Ch 2. Summary statistics and elementary data presentation

Outline of lecture

1. Motivating quantitative methods
2. Location and scale
3. Regression-type problems
4. Data presentation
5. Exercises

1.1 Motivating quantitative methods

- Many students often think they are afraid of mathematics
- Had very good student results in the past e.g. 97+% pass rates on modules with nearly 200 students. Would like to see students continue to do well

1.2 The importance of quantitative research methods ...

• If using quantitative research methods it is important to recognise that the purposes of quantitative research methods are often deceptively simple-minded

- **Need a feel for why you are ultimately trying to use quantitative methods**

1. Data display
2. Measures of location and spread
3. Regression-type approaches
4. Models for categories and groups based on survey-type data
5. Specialist financial time series models (important differences between accounting and finance research and general business research)

• Can't really do items 3-5 onwards justice without proper (inferential/hypothesis testing) statistics

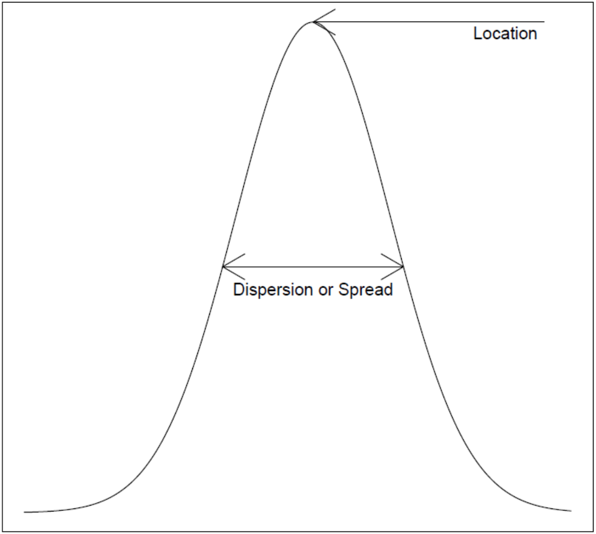
1.3 Core topics in quantitative methods

- Want to graphically motivate two things
  - Measures of locations and spread
  - Regression analysis and correlation

2.1 Two basic questions

- **Location**
  - What is a "typical" observation?
  - Mean, median, mode
- **Dispersion**
  - "How spread out is the data?"
  - Variance, standard deviation, inter-quartile range, range

2.2 Two basic questions



2.3 A quick note on skewness and kurtosis

• Skewness (asymmetry) and kurtosis (propensity for extreme values) are also important – **especially for finance** – see e.g. the popular books by Nicholas Taleb

• Methods involving skewness and kurtosis tend to be more difficult to apply and hence are inherently specialist

• **Disclaimer: Skewness and kurtosis will be of only background importance in this introductory course but will often be extremely important in financial problems in the real world!**

2.4 Mean

- The mean is the prototypical weighted average
- Simply add up all the observations and divide by the number of observations in the sample

$$\text{mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Numerical example.** If we have data $(19, 4, 3, 13, 10)$ calculate the mean as

$$\text{mean} = \bar{x} = \frac{19 + 4 + 3 + 13 + 10}{5 \quad \text{observations}} = \frac{49}{5} = 9.8$$

2.5 Median

- Median is the "mid-point" of the *ordered* data
- May often be a better summary of financial data than the mean

– will be less affected by extreme values!

$$\text{Median} = \frac{n+1}{2} \text{ordered data point}$$

- **Odd number of observations**
  - Median = "middle" data point
- **Even number of observations**
  - Median = Average of the two "middle" data points

2.6 Median – numerical example

- Example. Find the median of $(19, 4, 3, 13, 10)$
  1. Order the data from smallest to largest:
  $$(3, 4, 10, 13, 19)$$
  2. median $= (\frac{5+1}{2}) =$ 3rd data point $= 10$
- Example. Find the median of $(19, 4, 3, 13, 10, 19)$
  1. Order the data from smallest to largest:
  $$(3, 4, 10, 13, 19, 19)$$
  2. median $= (\frac{6+1}{2}) =$ 3.5th data point
  $$\text{median} = \frac{\text{3rd obs.} + \text{4th obs.}}{2} = \frac{10+13}{2} = 11.5$$

2.7 Mode

$$\text{Mode=Most commonly observed value}$$

- **WARNINGS!**
  - Mean and median are often better summaries
  - May not always exist unlike the mean and median
- **Numerical Example.**

Data

$(11, 14, 19, 18, 10, 13, 22, 18, 11, 14, 1, 12, 12, 18, 6, 12, 11, 19, 18, 2, 22, 18)$

$$\text{Mode} = \text{Most Common Value} = 18$$

2.8 Computation in Excel

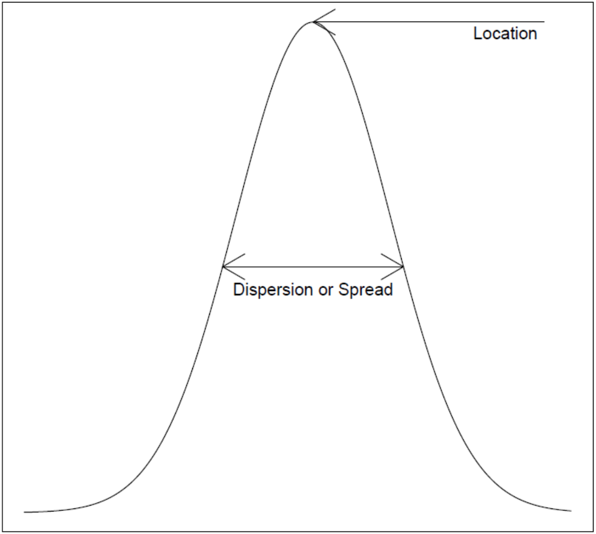
- Suppose you have data in cells $A1:A10$ and wish to calculate the mean.
  1. Click on the insert formula icon f_x
  2. Choose category *Statistical*
  3. Select *Average*. Press *OK*
  4. Enter $A1:A10$ in the box titled **Number 1**
- The median and mode can be calculated in exactly the same way as above using the commands *MEDIAN* and *MODE.SNGL*

2.9 A rule of thumb

- A useful rule of thumb, albeit not actually strictly technically correct is

$$\begin{aligned} \text{Mean} > \text{Median} &= \text{Positively skewed} \\ \text{Mean} < \text{Median} &= \text{Negatively skewed} \end{aligned}$$

2.10 Recall the following mental picture



2.11 Dispersion

• Want to measure the spread or dispersion of the data
• The most commonly used measure of *spread* or *dispersion* is the standard deviation or the variance

$$\begin{aligned} v &= \text{variance} = \text{standard deviation}^2 \\ s &= \text{standard deviation} = \sqrt{\text{variance}} \end{aligned}$$

• In practical examples use of the standard deviation may be preferred.
• E.g. if looking at weekly sales figures measured in units of £. The standard deviation will be measured in units of £and the variance will be measured in units of £^{2}
• **May thus be more convenient to communicate this information using the standard deviation**

$$v = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = s^2$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{v}$$

2.13 Variance and standard deviation – practical calculation formulae

- In practice, the formulae to use are

$$v = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}}$$

- **Numerical example.** Data $(19, 4, 3, 13, 10)$

$$\text{mean} = \bar{x} = \frac{19+4+3+13+10}{5} = \frac{49}{5} = 9.8$$

$$\sum_{i=1}^{n} x_i^2 = 19^2 + 4^2 + 3^2 + 13^2 + 10^2 = 655$$

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{655 - 5(9.8)^2}{4} = 43.7$$

$$\text{Standard Deviation} = \sqrt{Variance} = \sqrt{43.7} = 6.611$$

2.14 Quartiles and the inter-quartile range

• **The median is the 0.5 point or the half-way mark of the ordered data**

• We can also define other quartiles (and infinitely many other quantiles)

• The median is the second quartile and the 50 % quantile

$$\begin{aligned} \text{Lower Quartile} &= \text{0.25 point of the ordered data} \\ \text{Median} &= \text{0.5 point of the ordered data} \\ \text{Upper Quartile} &= \text{0.75 point of the ordered data} \end{aligned}$$

The inter-quartile range is the difference between the Upper and Lower Quartiles

$$\text{Inter-Quartile Range} = \text{Upper Quartile} - \text{Lower Quartile}$$

2.15 The inter-quartile range

- **The key formula to remember is simply**

$$\text{Inter-Quartile Range} = \text{Upper Quartile} - \text{Lower Quartile}$$

- The inter-quartile range is a measure of the dispersion or spread of the data
- If some of the data take extreme values the inter-quartile range may provide a more useful summary than the variance or standard deviation.
- **There are thus some occasions when the inter-quartile range may offer a better description of financial data than the variance or standard deviation**

2.16 Calculating the inter-quartile range

- **The key formula to remember is simply**

$$\text{Inter-Quartile Range} = \text{Upper Quartile} - \text{Lower Quartile}$$

- Calculating quartiles is deceptively involved. Many different formulas are given and the standard of business mathematics texts and other statistical study guides can be very varied.
- **On this course hand calculation of quartiles is not required**
- **It would be more important to ultimately be able to do this by computer**
- **Calculation of quartiles and the inter-quartile range can be deceptively involved, uses something called pivots, and I have seen published textbooks make mistakes on this**

2.17 The range

• The range is simply given by the difference between the maximum and minimum values:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

• The range gives a fairly crude measure of the spread of the data

• **The variance/standard deviation and the inter-quartile range may give better summaries**

• **Numerical Example.**

Data

$(11, 14, 19, 18, 10, 13, 22, 18, 11, 14, 1, 12, 12, 18, 6, 12, 11, 19, 18, 2, 22, 18)$

$$\text{Range} = \text{Maximum value} - \text{Minimum value} = 22 - 1 = 21$$

2.18 Calculation of variance and standard deviation using Excel

- Suppose you have data in cells $A1 : A10$ and wish to calculate the variance.
  1. Click on the insert formula icon f_x
  2. Choose category *Statistical*
  3. Select *VAR.S*. Press *OK*
  4. Enter $A1 : A10$ in the box titled **Number 1**
- The standard deviation can be calculated in exactly the same way as above using the command *STDEV.S*

2.19 Calculation of range and inter-quartile range

- Suppose you have data in cells $A1 : A10$ and wish to calculate the inter-quartile range.
  1. Click on the insert formula icon f_x
  2. Choose category *Statistical*
  3. Select *QUARTILE.EXC*. Press *OK*
  4. Enter $A1 : A10$ in the box titled **Array**
  5. In the box marked **QUART** insert the value 3 to calculate the upper quartile.
  6. Repeat these steps and in the box marked **QUART** insert the value 1 to calculate the lower quartile.
     item The difference between these two values then gives the inter-quartile range
- The range can be calculated in exactly the same way as above. Inserting 4 into the box marked **QUART** calculates the maximum value. Inserting 0 into the box marked **QUART** calculates the minimum value. The difference between these two values gives the range.

3.1 Simple graphical analysis

• Mathematics and statistics are hard but the questions asked are often deceptively simple-minded.

• For example if you have two variables X and Y (e.g. X=interest rate, Y=GDP) two obvious questions to ask are

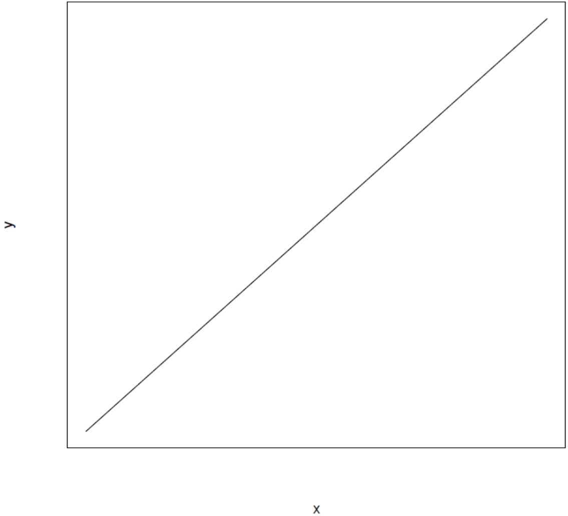
1. If X increases what happens to Y?
   - Increases?
   - Decreases?
   - Nothing?
2. If Y increases what happens to X?
   - Increases?
   - Decreases?
   - Nothing?

• These basic questions lead to a core area of statistics entitled regression

3.2 Correlation is not causation but...

- If two variables X and Y are positively correlated
  - As X increases Y increases
  - Equivalently, as X decreases Y decreases

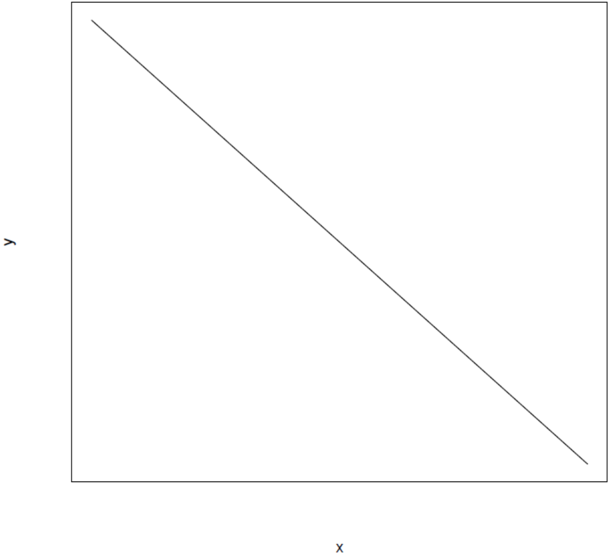
3.3 Positive correlation



3.4 Correlation is not causation but...

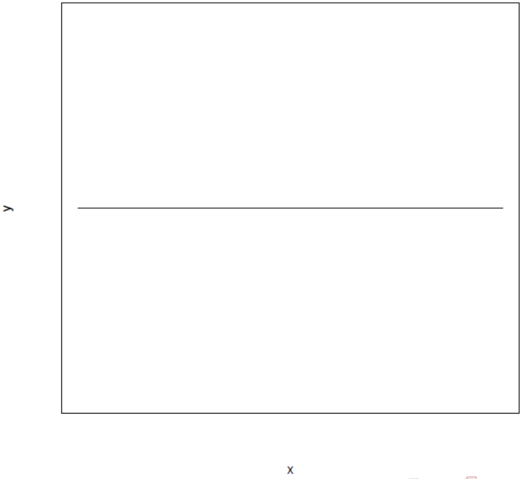
- If two variables X and Y are negatively correlated
  - As X increases Y decreases
  - Equivalently, as X decreases Y increases

3.5 Negative correlation



3.6 Zero correlation

- If X and Y are uncorrelated a change in X does not affect the value of Y:



3.7 Interpretation

- May have to use your imagination a bit. Real datasets often produce similar patterns to these graphs shown but are liable to be very imperfect.
- **This imperfect nature means that statistically rigorous regression methods are needed to fit lines to real data**
  - see later material in this course
- May also see examples where the graphs are curves rather than straight lines

4.1 Purpose of data presentations

- Building on the graphical methods discussed in the previous lecture
- Discuss different ways of presenting data – typically seen in reports and especially in dissertations
- **Our fundamental aim is to communicate information simply and effectively**
- Methods covered are
  - Stem and leaf plots
  - Frequency tables
  - Frequency polygons
  - Histograms

4.2 Professional Skills

1. Simple ways to communicate and display information
2. General numeracy
3. Attention to detail
4. A simple cross-check of more complex financial or statistical information – e.g. the bit that you might need an MSc, PhD (or more) in order to properly understand ...

4.3 Stem and leaf plots

- **An alternative way to write down the data**

Stem	=	Left-most digit
Leaves	=	Right-most digit
Leaves		Should be in size order

- **Advantages**

  - Simplicity
  - Gives an overview of the sample
  - Also gives an indication of the spread of the data within each subcategory
  - Retains the original values – unlike other methods it does not rely on the class mid-point

4.4 Stem and leaf plot – Example 1

- Suppose we have the following data for a set of exam scores: 70, 68, 72, 53, 56, 44, 64, 48, 46, 40, 64, 54, 46, 71, 61, 51, 50, 48, 35, 67, 29, 73, 61, 42, 53
- A stem and leaf plot for this data would be as follows

Stem	Leaf
2	9
3	5
4	0, 2, 4, 6, 6, 8, 8
5	0, 1, 3, 3, 4, 6
6	1, 1, 4, 4, 7, 8
7	0, 1, 2, 3

- Suppose we have the following data on wages ($ 000s): 51, 51, 48, 45, 45, 45, 44, 43, 42, 42, 41, 38, 37, 36, 35, 33, 33, 33, 32, 27, 23, 20, 18, 18
- A stem and leaf plot for this data would be as follows

Stem	**Leaf**
1	8, 8
2	0, 3, 7
3	2, 3, 3, 5, 6, 7, 8
4	1, 2, 2, 3, 4, 5, 5, 5, 8
5	1, 1

4.6 Frequency tables

- Used in order to organise data in a *systematic* way – especially for larger data sets
- May be used to create further graphs and tables e.g. frequency polygons, histograms
- Not amazingly exciting but does touch on important themes
  - General numeracy
  - Systematic analysis of data
  - Attention to detail

4.7 Frequency tables

- Suppose that we have the following data on mortgage interest rates: 7.29, 7.23, 7.11, 6.78, 7.47, 6.69, 6.77, 6.57, 6.80, 6.88, 6.98, 7.16, 7.30, 7.24, 7.16, 7.03, 6.90, 7.16, 7.40, 7.05, 7.28, 7.31, 6.87, 7.68, 7.03, 7.17, 6.78, 7.08, 7.12, 7.31, 7.40, 6.35, 6.96, 7.29, 7.16, 6.97, 6.96, 7.02, 7.13, 6.84
- This can then be rearranged into a frequency table as follows:

Interval	Frequency	Class Midpoint	Relative Frequency	Cumulative Frequency
6.30–6.50	1	6.40	0.025	1
6.50–6.70	2	6.60	0.05	3
6.70–6.90	7	6.80	0.175	10
6.90–7.10	10	7.00	0.25	20
7.10–7.30	13	7.20	0.325	33
7.30–7.50	6	7.40	0.15	39
7.50–7.70	1	7.60	0.025	40

4.8 Components of the frequency table I

- **Interval**

  - A convenient grouping of the observations – usually round numbers ending in 5 or 10

- **Frequency**

  - A raw count of the numbers in each category

  - E.g. 7 observations (6.78, 6.77, 6.80, 6.88, 6.87, 6.78, 6.84) on the previous slide fall into the category 6.70–6.90

  - Easy to make a mistake here if you do not pay close enough attention to detail!

- **Class Midpoint**

$$\text{Class Midpoint} = \frac{\text{Upper Boundary} + \text{Lower Boundary}}{2}$$

  - Often needed for subsequent additional computations

  - May sometimes give a useful way of visualising data in its own right

4.9 Components of the frequency table II

- **Relative Frequency**

$$\text{Relative Frequency} = \frac{\text{Class Frequency}}{\text{Total Number}}$$

- On the previous slide have 7 observations in the category 6.70–6.90. Relative Frequency=7/40=0.175. Similarly 13 observations in the category 7.10–7.30. Relative Frequency=13/40=0.325.
- The Relative Frequency is a proportion between 0 and 1 measuring how common each category is
- The sum of all the Relative Frequencies should equal 1

4.10 Components of the frequency table III

- **Cumulative Frequency**

  - To accumulate means "to gather"

  - The Cumulative Frequency is the running total of all the previous observations

- **Previous Mortgage Interest-Rate Example**

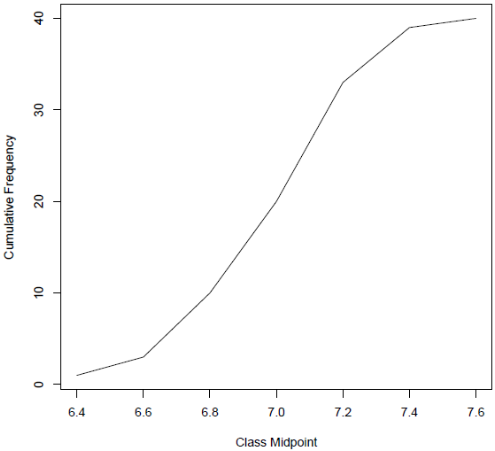
  - There is 1 observation in the first category 6.30–6.50 so the cumulative frequency is 1

  - There are 2 observations in the second category 6.50–6.70. The cumulative frequency is the previous value (1)+2=3

  - There are 7 observations in the third category 6.70–6.90. The cumulative frequency is the previous value (3)+7=10

4.11 Frequency Polygon

- **A frequency polygon is a plot class of class midpoint against cumulative frequency**
  - Will show how to construct frequency polygons in Excel



4.12 Frequency Polygons in Excel

- Need a column of midpoint values and a column of cumulative midpoints side-by-side. The column of class midpoint values needs to be on the left

1. Highlight the two columns
2. Insert→Scatter→Scatter with Straight lines

4.13 Histograms

- A histogram is essentially a bar chart of the grouped data
- In excel you need a column of category names to the left of a column of observed frequencies

1. Highlight the two columns
2. Insert→Bar→2-D Bar

4.14 Histograms with unequal class sizes

- What happens if not all the classes are of equal widths?
- Raw frequencies are not now so important
- What matters instead is *Frequency Density*

$$\text{Frequency Density} = \frac{\text{Frequency}}{\text{Class Width}} = \frac{\text{Frequency}}{\text{Upper bound} - \text{Lower bound}} \quad (1)$$

- **Frequency density measures how likely a class is to occur once the differing class widths have been taken into account**

4.15 Histograms with unequal class sizes

- Suppose we have the following data on people's ages
- We might think that there are clearly more people in the 60-100 category than in the 0-15 category
- However, using frequency density rather than the raw frequencies shows that this is not the case

Age	Frequency	Class Width	Frequency Density
0-15	15	15	$\frac{15}{15} = 1$
15-25	28	10	$\frac{28}{10} = 2.8$
25-40	30	15	$\frac{30}{15} = 2$
40-60	42	20	$\frac{42}{20} = 2.1$
60-100	20	40	$\frac{20}{40} = 0.5$

4.16 Histograms with unequal class sizes in Excel

- A histogram is essentially a bar chart of the grouped data
- In excel you need a column of category names to the left of a column of observed frequency densities

1. Highlight the two columns
2. Insert→Bar→2-D Bar

5. Exercises

1. Produce a stem and leaf plot for the following data: 54, 11, 91, 66, 92, 19, 1, 77, 83, 57, 30, 52, 100, 39, 62, 35, 99, 68, 53, 7, 79, 10, 13, 50, 9, 34, 74, 88, 18, 24, 24, 69, 40, 83, 32
2. What data corresponds to the following stem and leaf plot

Stem	0	2	4	6	10	16
Leaf	1 2 3 8	0 7	5	8	5	0

3. Construct a frequency polygon for the data on Slide 3.5
4. Construct a frequency table for the data in Q2.
5. In Q4 what is the midpoint of the highest group?
6. In Q4 what is the width of the lowest group?
7. Give the relative and cumulative frequencies of the 3rd class
8. If the cumulative frequency of Class $n-1$ is 10, the relative frequency of Class n is 0.031 and there are 128 observations what is the cumulative frequency of Class n?
9. If the relative and cumulative frequencies for adjacent classes are 0.067, 0.2 and 4, 16 respectively how big was the sample?