DA-STATS

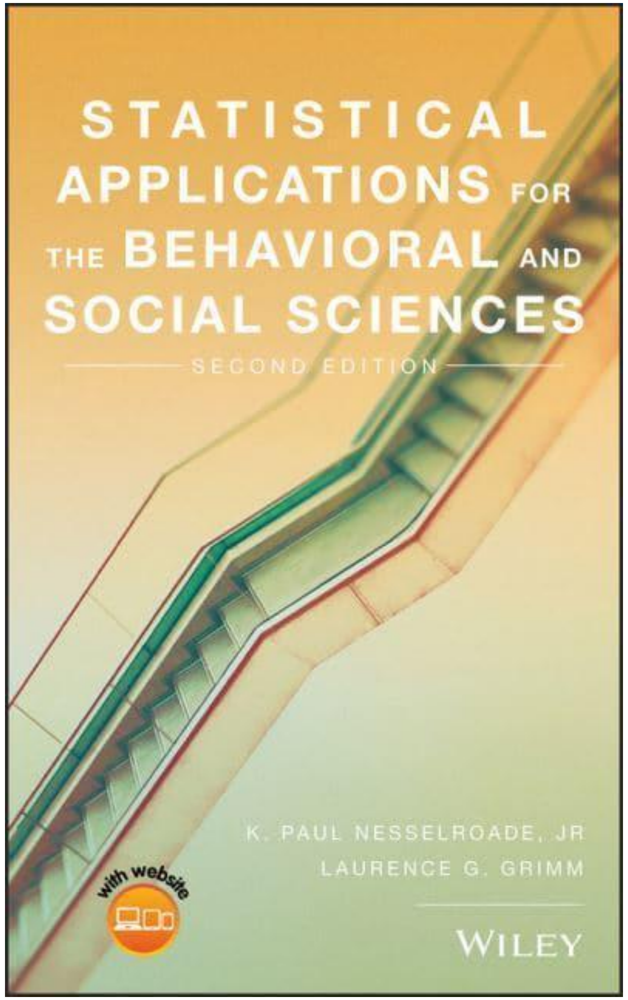
Topic 02: Introduction to Descriptive and Predictive Analysis

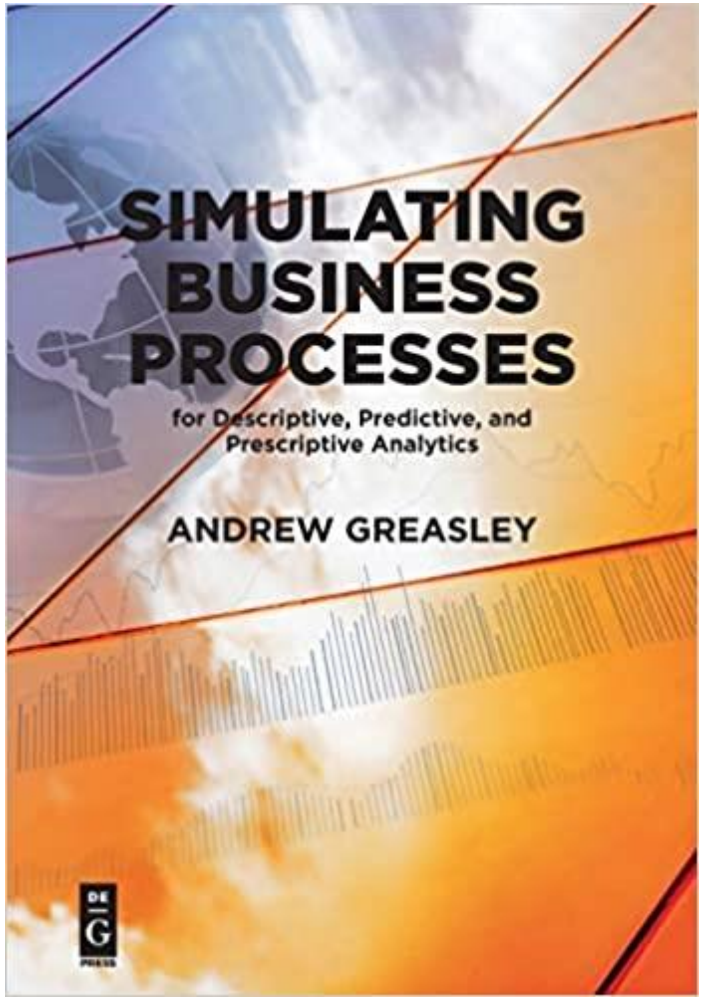
Shan E Ahmed Raza
Department of Computer Science
University of Warwick
shan.raza@warwick.ac.uk

Outline

- Introduction to Descriptive, Predictive and Prescriptive Techniques
- Data and Model Driven Modelling
- Descriptive Analysis
  - ➞Scales of Measurement and data
  - ➞Measures of central tendency, statistical dispersion and shape of distribution
- Correlation Techniques

Books – Lecture Material





Statistical Analysis

- Descriptive analysis – understand the past.
- Predictive analysis – predict the future.
- Prescriptive analysis – recommend an action.

Data driven vs Model driven approach

- Data-driven modelling approach
  - ➞ Aims to derive a **description of behavior from observations** of a system so that it can describe how that system behaves (its output) under different conditions or scenarios (its input).
  - ➞ Generally, **the more data (observations)** that can be used to form the description, **the more accurate the description** will be and thus the interest in big data analytics that uses large data sets.
  - ➞ Machine learning uses a selection of learning algorithms that use **large data sets and a desired outcome** to derive an algorithm

Data driven vs Model driven approach

- Model-driven modelling approach
  - ➞Aims to explain a system's behavior not just derived from its inputs but **through a representation of the internal system's** structure.
  - ➞a real system is simplified into its **essential elements** (its processes) and **relationships between these elements** (its structure).
  - ➞in addition to input data, information is required on the **system's processes**, the function of these processes and the essential parts of the relationships between these processes.
  - ➞are called **explanatory models** as they represent the real system and attempt to explain the behavior that occurs.
  - ➞generally, have **far smaller data needs than data-driven models** because of the key role of the representation of structure.

Descriptive Analysis

- Use of reports and visual displays to explain or understand past and current business performance
- Contain statistical summaries of metrics such as sales and revenue
- Intended to provide an outline of trends in current and past performance

What Happened?

Predictive Analysis

- Ability to predict future performance
- Detecting patterns or relationships in historical data
- Project these relationships into the future
- Domain knowledge to construct a simplified representation

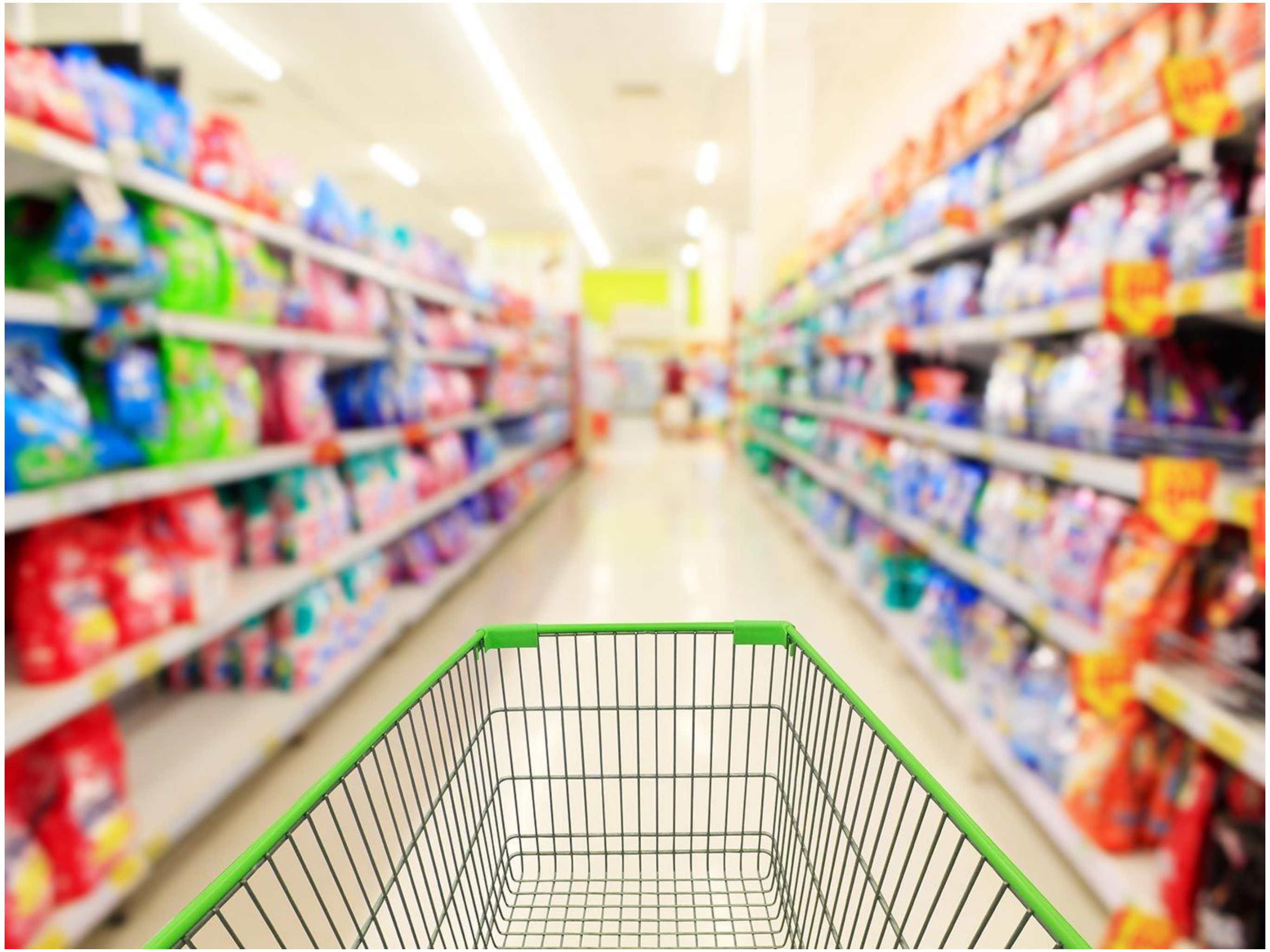
What could happen?

Prescriptive Analysis

- Recommend a choice of action from predictions of future performance
- Optimum decision based on the need to maximize (or minimize) some aspect of performance
- Many different scenarios can be tested until one is found that best meets the optimization criteria

What should be done?

Supermarket



Data Driven Modelling

- Data-driven modeling aims to derive a description of behavior from observations of a system.
- Describe the **relationship between input and output**
- Also known as descriptive models
  - ➞Note this is different from *descriptive analysis*
- **Imitates real behavior**
- More data (observations) to form the description → more accurate the description

Model Driven Modelling

- Explain a system's behavior not just derived from its inputs but through a representation of the internal system's structure.
- A real system is simplified into its essential elements (its processes) and relationships between these elements (its structure)
- The **effect of a change on design of the process can be assessed** by changing the structure of the model.
- Generally, we have far smaller data needs than data-driven models

Pros and Cons

- Data driven models
  - ➞ built in error terms.
  - ➞ errors can be quantified, and confidence levels can be estimated
  - ➞ large amount of data to estimate the parameters to fit the model

- Process driven models
  - ➞ built using mathematical equations
  - ➞ errors may be introduced during simplifications
  - ➞ real-world observations are used to evaluate the model

Descriptive Statistics

- Usually, the first step prior to more complex analysis
- Scales of Measurement
- Using Tables to Organize Data
- Measures of
  - ➞ Location or central tendency
  - ➞ Statistical dispersion
  - ➞ Shape of distribution

Scales of Measurement

- Measurement, is the assignment of numbers to attributes, objects, or events according to predetermined rules.
- Four different measurement scales
  - ➞Nominal Scales
  - ➞Ordinal Scales
  - ➞Interval Scales
  - ➞Ratio Scales

Nominal Scale

- No quantitative information
- Numbers merely to distinguish one type of thing from another type of thing or one event from another event
- As no quantitative information being communicated, we are free to exchange one number for any other currently unused number
- For instance, the numbers assigned to the members of a football team do not carry any quantitative value.

Ordinal Scales

- Adds relative position information to nominal scale
- Reflects a quantitative relationship between the various categories
- Rankings
  - ➙Comparatively more or less
  - ➙Not how much

Interval Scales

- Set of quantitatively ordered categories
- Intervals between the categories are held constant
- Do not possess a true zero point.
- Different interval scale can be used to measure the same amount

Ratio Scale

- Addition of an absolute zero point to interval scale
- Zero marks the absence of a quantity

Discrete and Continuous Quantities

- An important feature of measuring variables concerns how many different values can be assigned

Discrete and Continuous Quantities

- Discrete Variables
  - ➞Can take on only a finite number of values
  - ➞No meaningful values exist between any two adjacent values
  - ➞To find statistical features of sets of discrete data it is permissible to use “in-between” values

Discrete and Continuous Quantities

- Continuous Variables
  - ➞ Theoretically have an infinite number of points between any two numbers
  - ➞ Variables do not have gaps between adjacent numbers

Unorganised Raw Data - Example

Scores from 90 participants who completed a questionnaire measuring their achievement.

15	8	20	16	12	18	14	22	17	5
19	15	18	29	6	13	16	19	10	24
15	3	26	30	13	17	7	16	23	25
1	15	18	14	5	27	16	20	14	6
24	14	20	25	21	15	17	8	23	21
17	14	10	13	18	16	21	9	11	22
15	12	9	16	20	11	13	22	17	13
9	22	16	12	19	17	14	10	19	18
11	16	12	18	13	17	15	14	15	28

Simple Frequency Distribution

- How many participants received each score?
- Frequency is the number of occurrences of a repeating event
- Another example is frequency of radio waves defined per unit time. Number of times a wave completes its cycle in a unit time (per sec)

$$f = \frac{1}{T}$$

X	*f*	*X*	*f*
30	1	14	7
29	1	13	6
28	1	12	4
27	1	11	3
26	1	10	3
25	2	9	3
24	2	8	2
23	2	7	1
22	4	6	2
21	3	5	2
20	4	4	0
19	4	3	1
18	6	2	0
17	7	1	1
16	8	0	0
15	8		

Grouped Frequency Distribution

- The number of scores that fall into each of several ranges of scores
- Some sets of data cover a wide range of possible scores which can make the resulting frequency distributions long and cumbersome
- Exchange loss of information with a table that is easy to understand

Class interval	Midpoint	*f*
30–32	31	1
27–29	28	3
24–26	25	5
21–23	22	9
18–20	19	14
15–17	16	23
12–14	13	17
9–11	10	9
6–8	7	5
3–5	4	3
0–2	1	1

Cumulative Frequency Distributions

- The sum of frequencies found at that interval plus all preceding intervals
- Keeps a running tally of all scores up through each given interval

Class interval	*f*	*Cum f*
30–32	1	90
27–29	3	89
24–26	5	86
21–23	9	81
18–20	14	72
15–17	23	58
12–14	17	35
9–11	9	18
6–8	5	9
3–5	3	4
0–2	1	1

Measures of Central Tendency

- Statistical indices designed to communicate what is the “center” or “middle” of a distribution
- Three measures of central tendency
  - Mean
  - Median
  - Mode

The Mean

- The mean, colloquially referred to as the “average,” is the most frequently used measure of central tendency.

$$\mu = \frac{\Sigma X}{N}$$

μ = the symbol for the mean of a population

X = a score in the distribution

N = population size

Σ = sum up a set of scores, $\Sigma X = X_1 + X_2 + X_3 + \cdots + X_N$

The Mean

- What is the mean of this population of scores?

$$5, 8, 10, 11, 12$$

$$\mu = \frac{5 + 8 + 10 + 11 + 12}{5} = \frac{46}{5} = 9.20$$

The Mean of a Frequency Distribution

- Mean of a frequency distribution

$$\mu = \frac{\Sigma X f}{\Sigma f}$$

f $=$ frequency with which a score appears

Calculating the mean from a frequency distribution

X	f	Xf
7	1	7
6	3	18
5	2	10
4	5	20
3	4	12
2	1	2
1	1	1
	$n = \Sigma f = 17$	$\Sigma Xf = 70$
	$M = \dfrac{\sum Xf = 70}{\sum f = 17} = 4.12$	

The Weighted Mean

WARWICK

- Weighted Mean

$$M = \frac{n_1 V_1 + n_2 V_2 + n_3 V_3 + \cdots n_4 V_5}{n_1 + n_2 + n_3 + \cdots n_n}$$

Weighted Mean - Example

- Let's consider score of students in a subject from three schools

School A

Student	Score
A	60
B	40

School B

Student	Score
C	80
D	70
E	60

School B

Student	Score
F	75
G	60
H	60

Mean: $M1 = 50$ $M2 = 70$ $M3 = 65$

Mean of all scores: ~63.125

Mean of $M1, M2, M3$: ~61.667

Weighted Mean of $M1, M2, M3$ where $n_1 = 2, n_2 = 3, n_3 = 3 = \frac{2(50) + 3(70) + 3(65)}{2 + 3 + 3} = 63.125$

The Median

- The median divides the distribution based on the frequency or number of scores above and below a given point.
- Not algebraically defined but there are algorithms to calculate the median

Calculating the Median – Example 1

- What is the median of this distribution?

40, 1, 4, 42, 6, 8, 43, 45, 47

↓ Sort

1, 4, 6, 8, 40, 42, 43, 45, 47

4 scores | Median | 4 scores

Calculating the Median – Example 2

WARWICK

- What is the median of this distribution? (with even number of scores)

3, 9, 15, 16, 19, 22

2 scores

2 scores

$$\text{Median} = \frac{15+16}{2} = 15.5$$

The Mode

- The most typical or most frequent score in the distribution.

The Mode - Example

- What is the mode of this distribution?

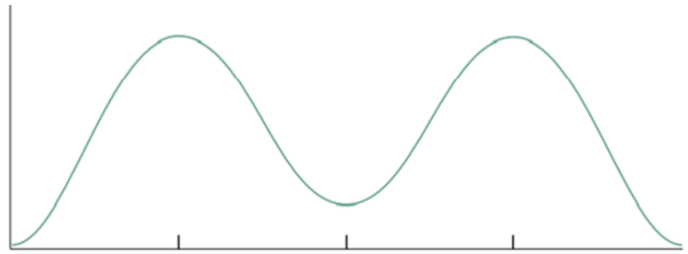
100, 101, 105, 105, 107, 108

105

The Mode from Frequency Distribution

Mode: 15 & 16

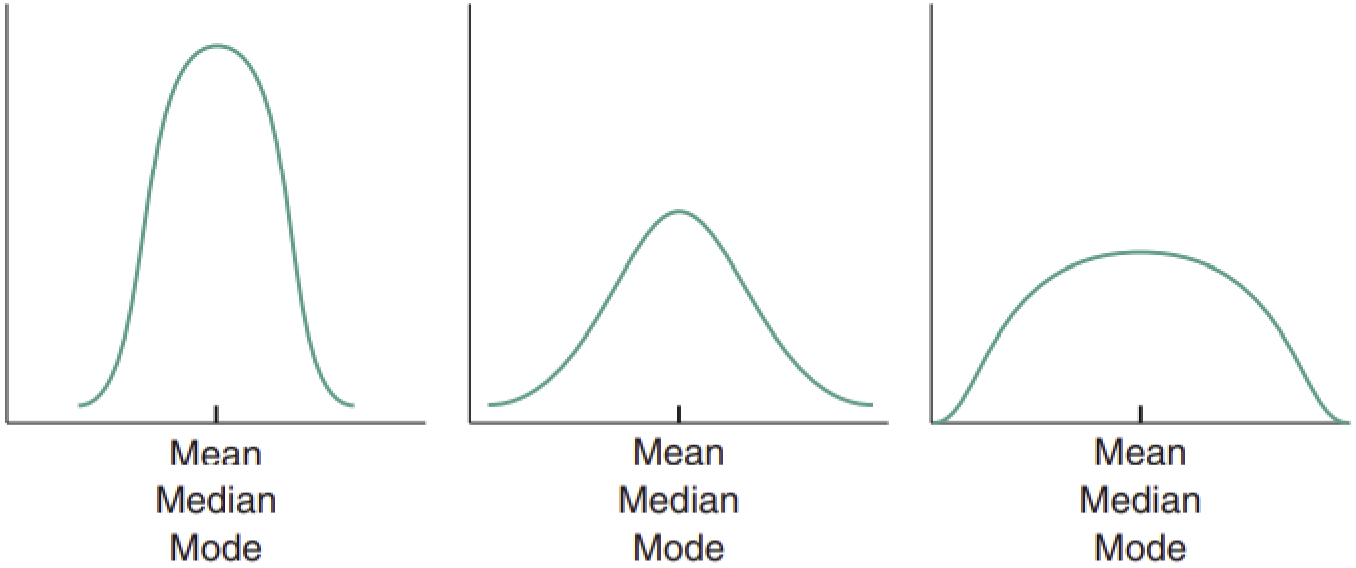
- This is known as a bimodal distribution ("bi," meaning two).



- A distribution with a single mode is termed unimodal ("uni," meaning one).
- Having more than two modes is called "multimodal".

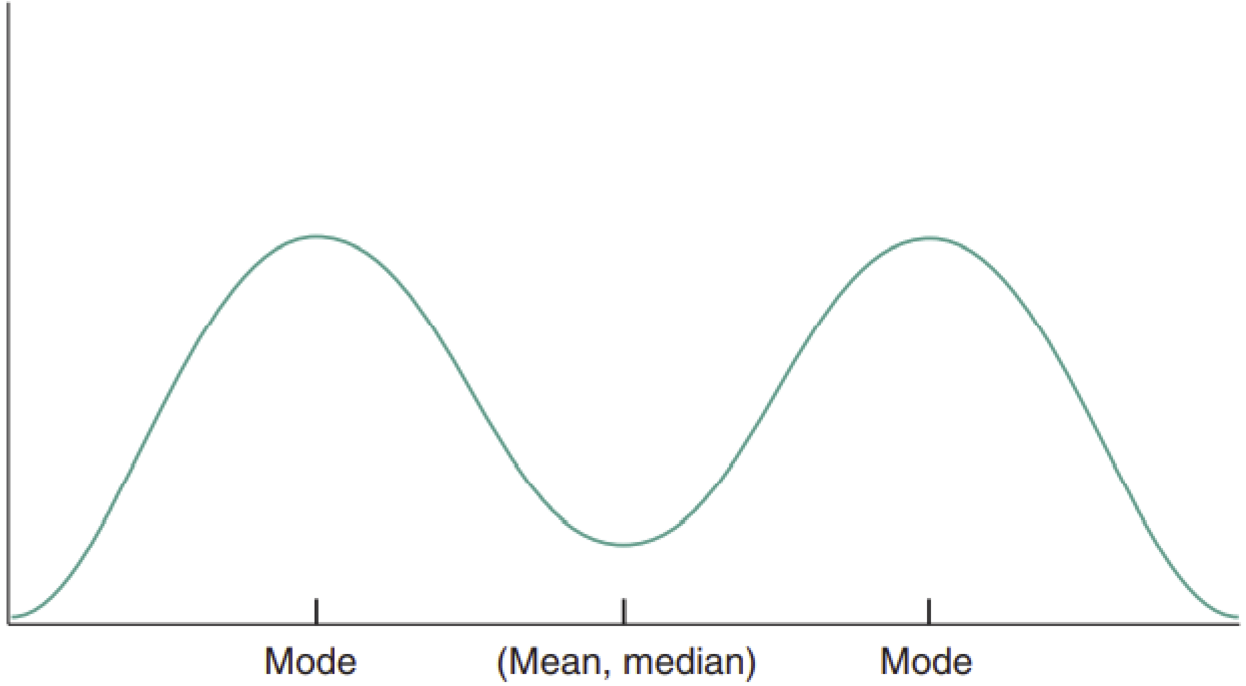
X	f	X	f
30	1	14	7
29	1	13	6
28	1	12	4
27	1	11	3
26	1	10	3
25	2	9	3
24	2	8	2
23	2	7	1
22	4	6	2
21	3	5	2
20	4	4	0
19	4	3	1
18	6	2	0
17	7	1	1
16	8	0	0
15	8		

How the Shape of Distributions Affects Measures of Central Tendency



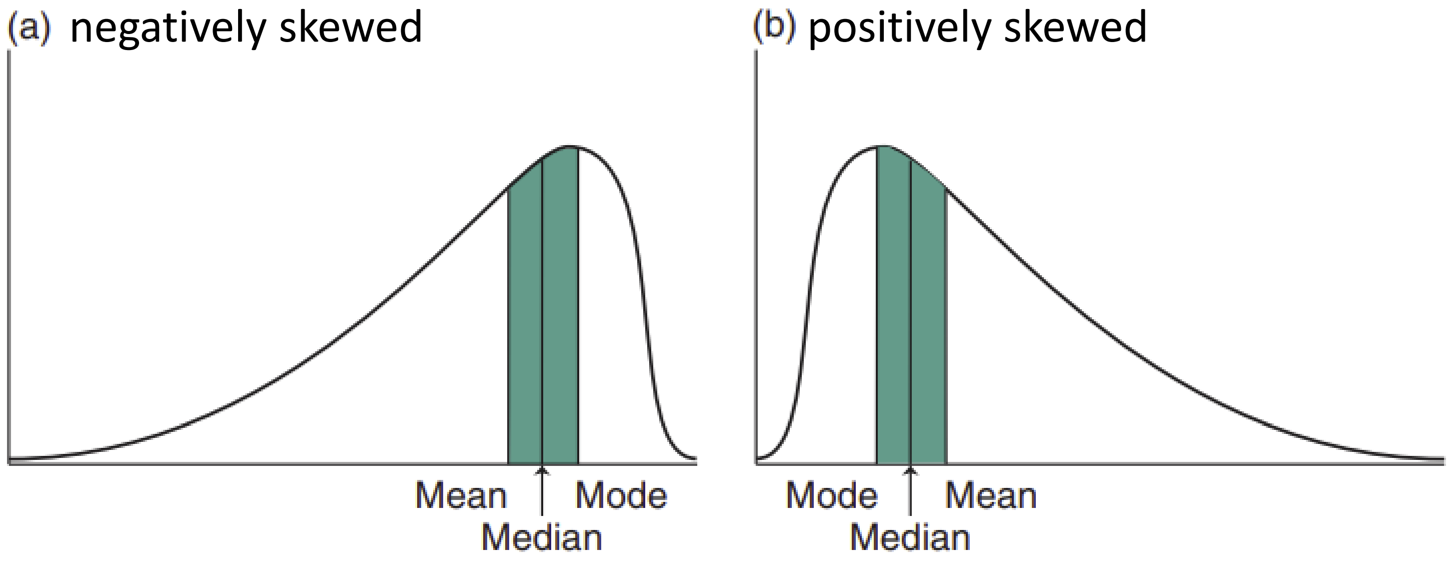
- If the distribution is symmetrical, then all three measures of central tendency will be identical.

How the Shape of Distributions Affects Measures of Central Tendency



- If the distribution is symmetrical but bimodal, then mean and median are identical but not the mode.

How the Shape of Distributions Affects Measures of Central Tendency

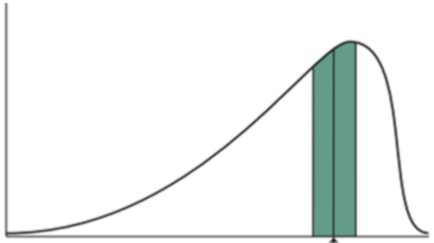


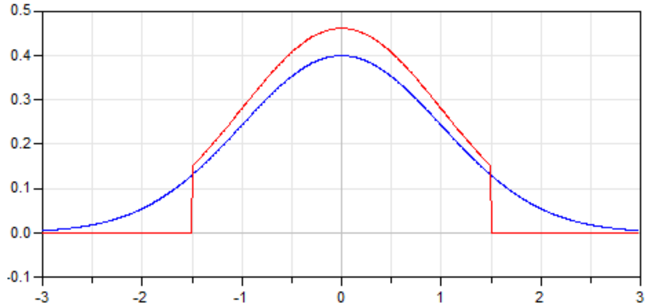
- If a distribution is skewed, then the mean, median, and mode will all be different.

When to Use the Mean, Median, and Mode?

- Mean – no meaningful quantitative information for nominal or ordinal scales
- Median – measure of choice for ordinal scale
- Mode – preferred for nominal scale
- Sensitivity to extreme scores
- Skewed Distributions
- Scores of a distribution are truncated

Family income (beginning of Congressman Windblows' term)	Family income (end of Congressman Windblows' term)
\$44 000	\$44 000
\$48 000	\$48 000
\$50 000	\$50 000
\$52 000	\$52 000
\$56 000	\$56 000
μ = \$50 000	\$250 000
	μ = \$83 333





Measures of Variability

- Convey the degree to which scores are spread out and dispersed around a central point
  - ➙ Range
  - ➙ Mean Deviation
  - ➙ The Variance
  - ➙ The Standard Deviation

Range

- Overall span of the scores in a distribution – from the lowest value up to the highest value

- Range

$$Range = X_H - X_L$$

X_H = Highest score in the distribution

X_L = Lowest score in the distribution

The Range - Example

- What is the range of this distribution?

17, 44, 50, 23, 42

Range = 50 – 17 = 33

The Interquartile Range and Semi-Interquartile Range

- Every distribution can be divided into four equal sections or quartiles.
- A quartile is one-fourth of a distribution of scores.
- The bottom 25% of the values in a distribution make up the first quartile.

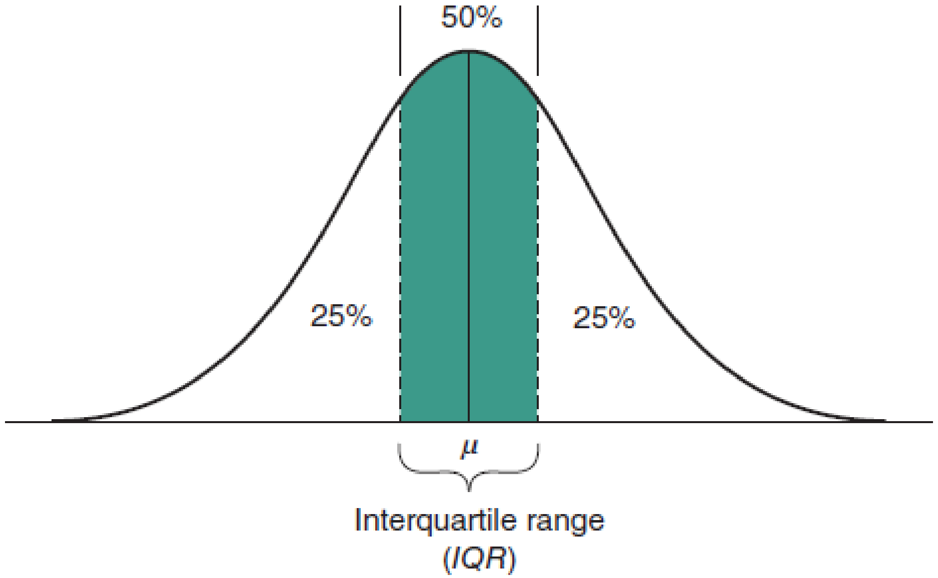
The Interquartile Range and Semi-Interquartile Range

- Interquartile range, IQR

$$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$$

Q_3 = the third quartile (75th percentile)

Q_1 = the first quartile (25th percentile)



- Semi-interquartile range, SIQR

$$SIQR = \frac{\mathrm{Q}_3 - \mathrm{Q}_1}{2}$$

Mean Deviation

- The degree to which scores deviate from the mean

$$MD = \frac{\Sigma|X - \mu|}{N}$$

X = a raw score

μ = the population mean

N = the number of scores in the population

Mean Deviation – Example (1)

- Consider these two distributions:

  Distribution A: 11, 12, 13, 14, 15, 16, 17

  Distribution B: 5, 8, 11, 14, 17, 20, 23

$$\mu = 14$$

Mean Deviation – Example (2)

Distribution A				Distribution B			
Scores	μ	$(X-\mu)$	$\|X-\mu\|$	Scores	μ	$(X-\mu)$	$\|X-\mu\|$
11	14	–3	3	5	14	– 9	9
12	14	–2	2	8	14	– 6	6
13	14	–1	1	11	14	– 3	3
14	14	0	0	14	14	0	0
15	14	1	1	17	14	3	3
16	14	2	2	20	14	6	6
17	14	3	3	23	14	9	9
$N = 7$			$\sum\|X-\mu\| = 12$	$N = 7$			$\sum\|X-\mu\| = 36$

$$MD_A = \frac{\Sigma|X-\mu|}{N} = \frac{12}{7} = 1.71 \quad MD_B = \frac{\Sigma|X-\mu|}{N} = \frac{36}{7} = 5.14$$

The Variance

- **Variance**

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

$\sigma^2 =$ the symbol for the population variance

X = a raw score

μ = the population mean

N = the number of scores in the population

The Variance - Example

- What is the variance of this sample of scores?

3, 4, 6, 8, 9

X	M	$X - M$	$(X - M)^2$
3	6	–3	9
4	6	–2	4
6	6	0	0
8	6	2	4
9	6	3	9
0	0	0	26

$$s^2 = \frac{\sum(X-M)^2}{n-1} = \frac{26}{4} = \mathbf{6.5} \; \blacksquare$$

The Standard Deviation

- Variance is a squared value
  - ➞ not stated in the original units of the measured variable
- Standard deviation
  - ➞ the square root of the variance

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

$\sigma =$ the symbol for the standard deviation

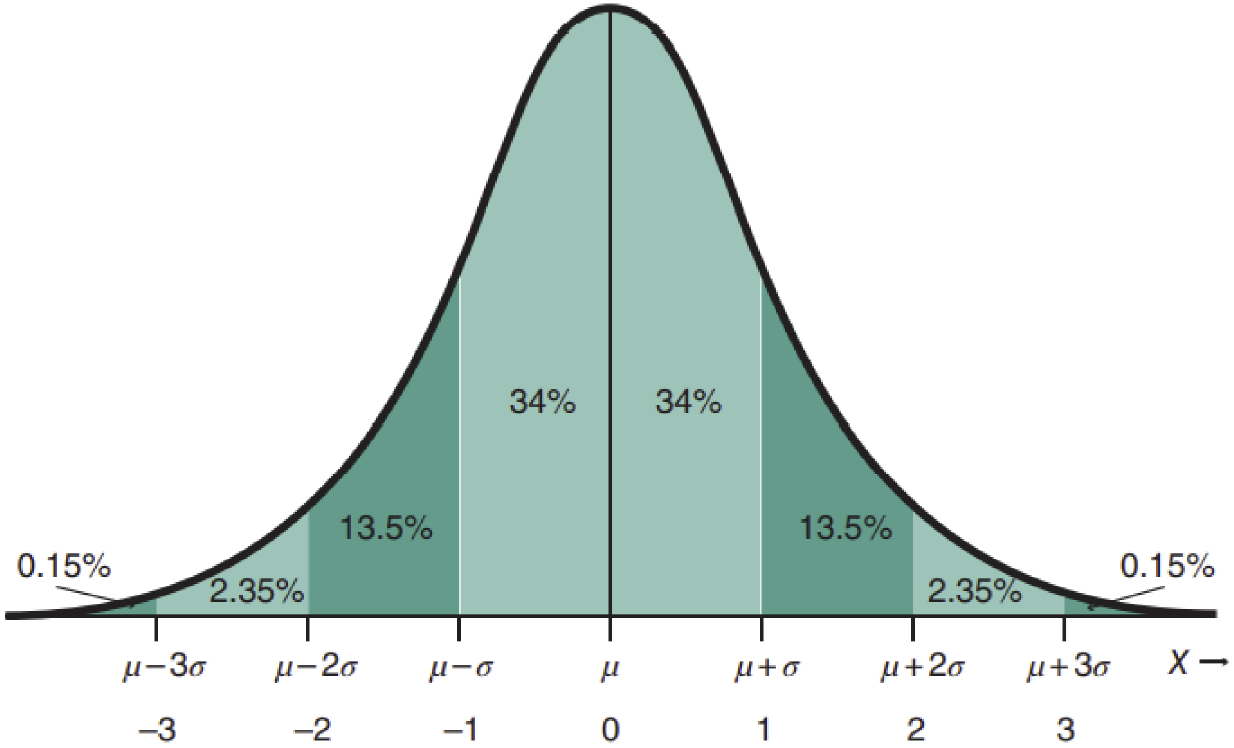
X = a raw score

μ = the population mean

N = the number of scores in the population

The Standard Deviation and the Normal Curve

- Approximately **68% (95%, 99%)** of the scores will fall between plus and minus **one (two, three)** standard deviation from the mean

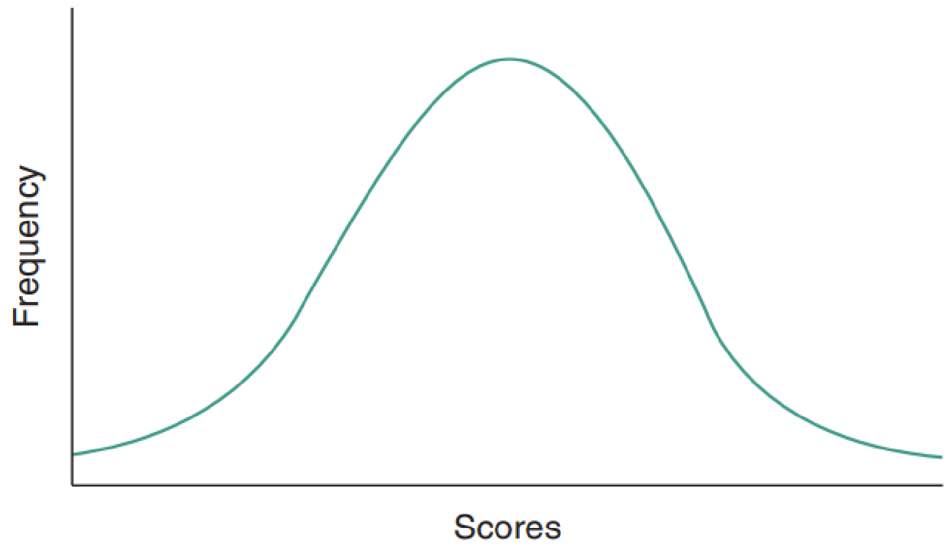


Deciding Which Measure of Variability to Use

- Range is most vulnerable to extreme scores.
- IQR and SIQR are not much influenced by a *small* number of extreme scores
- Variance and standard deviation are also affected by extreme scores, since squared deviations
- For a skewed distribution, IQR and SIQR best describe variability
- If the scale of measurement does not allow for the calculation of a mean, then deviation scores cannot be calculated.

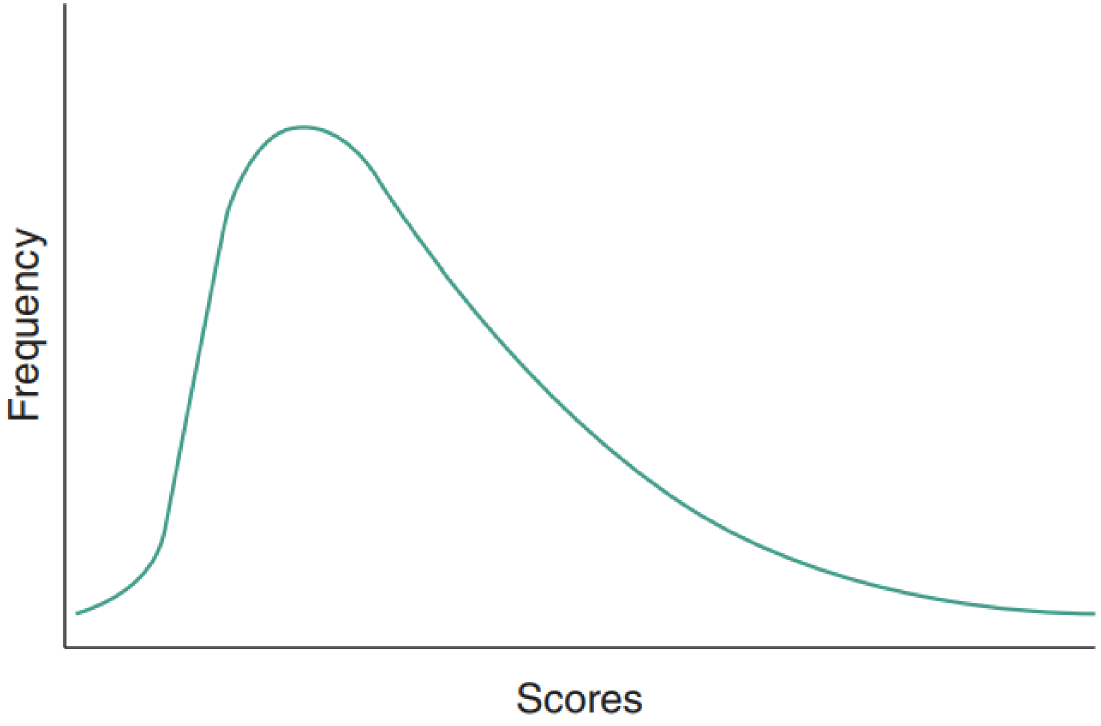
Measures of Shape Distribution

- "Deviation from Normal Distribution" described by
  - ➞Number of peaks
  - ➞Possession of Symmetry
  - ➞Tendency to Skew
  - ➞Uniformity

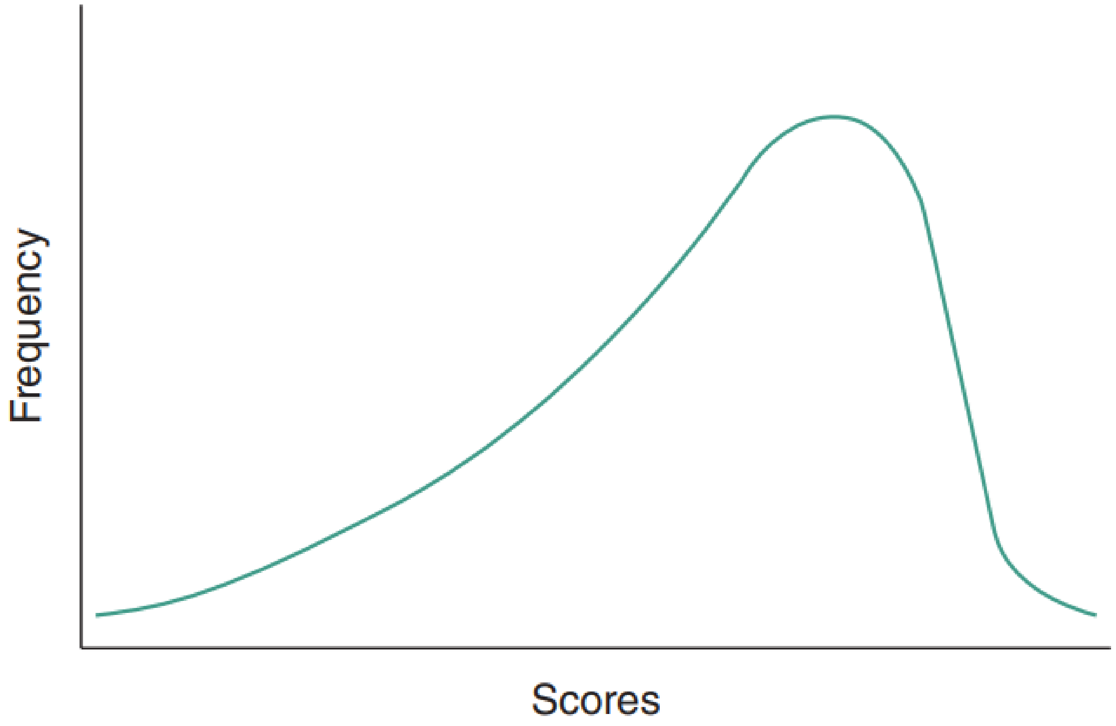


Skewed Distribution

- Most of the scores near one end of a distribution.



A positively skewed distribution



A negatively skewed distribution

Skewed Distribution

- For univariate data $X_1, X_2, \dots, X_N$, the formula for skewness is:

$$g_1 = \frac{\sum_{i=1}^{N}(X_i - \mu)^3/N}{s^3}$$

s = standard deviation of the distribution

- The above formula for skewness is referred to as the Fisher-Pearson coefficient of skewness

Kurtosis

- The quality of the peak of the curve
- **Leptokurtic** – narrow and accentuated peak
- **Platykurtic** – broad and muted peak

Kurtosis

WARWICK

- For univariate data $X_1, X_2, \dots, X_N$, the formula for kurtosis is:

$$g_1 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{X_i} - \mu)^4 / \mathrm{N}}{\mathrm{s}^4}$$

s = standard deviation of the distribution

Correlation

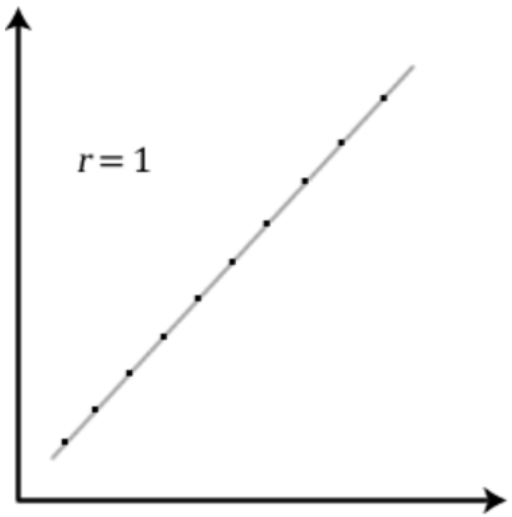
- Measure of the strength of association between two variables.
- Correlational analyses are often applied to two variables or scores – bivariate distribution
- A correlation coefficient can range from −1 to +1.

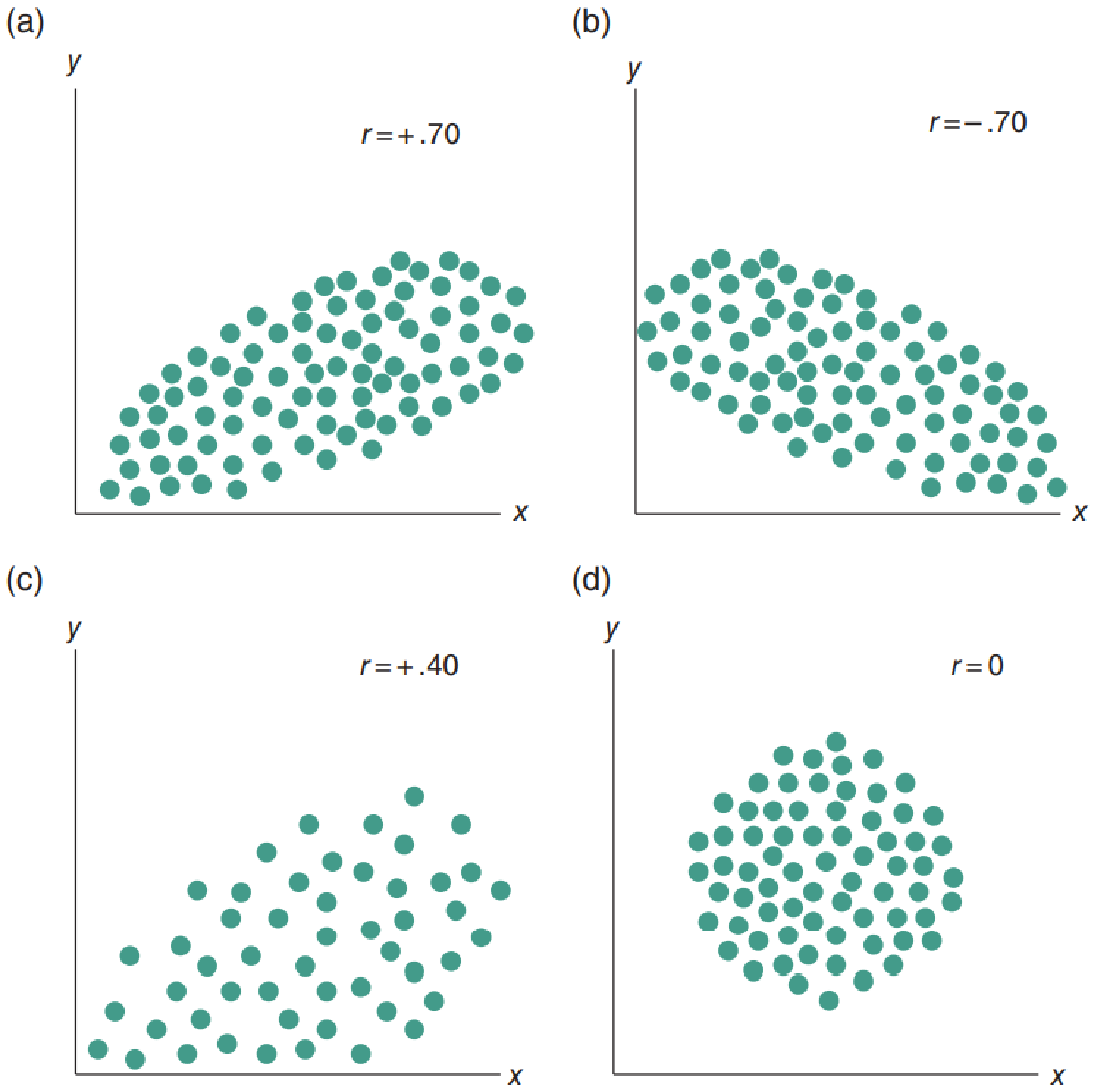
Correlation

WARWICK

- Larger the absolute value of the correlation → stronger the association
- Correlation coefficient → 0, weaker relationship between variables.
- A positive sign indicates a positive relationship
- A negative sign indicates a negative relationship

Correlation





Correlation

- Four types of correlations:
  - ➞ Pearson correlation
  - ➞ Kendall rank correlation
  - ➞ Spearman correlation
  - ➞ Point-Biserial

Pearson Correlation

- Most powerful and most frequently used version of the correlation measure

$$\rho = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\Sigma(x_i - \mu_x)^2 \Sigma\left(y_i - \mu_y\right)^2}}$$

ρ = correlation coefficient

x_i = values of the x variable in a sample

y_i = values of the y variable in a sample

μ_x = mean of x values

μ_y = mean of y values

Kendall Rank τ

- Kendall rank correlation is a non-parametric measure of relationship between two ranked variables.

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

$\tau = \text{Kendall rank } \tau$

n_c = number of concordant pairs
n_d = number of discordant pairs

https://www.statisticshowto.com/kendalls-tau/

Spearman Rank Correlation

- Spearman rank correlation is a non-parametric measure of relationship between two ranked variables.
- Suited for correlation analysis of variables on ordinal scale.

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

ρ = Spearman Rank Correlation

d_i = difference between the ranks of corresponding pairs
n = number of observations

https://www.youtube.com/watch?v=DE58QuNKA-c

Spearman Rank Correlation

- Assumptions
  - ➞ data must be at least ordinal
  - ➞ the scores are monotonically related