Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

BU7142 Foundations for Business Analytics

Lecture 1

Introduction, Central Tendency and Spread Measure

Dr Yufei Huang

Contact

Yufei Huang

- Email: yufei.huang@tcd.ie
- Tel: +353 1896 8228
- Office Hour: **by appointment**

Content

- About the course
- Understanding the data:
  - Central tendency
  - Spread
- Seminar

Course Structure

- 5 days, 6 sessions
- lecture + seminar
- Please bring your laptop

Topics Covered (subject to minor changes along the way)

1. *Introduction to Business Analytics*
2. *Set Theory Probability Theory*
3. *Discrete and Continuous Random variables*
4. *Normal Distribution*
5. *Sampling, Central Limit Theorem, and Confidence Intervals*
6. *Hypothesis Testing*
7. *Simple Linear Regression*
8. *ANOVA Multi-variable Regression and Beyond*

Recommended Books

Aczel, A.D., Sounderpandian, J., Patille, L. Complete Business Statistics, 7th Edition. McGraw-Hill, Irwin.

Healey, J.F. (2015). Statistics: A Tool for Social Research. 10th Edition. Stanford: Cengage Learning. ISBN: 978-1-285.45885-4.

Miles, J. and Field, Z. (2012). Discovering Statistics Using R. London: Sage Publications.

Grolemund, G. and Wickham, H. (2016). R for Data Science. O'Reilly. http://r4ds.had.co.nz/ Creative Commons.

Software

- This module uses Microsoft Excel to demonstrate examples, exercises and case studies during the lectures

- You are free to use any other software for assignments.

Assessment

- Individual Assignment: 50%
  Problem sets and case studies.

- Group Assignment: 50%
  Data analysis project

- Both assignments will be provided at a later stage during the module

Forming Groups using Google Doc

- Ideally, find your group on 12 Sep, but no later than 16 Sep.
- 4-5 Students, Max 6
- Name and student no. of all members
- Link:

https://docs.google.com/spreadsheets/d/13NOMmmCuqb_dIKWFnXtdLt99KxbFEjZ9U6kOv3DPjiM/edit?usp=sharing

Using Google Docs

- Google Doc Important Notes
  - Please find your group members and form your groups first, then go to google doc to fill in the form
  - When using google doc, please only enter the information for your own group
  - Do NOT change anything else in the form, especially the information about other groups
  - We can track the changes, if we find any misuse, we reserve the right to stop using google doc, and assign groups randomly

# (Un)Equal Contribution

- All students within each group are expected to contribute equally to the group project.
- In the event that students want to report that contributions to groupwork have been unequal,
- inform the lecturer via e-mail (cc all group members) at least one week in advance of the submission deadline of the group project.
- wherever possible, the students should agree between themselves in a friendly manner what they feel to be the most appropriate variations in individual marks based entirely upon individual engagement with the work process.
- Members of the group should then indicate the percentage contribution to the work process that each person has made on the first page of the report. All students need to sign.
- If students cannot reach an agreement, students will need to meet with the lecturer to present their case. The proposed percentage contribution must be supported by verifiable evidence, for example notes of work carried out, drafts and minutes of meetings.
- The final decision with regards to the unequal contributions for a group will be determined by the unit lecturer in close collaboration with the Director(s) of Studies of the relevant programme(s).

# (Un)Equal Contribution

Some additional information:

- A Team Agreement Form is provided on Blackboard to help the groups plan for their group assignment (This form is not compulsory and will not be assessed, but we advise you to use the form to plan for your group assignment.)

- If an agreement on contribution can not be reached within the group. We may ask each group member fill in a Peer Review Form so that we can help you decide the contribution of each group member

Plagiarism

- **Make sure you know 'THE RULES'**

- The Oxford English Dictionary defines PLAGIARISM as:

  "The action or practice of taking someone else's work, idea, etc., and passing it off as one's own; literary theft"

- **Plagiarism** arises from poor academic practice – for example, copying information, text or diagrams from an unacknowledged source, without quotation marks or any indication that the presenter is not the original author or researcher.

- **Plagiarism Detection** - the University uses the electronic detection software **TURNITIN** to monitor all students' work – this is a very powerful and flexible tool which checks text based submissions against available, online resources, and its own database of previously submitted work, and will highlight any matching text found.

- **Plagiarism Offences** - Plagiarism is a serious academic offence. Any student found to have used unfair means, such as plagiarism, in any examination or assessment procedure may be penalised. Proven cases of plagiarism may also lead to disciplinary proceedings

Introduction, Central Tendency and Spread Measure

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

  https://www.youtube.com/watch?v=6HJqPZ-KmZs

- 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!

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:

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

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

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

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

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

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

Example: Product Trial Data

- The mean product trial rating is:

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

- The mean willingness to buy:

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

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

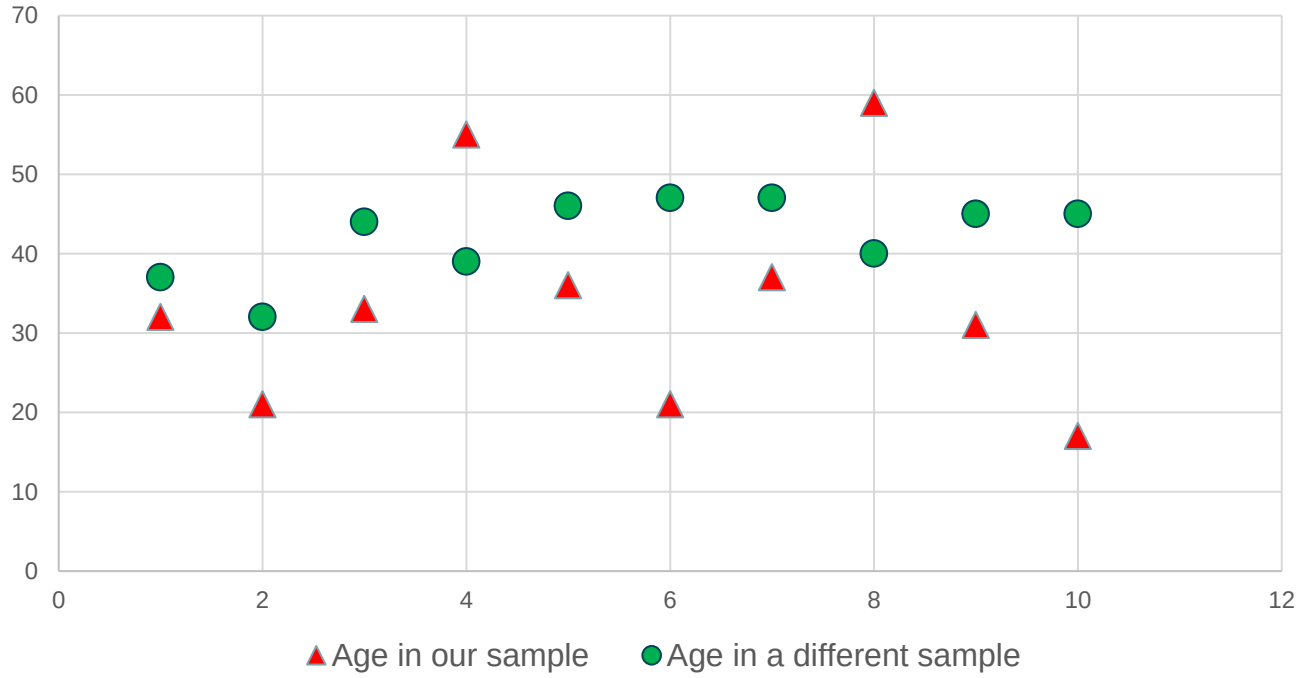
- Mode is the most frequent value: the value that appears the largest number of times in our sample (if there are two modes, we can refer to the first one)

- 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

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:

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

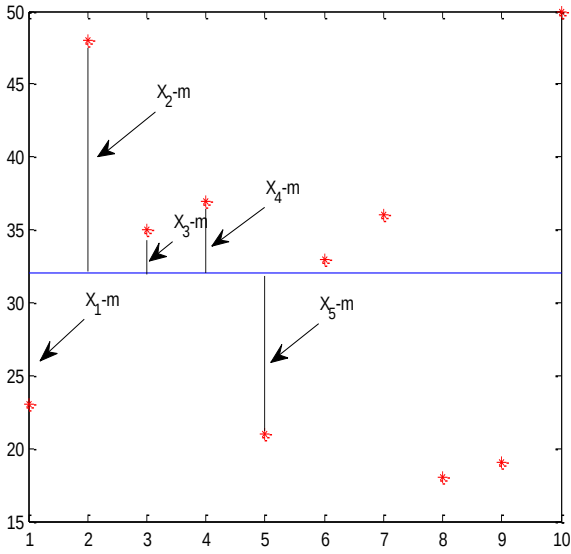
- The variance of Data Series 2 is:

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

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

- 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

Seminar -Exercises

Exercise 1: employment in fashion industry in the US, 1990-2008

The data

Employment in apparel manufacturing and component industries, 1990–2008

Year	Apparel manufacturing, total	Cut and sew apparel contractors	Men's and boys' cut and sew apparel	Women's and all other cut and sew apparel	All other apparel manufacturing
1990	902,800	268,600	234,400	246,700	153,200
1991	876,900	262,600	228,400	237,400	148,500
1992	879,300	262,400	230,700	233,200	153,000
1993	857,300	253,000	226,700	223,800	153,900
1994	831,900	245,300	219,700	214,400	152,500
1995	791,100	233,300	207,100	202,400	148,200
1996	722,300	214,500	185,200	184,200	138,500
1997	680,800	204,100	171,200	173,300	132,200
1998	621,400	186,800	155,100	156,100	123,300
1999	540,500	163,500	129,300	136,400	111,300
2000	483,500	146,600	111,600	122,000	103,300
2001	415,200	123,400	95,200	104,800	91,800
2002	350,000	106,400	74,200	92,600	76,800
2003	303,900	94,200	61,200	79,300	69,300
2004	278,000	88,300	54,600	70,600	64,500
2005	250,500	79,500	48,700	65,200	57,100
2006	232,400	74,600	43,100	60,700	54,000
2007	214,600	66,300	39,400	59,900	49,000
2008	199,000	60,700	37,400	57,300	43,600

Questions

- Q1: What was the mean of the total employment?
- Q2: To estimate the decrease in the employment, calculate the employment mean in the years1990-1993 and in the years 2005-2008.
- Q3: What was the median of the employment between 1990 and 1994?
- Q4: What was the median of the employment between 2005 and 2008?
- Q5: The range of employment level between 1990 and 1994 was?
- Q6: What was employment range between 2005 and 2008?
- Q7:Calculate the variance and standard deviation of employment level in the period 1990-1993.
- Q8:Calculate the variance and std. of the employment in 2005-2008.

Exercise 2

Fortune published a list of the 10 largest “green companies”—those that follow environmental policies. Their annual revenues, in $ billions, are given below.

Company	Revenue $ Billion
Honda	84.2
Continental Airlines	13.1
Suncor	13.6
Tesco	71.0
Alcan	23.6
PG&E	12.5
S.C. Johnson	7.0
Goldman Sachs	69.4
Swiss RE	24.0
Hewlett-Packard	91.7

Find the mean, variance, and standard deviation of the annual revenues.

Thank You!

Any Questions?