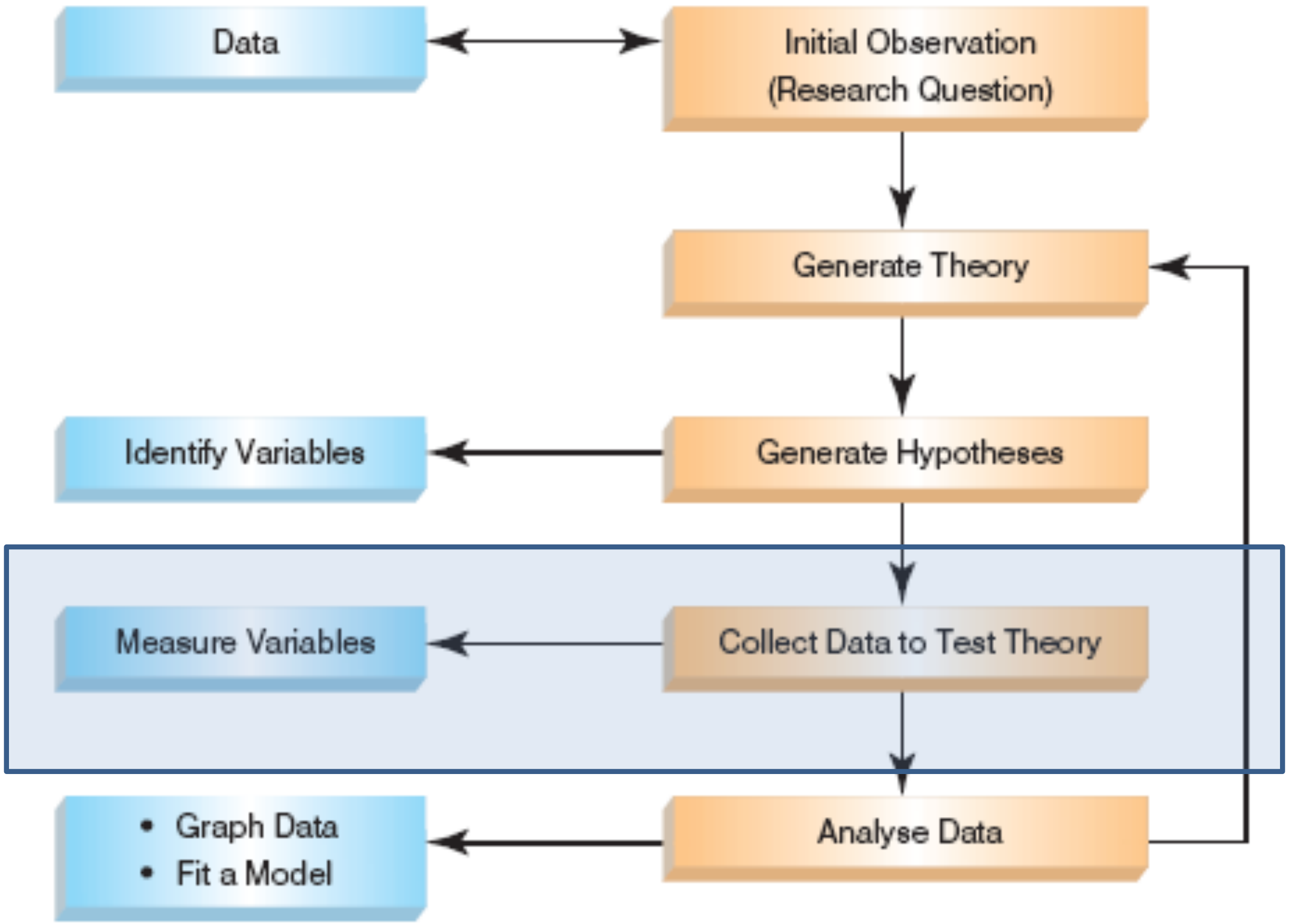


FIGURE 1.2 The research process

ANDY FIELD

Data Collection 1: What to Measure?

- Hypothesis:
  - *Coca-Cola kills sperm.*
- Independent Variable
  - The proposed cause
  - A **predictor** variable
  - A manipulated variable (in experiments)
  - Coca-Cola in the hypothesis above
- Dependent Variable
  - The proposed effect
  - An **outcome** variable
  - Measured not manipulated (in experiments)
  - Sperm in the hypothesis above

Levels of Measurement

- Categorical (entities are divided into distinct categories):
  - **Binary variable**: There are only two categories
    - e.g. dead or alive.
  - **Nominal variable**: There are more than two categories
    - e.g. whether someone is an omnivore, vegetarian, vegan, or fruitarian.
  - **Ordinal variable**: The same as a nominal variable but the categories have a logical order
    - e.g. whether people got a fail, a pass, a merit or a distinction in their exam.
- Continuous (entities get a distinct score):
  - **Interval variable**: Equal intervals on the variable represent equal differences in the property being measured
    - e.g. the difference between 6 and 8 is equivalent to the difference between 13 and 15.
  - **Ratio variable**: The same as an interval variable, but the ratios of scores on the scale must also make sense
    - e.g. a score of 16 on an anxiety scale means that the person is, in reality, twice as anxious as someone scoring 8.

Measurement Error

- Measurement error
  - The discrepancy between the actual value we're trying to measure, and the number we use to represent that value.
- Example:
  - You (in reality) weigh 80 kg.
  - You stand on your bathroom scales and they say 83 kg.
  - The measurement error is 3 kg.

Validity

- Whether an instrument measures what it set out to measure.
- Content validity
  - Evidence that the content of a test corresponds to the content of the construct it was designed to cover
- Ecological validity
  - Evidence that the results of a study, experiment or test can be applied, and allow inferences, to real-world conditions.

Reliability

- Reliability
  - The ability of the measure to produce the same results under the same conditions.
- Test–Retest Reliability
  - The ability of a measure to produce consistent results when the same entities are tested at two different points in time.

Data Collection 2: How to Measure

- Correlational research:
  - Observing what naturally goes on in the world without directly interfering with it.
- Cross-sectional research:
  - This term implies that data come from people at different age points, with different people representing each age point.
- Experimental research:
  - One or more variable is systematically manipulated to see their effect (alone or in combination) on an outcome variable.
  - Statements can be made about cause and effect.

Experimental Research Methods

- Cause and Effect (Hume, 1748)
  1. Cause and effect must occur close together in time (contiguity).
  2. The cause must occur before an effect does.
  3. The effect should never occur without the presence of the cause.
- Confounding variables: the '*Tertium Quid*'
  - A variable (that we may or may not have measured) other than the predictor variables that potentially affects an outcome variable.
- Ruling out confounds (Mill, 1865)
  - An effect should be present when the cause is present and that when the cause is absent the effect should be absent also.
  - Control conditions: the cause is absent.

Methods of Data Collection

- Between-group/between-subject/independent
  - Different entities in experimental conditions
- Repeated-measures (within-subject)
  - The same entities take part in all experimental conditions.
  - Economical
  - Practice effects
  - Fatigue

Types of Variation

- Systematic Variation
  - Differences in performance created by a specific experimental manipulation.
- Unsystematic Variation
  - Differences in performance created by unknown factors.
    - Age, gender, IQ, time of day, measurement error, etc.
- Randomization
  - Minimizes unsystematic variation.
  - Practise effects – familiarity with task
  - Boredom effects – tired of repeating task
  - Counterbalance – juggle order of tasks

The Research Process

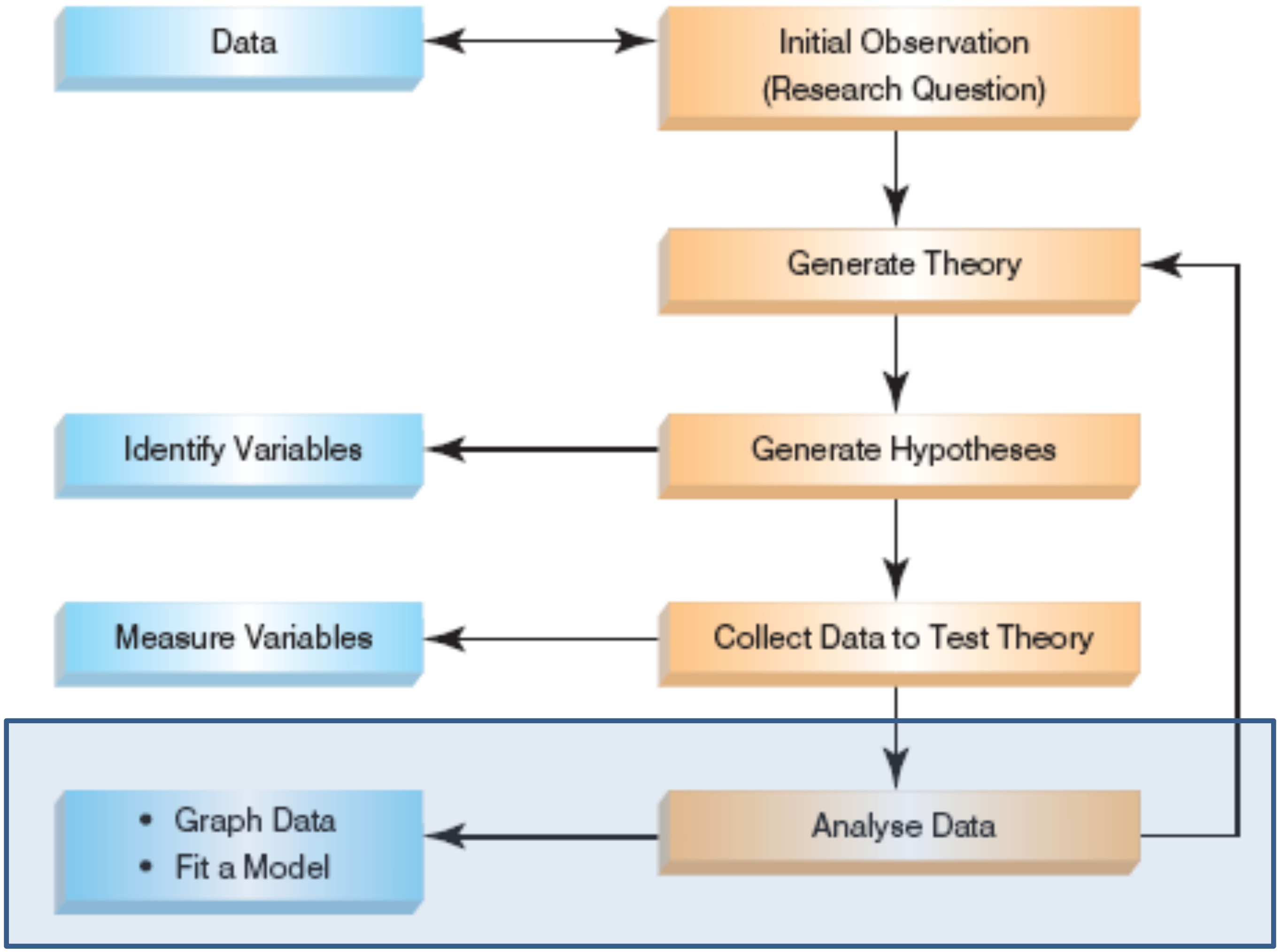


FIGURE 1.2 The research process

Analysing Data: Histograms

- Frequency Distributions (aka Histograms)
  - A graph plotting values of observations on the horizontal axis, with a bar showing how many times each value occurred in the data set.
- The ‘Normal’ Distribution
  - Bell-shaped
  - Symmetrical around the centre

The Normal Distribution

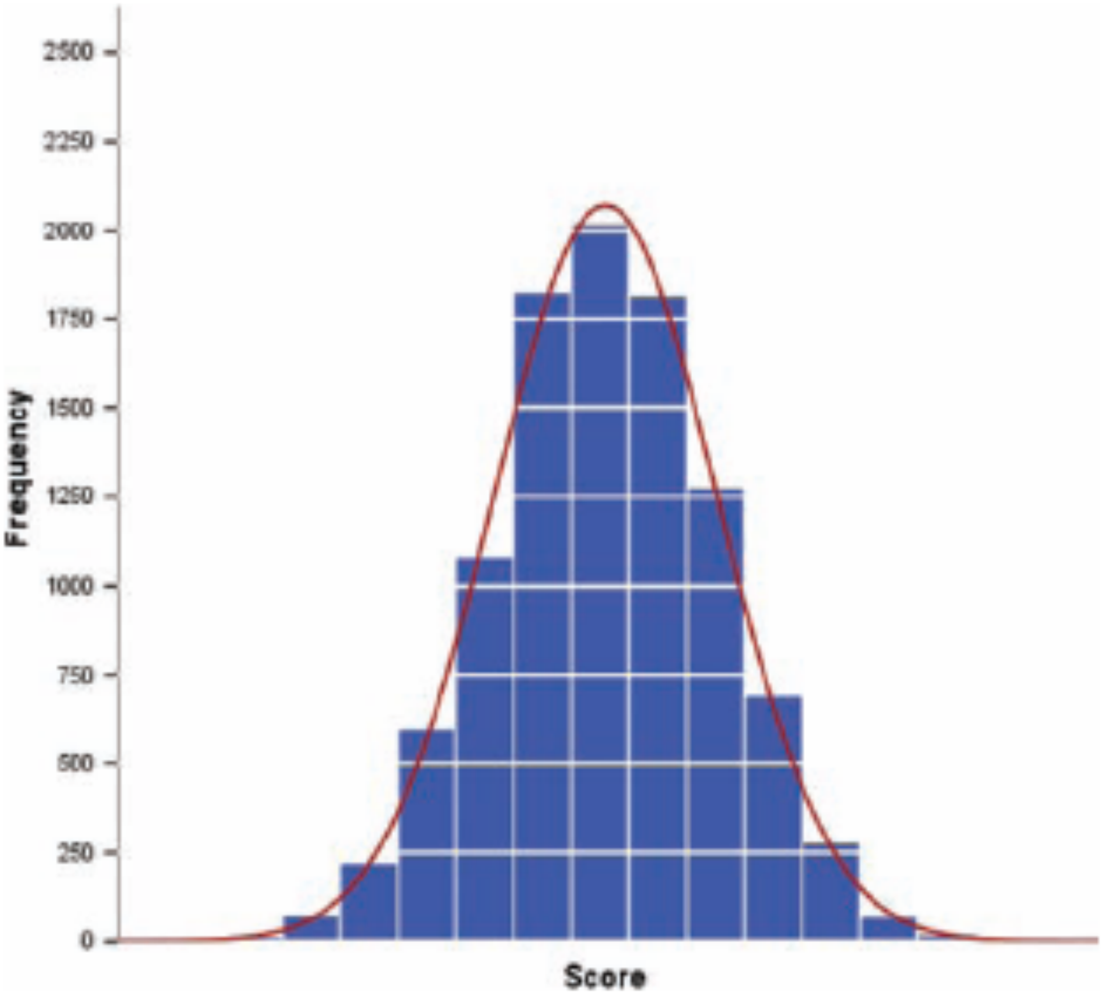


FIGURE 1.3 A ‘normal’ distribution (the curve shows the idealized shape)

Properties of Frequency Distributions

- Skew
  - The symmetry of the distribution.
  - Positive skew (scores bunched at low values with the tail pointing to high values).
  - Negative skew (scores bunched at high values with the tail pointing to low values).
- Kurtosis
  - The 'heaviness' of the tails.
  - Leptokurtic = heavy tails.
  - Platykurtic = light tails.

Skew

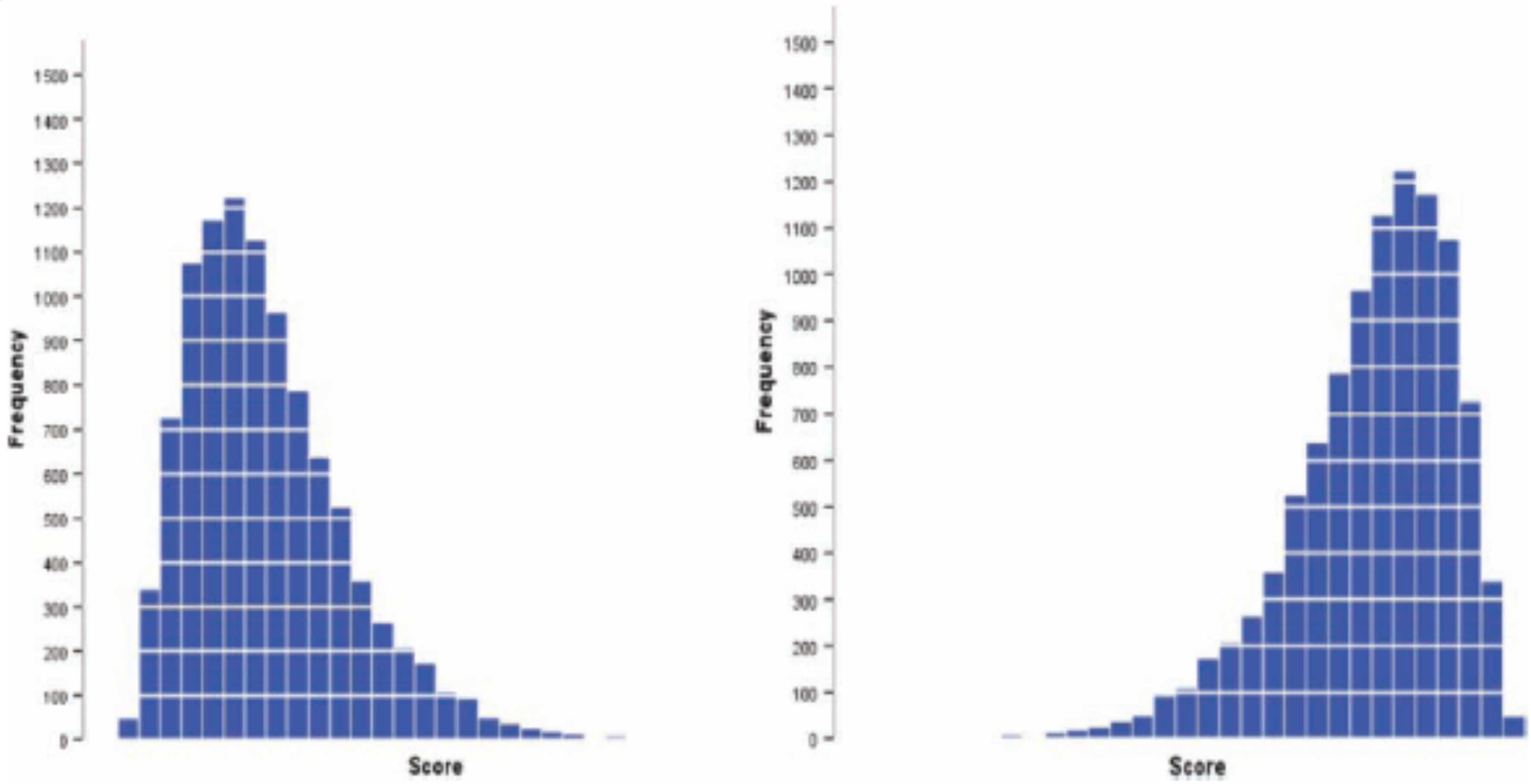


FIGURE 1.4 A positively (left figure) and negatively (right figure) skewed distribution

Kurtosis

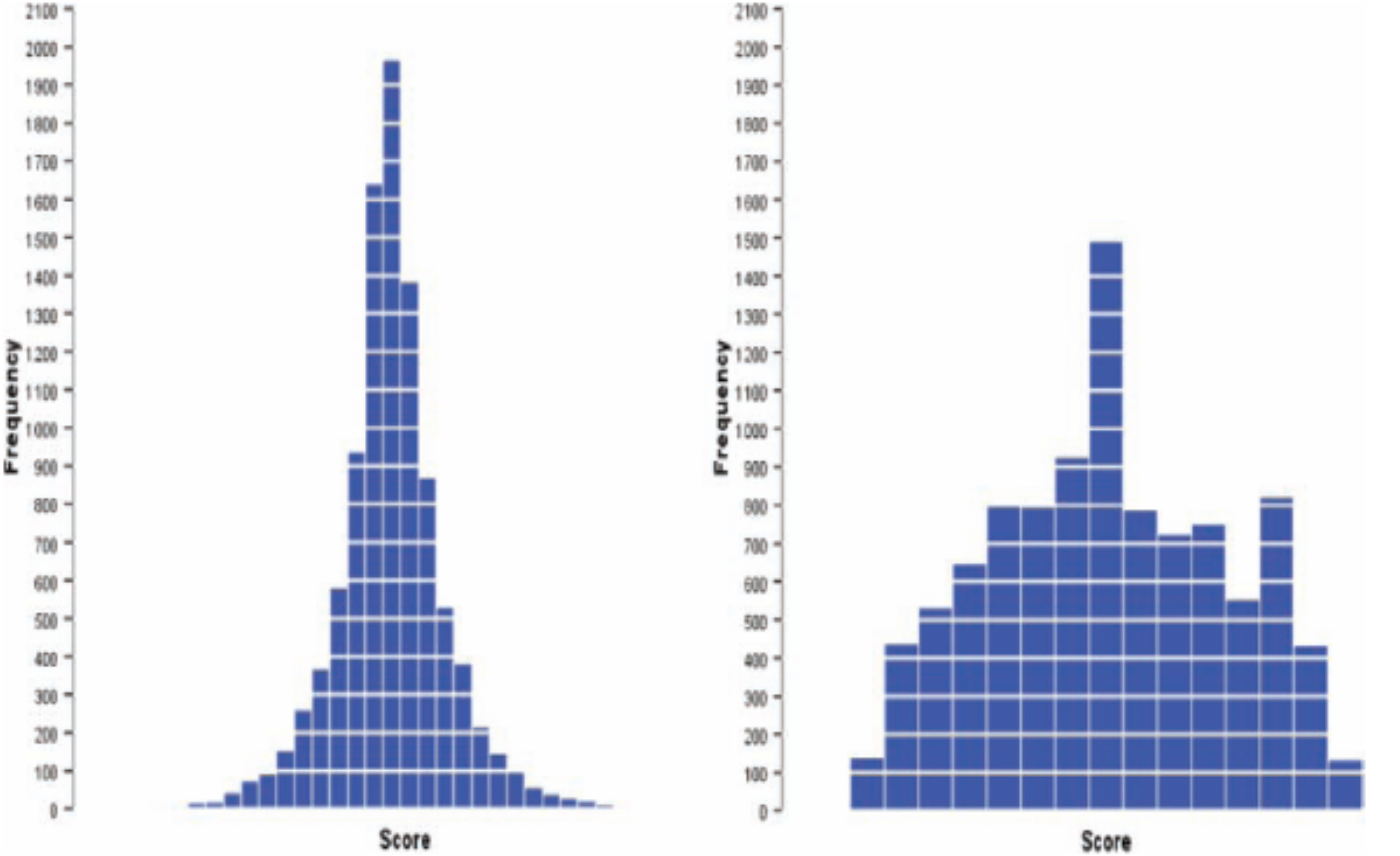


FIGURE 1.5 Distributions with positive kurtosis (leptokurtic, left figure) and negative kurtosis (platykurtic, right figure)

Central tendency: The Mode

- Central tendency
  - Centre of the frequency distribution
- Mode
  - The most frequent score
  - Tallest bar in the frequency distribution
- Bimodal
  - Having two modes
- Multimodal
  - Having several modes

A Bimodal Distribution

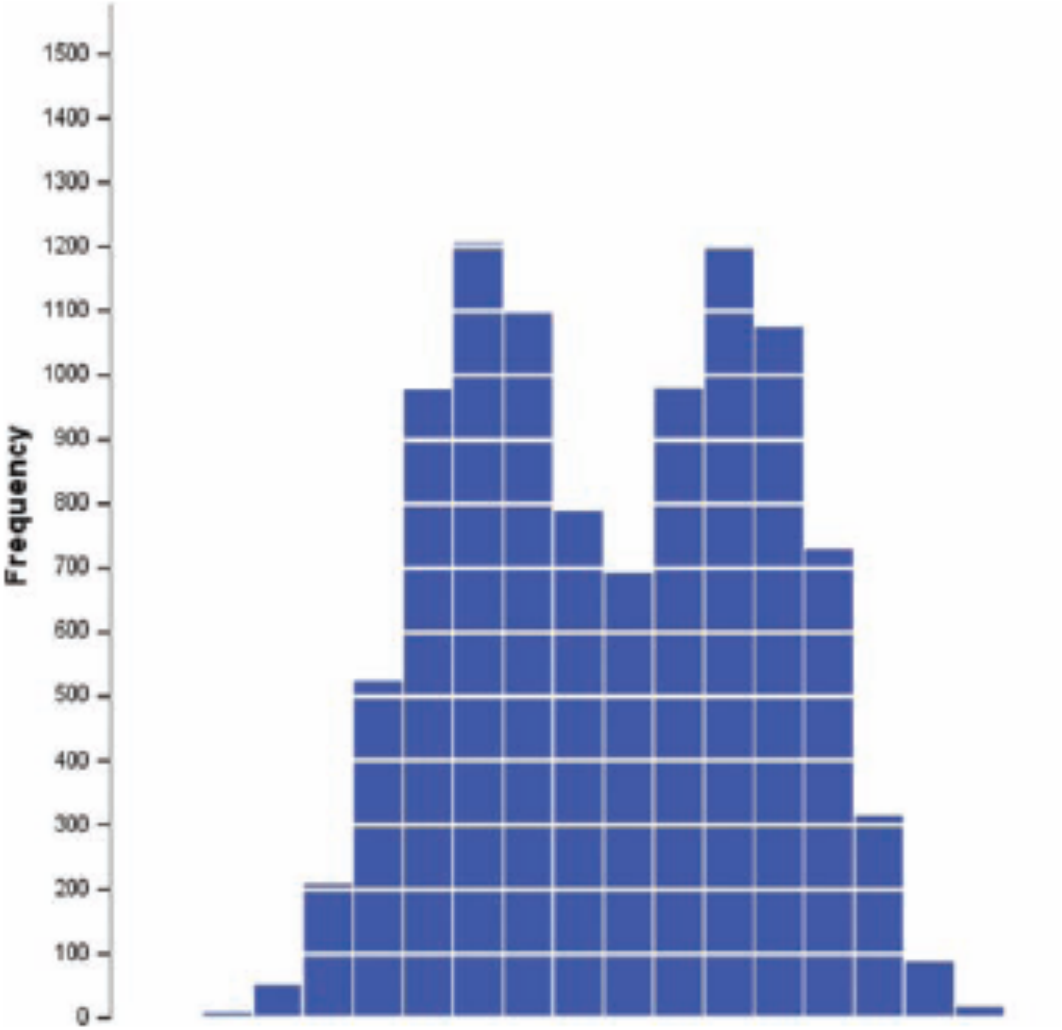
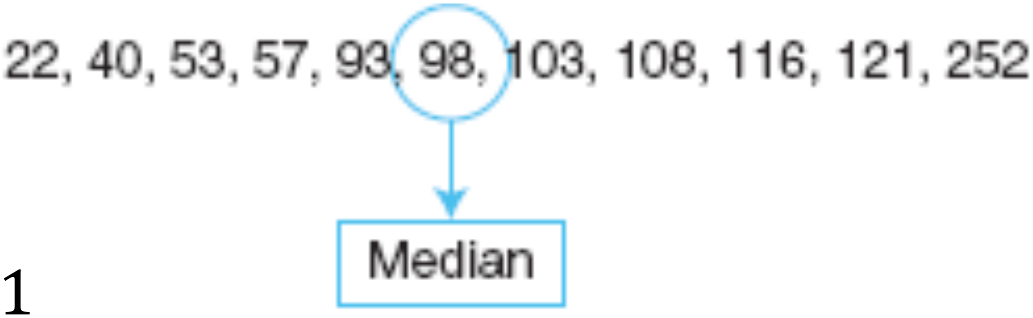


FIGURE 1.6
A bimodal distribution

Central Tendency: The Median

- Median
  - The middle score when scores are ordered.
- Example
  - Number of friends of 11 Facebook users.



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

Median

- Equation: $$\frac{n+1}{2}$$
  - n is the number of samples

Central Tendency: The Mean

- Mean
  - The sum of scores divided by the number of scores.
  - Number of friends of 11 Facebook users.

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\sum_{i=1}^{n} x_i = 22 + 40 + 53 + 57 + 93 + 98 + 103 + 108 + 116 + 121 + 252$$

$$= 1063$$

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1063}{11} = 96.64$$

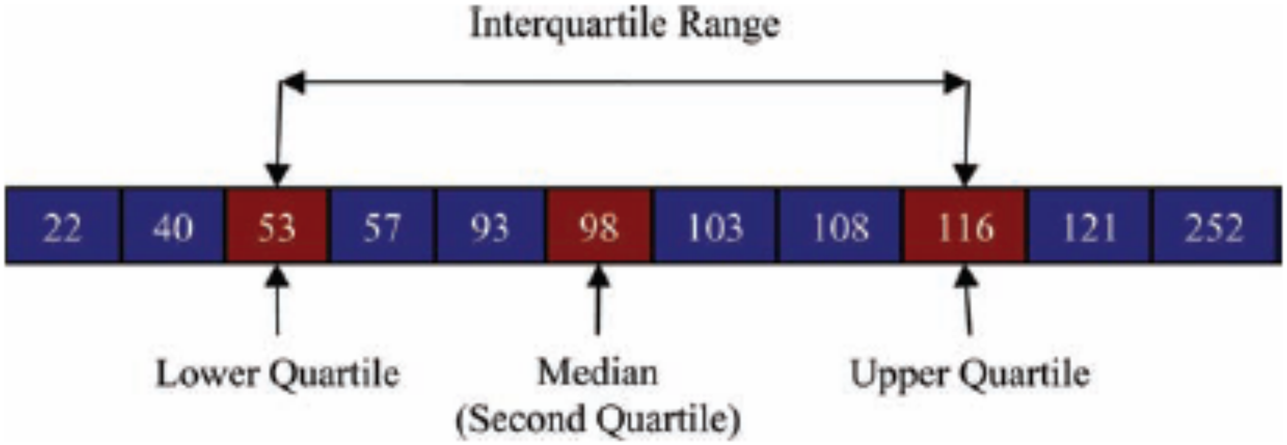
The Dispersion: Range

- The Range
  - The smallest score subtracted from the largest
- Example
  - Number of friends of 11 Facebook users.
  - 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252
  - Range = 252 – 22 = 230
  - Very biased by outliers

The Dispersion: The Interquartile range

- Quartiles
  - The three values that split the sorted data into four equal parts.
  - Second quartile = median.
  - Lower quartile = median of lower half of the data.
  - Upper quartile = median of upper half of the data.
  - Interquartile = Upper quartile – Lower quartile
  - We lose half of the data, but are not affected by extreme values

FIGURE 1.7 Calculating quartiles and the interquartile range



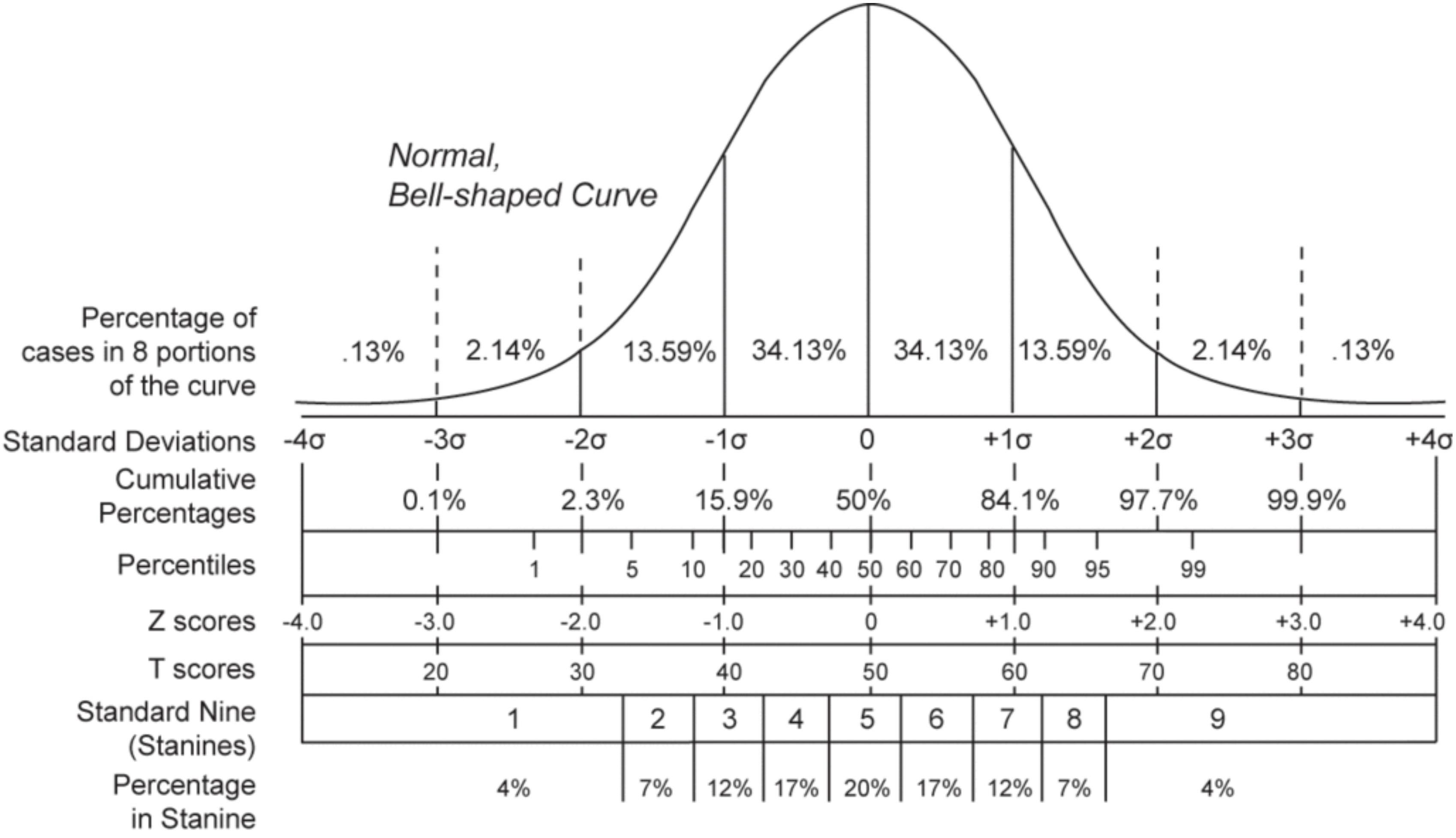
Going beyond the data: *z*-scores

- *z*-scores (standard score)
  - Standardising a score with respect to the other scores in the group.
  - Expresses a score in terms of how many standard deviations it is away from the mean.
  - The distribution of *z*-scores has a mean of 0 and *SD* = 1.

$$z = \frac{X - \overline{X}}{s}$$

Properties of *z*-scores

- 1.96 cuts off the top 2.5% of the distribution.
- −1.96 cuts off the bottom 2.5% of the distribution.
- As such, 95% of *z*-scores lie between −1.96 and 1.96.
- 99% of *z*-scores lie between −2.58 and 2.58.
- 99.9% of them lie between −3.29 and 3.29.


Normal,
Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Percentiles
1
5
10
20
30
40
50
60
70
80
90
95
99
Z scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T scores
20
30
40
50
60
70
80
Standard Nine (Stanines)
1
2
3
4
5
6
7
8
9
Percentage in Stanine
4%
7%
12%
17%
20%
17%
12%
7%
4%

Types of Hypotheses

- Null hypothesis, H_0
  - There is no effect.
  - E.g. *Big Brother* contestants and members of the public will not differ in their scores on personality disorder questionnaires
- The alternative hypothesis, H_1
  - Aka the experimental hypothesis
  - E.g. *Big Brother* contestants will score higher on personality disorder questionnaires than members of the public