MSIN0017 Business Analytics

Lecture 1

Introduction

Dr Yufei Huang

Course Teacher

Dr Yufei Huang

- **Appointments**: Associate Professor, Trinity College Dublin (2018~)
  Assistant Professor, University of Bath (2016-2018)
  Honorary Research Associate, UCL (2016~)

- **Education:** PhD in Management (2016), UCL School of Management, UK
  MS in Physics (2010), Xi'an Jiaotong University, China
  BBA in Marketing (2005), Xi'an Jiaotong University, China

- **Teaching:**@Trinity: BU1550 Information Systems and Data Management, BU7582 Research Methods
- @Bath: MN50482 Supply Management, MN50205 Project Management
  MN50166 Research Method, MN50550 Business Analytics
  MN50637 Global Supply Chain and Logistics Management
  @UCL: MSIN0017 Business Analytics, MSIN0110 Big Data Analytics

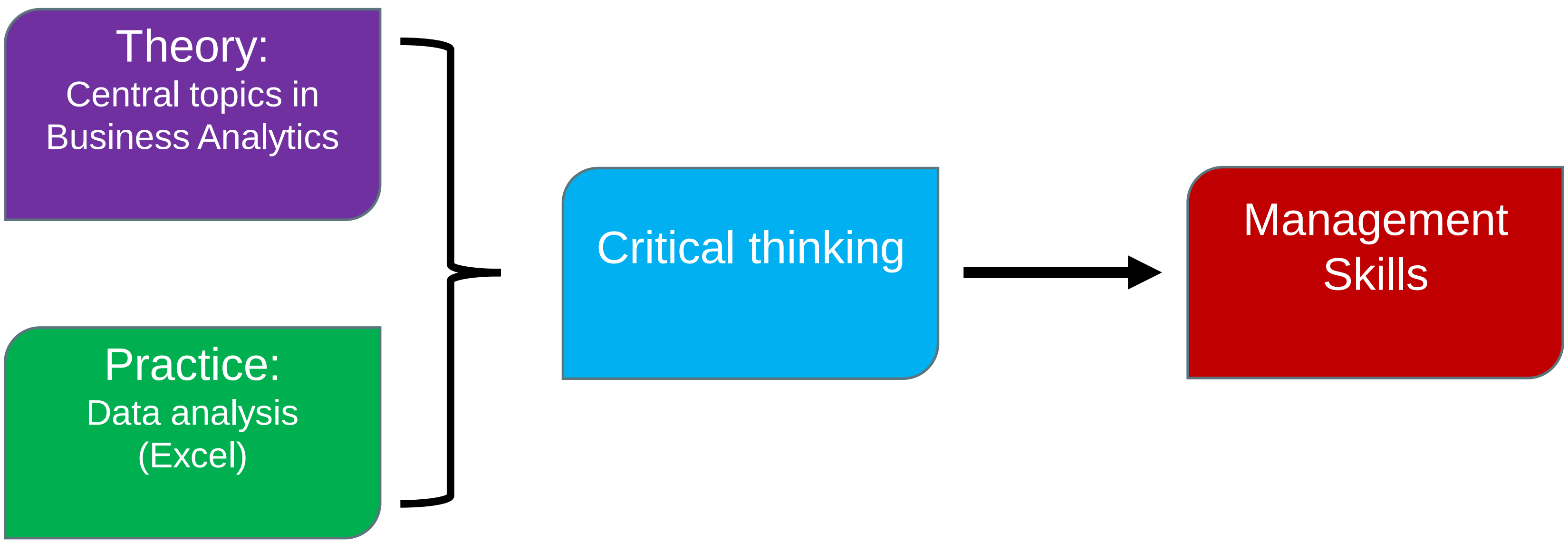
- **Research:** New Product Development and Introduction, Supply Chain Management, Quantitative Marketing

UCL SCHOOL OF MANAGEMENT

Content

- About the course
- Introduction to Business Analytics
- Understanding the data:
  - Central tendency
  - Spread

Course Goals



Course Structure

- 10 lectures (on Mondays, 16:00-18:00, Medical Sciences and Anatomy Anatomy G29 J Z Young LT)

- 10 seminars (on Thursdays)
  - There are 6 seminar groups, please go to your own seminar

- Please check timetable regularly for time and location changes https://timetable.ucl.ac.uk/tt/moduleTimet.do?firstReq=Y&moduleId=MSIN0017

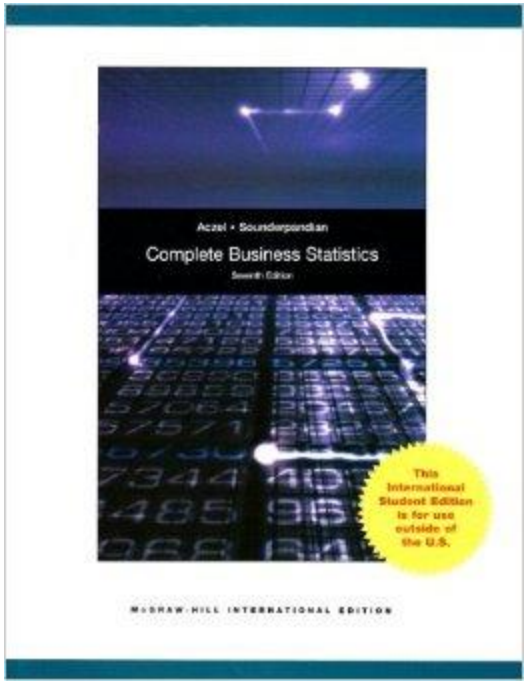
- Whenever needed, lectures will start with the required mathematical background.

- Please bring your laptop to the lectures and seminars

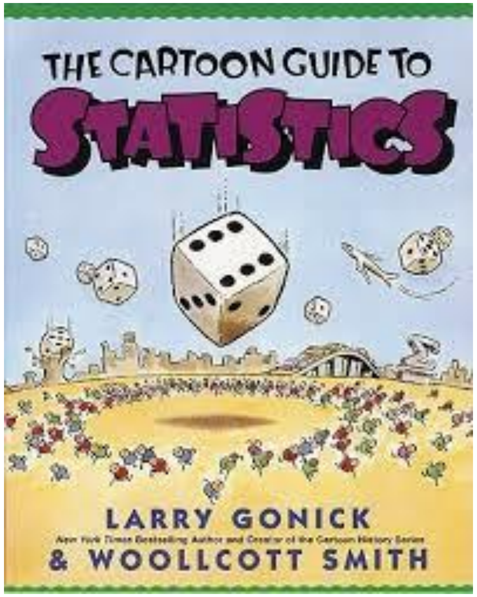
Syllabus (subject to minor changes along the way)

1. *Introduction to Business Analytics*
2. *Probability Theory*
3. *Random Variables*
4. *Normal Distribution*
5. *Sampling, Central Limit Theorem, and Confidence Intervals*
6. *Hypothesis Testing*
7. *Simple Linear Regression*
8. *Multidimensional and Nonlinear Regression*
9. *Revision*
10. *Application of Business Analytics and Summary*

Textbook



Complete Business Statistics



The Cartoon Guide to Statistics

UCL SCHOOL OF MANAGEMENT

Software

- This module uses Microsoft Excel for examples, exercises and coursework
- Excel tutorial will be provided during lectures or seminars whenever needed

Seminar Teachers and TAs

Ms Zejing Shao, PhD student, UCL Stats Dept.
Ms Chiara Cecilia Maiocchi, PostDoc in Math, University of Reading

Seminar Content:

- Emphasize important points from the lecture
- Provide more exercises
- Q&A

Assessment

- 80% is awarded on the basis of your examination result of an unseen 2-hour exam.

- 20% is awarded for coursework.
  - There will be 2 coursework submissions (10% and 10%)
  - Deadline for Coursework 1: 10/11/2023
    Deadline for Coursework 2: 08/12/2023
  - There will be one question from the coursework after each lecture.
  - Please start working on the question during the week.
  - Combine your solutions to form a coursework report, then submit

Coursework Submission

- Submit one single PDF file containing answers to coursework.
- Briefly explain your results.
- You can include figures or tables in your report.
- This is not group work, finish and submit report by yourself.
- Do not exceed 10 pages.
- Do not submit Excel file.
- Do not submit multiple files.

Final Exam

- 2-hour unseen final exam in Term 3 2024
  (Time & Format TBC: likely to be in-person exam)

- You can bring a calculator (check UCL regulation)
  http://www.ucl.ac.uk/current-students/exams_and_awards/regulations/candidate_guide.pdf

- You can bring 1 piece of A-4 paper, and write whatever you want on it
  - Double-sided if you need
  - Print if you want

- You can NOT use laptop, smart phone, textbook, lecture slides, seminar slides, your own notes, etc.

Additional Help

Contact us:

- Yufei Huang: yufei.huang.10@ucl.ac.uk
- Zejing Shao: zejing.shao.15@ucl.ac.uk
- Chiara Cecilia Maiocchi: c.maiocchi@ucl.ac.uk

Introduction to Business Analytics

What is uncertainty?

- Cambridge dictionary:

  a situation in which something is not known, or something that is not known or certain

- In simple words: we don't know exactly what is going to happen

- People attempt to try to interpret an uncertain world using mathematical tools (what we will learn)

Why are Probability, Statistics and Business Analytics important?

- Facing uncertainty, your intuition sometimes is wrong!

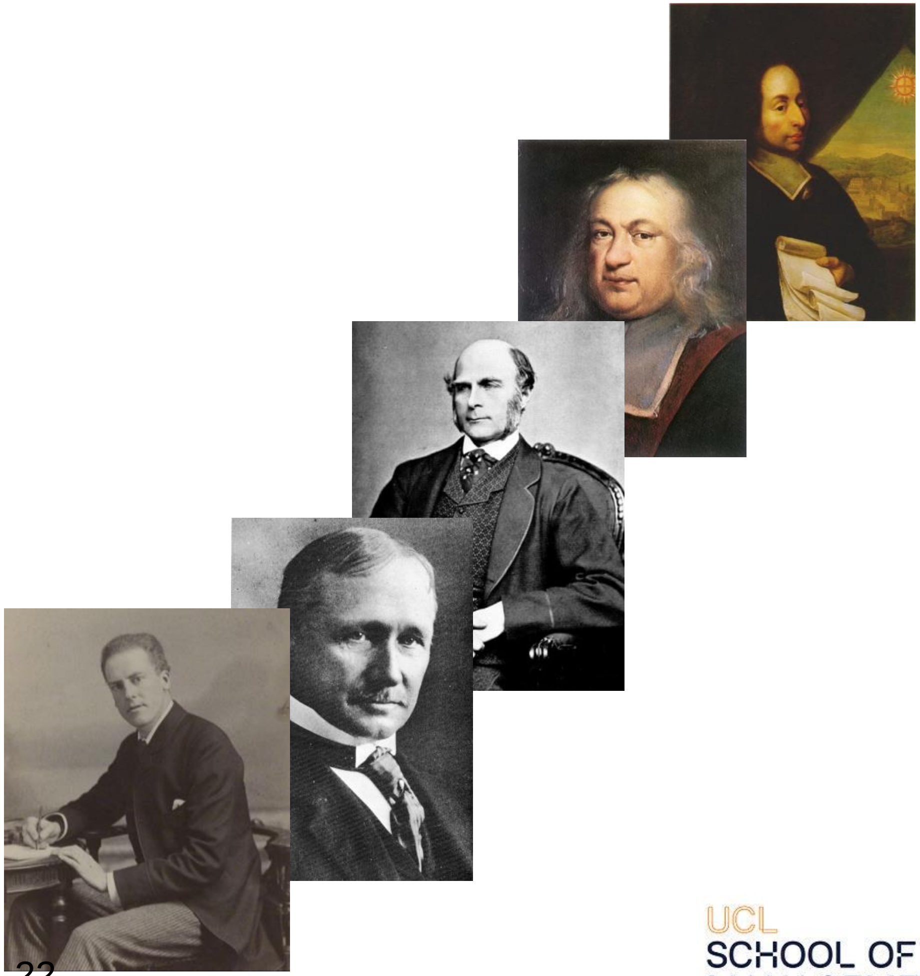
- Probability, Statistics and Business Analytics help handle data
  - Data Collection
  - Data Analysis
  - Data Interpretation

- Thereby support judgement and decision making

History of Business Analytics

- Pascal 1623-1662

  Fermat 1601-1665
  - Mean, expectation
- Galton 1822-1911
  - Regression
  - Correlation
- Taylor 1856-1915
  - Business analytics
- Pearson 1857-1936
  - Standard deviation,
  - Hypothesis testing and p values
  - Established the first Statistics department in the world at UCL(!!!)



Example

- Imagine that you are a product manager of a software company in the UK. You are going to launch a new App in the market. You have got some data* after conducting product trial.

- What can you infer from the following data table?

Example: Product Trial Data

Participant NO.	Product Trial Rating	Willingness to Buy	Previous Experience	Gender	International	Age
1	84	54	N	M	I	32
2	80	69	N	F	D	21
3	71	47	Y	F	I	33
4	65	48	N	M	D	55
5	64	74	Y	M	D	36
6	62	41	N	F	D	21
7	84	62	N	M	D	37
8	73	69	Y	F	I	59
9	71	64	N	F	I	31
10	71	79	N	F	I	17

* Disclaimer: The data is randomly generated by the lecturer, and is only used as a demonstration example. Therefore the conclusions from the data neither represent the reality nor indicate the lecturer’s own opinion.

Key Measures

- Measures of central tendency
  - Mean/average
  - Median
  - Mode

- Measures of dispersion/spread of a sample
  - Range
  - Variance
  - Standard deviation

Measures of Central Tendency: Mean

The mean of N measurements $x_1, \ldots, x_N$ is given by:

$$m = \bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Example: Product Trial Data

- The mean product trial rating is:

$$m_{\text{rating}} = \frac{84+80+71+65+64+62+84+73+71+71}{10} = 72.5$$

- The mean willingness to buy:

$$m_{\text{will}} = \frac{54+69+47+48+74+41+62+69+64+79}{10} = 60.7$$

Participant NO.	Product Trial Rating	Willingness to Buy
1	84	54
2	80	69
3	71	47
4	65	48
5	64	74
6	62	41
7	84	62
8	73	69
9	71	64
10	71	79

Understanding the Data: Gender

- Mean product trial rating for female:

$$m_{r,f} = \frac{80+71+62+73+71+71}{6} = 71.3$$

- Mean willingness to buy for female:

$$m_{w,f} = \frac{69+47+41+69+64+79}{6} = 61.5$$

- Mean product trial rating for male:

$$m_{r,m} = \frac{84+65+64+84}{4} = 74.25$$

- Mean willingness to buy for male :

$$m_{w,m} = \frac{54+48+74+62}{4} = 59.5$$

Participant NO.	Product Trial Rating	Willingness to Buy	Gender
1	84	54	M
2	80	69	F
3	71	47	F
4	65	48	M
5	64	74	M
6	62	41	F
7	84	62	M
8	73	69	F
9	71	64	F
10	71	79	F

Results So Far

- Men in our sample give higher rating than women for the trial product, but the mean willingness to buy for men tends to be lower than women.

- We can do similar analysis for other variables, such as “Previous Experience”, “age” and “international”.

Understanding the Data: International

- What are the mean product trial rating for international and domestic participants?

  74/71

- What are the mean willingness to buy for domestic and international participants?

  62.6 / 58.8

- What conclusions can you draw?

Participant NO.	Product Trial Rating	Willingness to Buy	International
1	84	54	I
2	80	69	D
3	71	47	I
4	65	48	D
5	64	74	D
6	62	41	D
7	84	62	D
8	73	69	I
9	71	64	I
10	71	79	I

Measures of Central Tendency: Median

Age	Age, sorted
32	17
21	21
33	21
55	31
36	32
21	33
37	36
59	37
31	55
17	59

- The median of a sample is the data point below which lie half of the data in the sample.
- To calculate it:
  1. Sort the data according to its order
  2. If there is an odd number of points, choose the middle data point
  3. If there is an even number, choose the mean of the two middle values.

Example: the median age of our sample is:

$$\text{median} = \frac{32+33}{2} = 32.5$$

Measures of Central Tendency: Mode

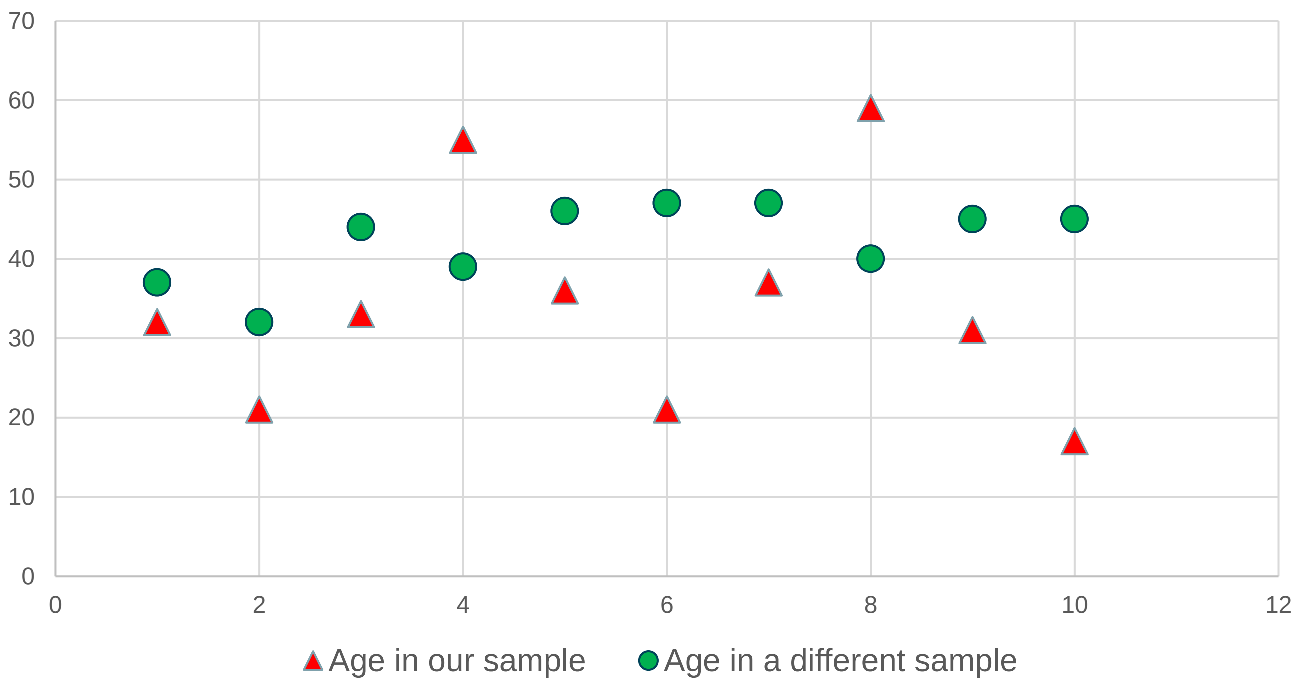
- Mode is the most frequent value: the value that appears the largest number of times in our sample (if there are two modes, we can refer to the first one)

- Example: the mode for product trial rating is 71.

Product Trial Rating
84
80
71
65
64
62
84
73
71
71

Measures of Spread

- Example: ages in our sample seem very different from another sample
- We see that the spread of age in our sample is larger than that in the second sample
- How can we characterize spread?



Age in our sample	Age in a different sample
32	37
21	32
33	44
55	39
36	46
21	47
37	47
59	40
31	45
17	45

Measures of Spread: Range

- The range of a sample of N elements,

  $$x_1, x_2, \ldots, x_N$$

  is the difference between the largest and smallest data value:

  $$\text{Range} = \max(\{x_1, \ldots, x_N\}) - \min(\{x_1, \ldots, x_N\})$$

- Example:

  The range of ages in our sample:

  59-17=42

  The range of ages in the other sample:

  47-32=15.

Age in our sample	Age in a different sample
32	37
21	32
33	44
55	39
36	46
21	47
37	47
59	40
31	45
17	45

Measures of Spread: Variance

- The variance of a sample of N elements, $X_1, X_2, \ldots, X_N$ with mean m is given by:

$$s^2 = \frac{1}{N-1}\left[(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_N - m)^2\right] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2$$

- The variance of the population of N elements, $X_1, X_2, \ldots, X_N$ with mean m is given by:

$$\sigma^2 = \frac{1}{N}\left[(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_N - m)^2\right] = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2$$

- It gives information about the extent to which the measurements are different than its mean and how spread they are.

- We usually use the formula for a sample, as the data for a whole population is difficult to obtain

UCL SCHOOL OF MANAGEMENT

Measures of Spread: Standard Deviation

- The standard deviation of a sample is the square root of its variance:

$$s = \sqrt{\frac{1}{N-1}\left[(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_N - m)^2\right]}$$

- The standard deviation of the population is the square root of its variance:

$$\sigma = \sqrt{\frac{1}{N}\left[(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_N - m)^2\right]}$$

Calculation of Variance: Example

No.	Data Series 1	Data Series 2
1	23	32
2	48	33
3	35	31
4	37	32.5
5	21	31.5

- The mean of Data Series 1 is:

$$m_1 = \frac{1}{5}(23 + 48 + 35 + 37 + 21) = 32.8$$

- The variance of Data Series 1 is:

$$s_1^{\,2} = \frac{1}{4}\left[(23-32.8)^2 + (48-32.8)^2 + (35-32.8)^2 + (37-32.8)^2 + (21-32.8)^2\right]$$

$$= \frac{1}{4}\left(96.04 + 231.04 + 4.84 + 17.64 + 139.24\right) = 122.2$$

- The mean of Data Series 2 is:

$$m_2 = \frac{1}{5}(32 + 33 + 31 + 32.5 + 31.5) = 32$$

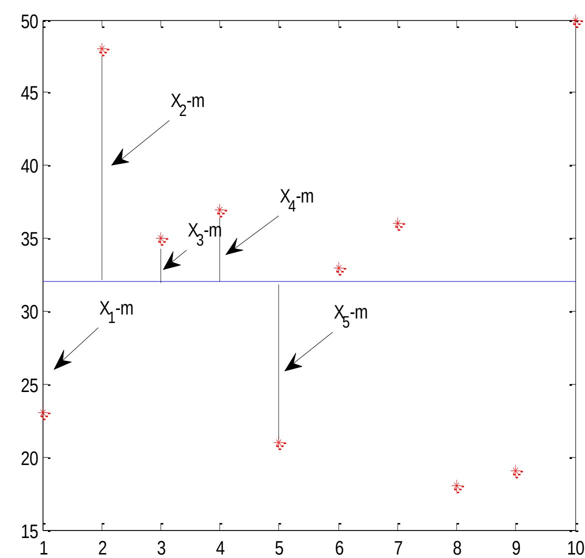
- The variance of Data Series 2 is:

$$s_2^{\,2} = \frac{1}{4}\left[(32-32)^2 + (33-32)^2 + (31-32)^2 + (32.5-32)^2 + (31.5-32)^2\right]$$

$$= \frac{1}{4}[0 + 1 + 1 + 0.25 + 0.25] = 0.625$$

- The standard deviation is the square root of its variance:

$s_1 = 11.05 \quad s_2 = 0.79$



Application: Return-to-Risk Ratio

- Return-to-risk ratio is defined as: ***Return-to-Risk Ratio= Return / sd***, where:
  - Return is a profit of an investment (change of value)
  - sd is the standard deviation of the sample

- Example: if the expected return of a company is 25%, and the standard deviation of the return is 12.5, then the return-to-risk ratio is 25/12.5=2.

Example:

The historical returns of a high-tech company are given in the following table.

Year	1	2	3	4
Returns	20%	10%	30%	20%

Assume that the expected return for year 5 is the average return.

- Calculate the expected return for year 5
- Calculate the variance
- Calculate the return-to-risk ratio

Solution

Year	1	2	3	4
Returns	20%	10%	30%	20%

- The mean return is:

$$\bar{x} = \frac{20+10+30+20}{4} = \frac{80}{4} = 20.$$

- The Variance is:

$$S^2 = \frac{1}{4-1}\left[(20-20)^2 + (10-20)^2 + (30-20)^2 + (20-20)^2\right]$$

$$= \frac{1}{3}\left(0^2 + 10^2 + 10^2 + 0^2\right) = \frac{200}{3} = 66.6667$$

- The standard deviation is:

$$\text{sd} = \sqrt{S^2} = \sqrt{\frac{200}{3}} = 10\sqrt{\frac{2}{3}} = 8.165$$

- The return-to-risk ratio is:

$$\frac{20}{8.165} = 2.4495$$

Dimensionless Measure

- Can the units of the measurement affect the mean and standard deviation?

  Yes.

- Can you think about an example?

- What can we do in order to get a measure of dispersion which is independent of the units of the measurement?

  Coefficient of Variation (CV)= sd / mean

- CV is the inverse of Return-to-Risk Ratio

Example:

The historical returns of a high-tech company are given in the following table.

Year	1	2	3	4
Returns	20%	10%	30%	20%

Assume that the expected return for year 5 is the average return.

- Calculate coefficient of variation

  The mean return is: 20%

  The standard deviation is: 8.165%

  The coefficient of variation is: CV=s.d./mean=0.408

Reference

Chapter 1 of:

Aczel, A., & J. Sounderpandian. 2008. Complete Business Statistics. McGraw-Hill/Irwin, Seventh Edition.

Mathematical background: the ∑ notation

- The sum $a_1 + a_2 + \ldots + a_n$ can be written using the sigma notation:

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \ldots + a_n$$

- Examples:

1. $1^2 + 2^2 + 3^2 + \ldots + 100^2 = \sum_{i=1}^{100} i^2$
2. $3^2 + 4^2 + 5^2 + \ldots + 100^2 = \sum_{i=3}^{100} i^2$
3. $R + R + R + R + R = \sum_{i=1}^{5} R$
4. $R + 2R + 3R + 4R + 5R = \sum_{i=1}^{5} iR$

Question

How can you write, using the ∑ notation:

1. $\dfrac{x_1 + x_2 + \ldots + x_N}{N}$?

$$\frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}(x_1 + x_2 + \ldots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

2. $\dfrac{1}{N-1}\left[(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_N - m)^2\right]$?

$$\frac{1}{N-1}\left[(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_N - m)^2\right] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2$$

Rules of the $\sum$ notation

1. If c is a constant then $\sum_{i=1}^{n} c = nc.$

Example. $\sum_{i=1}^{5} R = R + R + R + R + R = 5R.$

2. If c is a constant then $\sum_{i=1}^{n} ca_i = c\sum_{i=1}^{n} a_i$

Example. $\sum_{i=1}^{5} Ri = R\cdot 1 + R\cdot 2 + R\cdot 3 + R\cdot 4 + R\cdot 5 = R(1+2+3+4+5) = R\sum_{i=1}^{5} i$

3. $\sum_{i=1}^{n} a_i \pm b_i = \sum_{i=1}^{n} a_i \pm \sum_{i=1}^{n} b_i.$

4. $\sum_{i=1}^{n} a_i = \sum_{k=1}^{n} a_k$