MA208 Quantitative Techniques for Business

Lecture 2: Statistics ctd.

Dr Kirsten Pfeiffer

School of Mathematics, Applied Mathematics and Statistics
NUI Galway

Lecture 2 - Outline

- Stem and Leaf Displays
- **Numerical** methods to describe and analyse data sets
  - Measures of centre
  - Measures of spread
  - Five Number Summary
  - Boxplots

In our first lecture we have introduced some graphical tools to describe data sets: for example dot plot diagrams, bar charts, pie charts, histrograms.

Each type of diagram has its own advantages and disadvantages. For example dot pots are less useful than histograms when we have large collections of data and therefore too many dots for each case in the data.

Another useful diagram is the Stem and Leaf Display. This diagram takes into account individual data and is also useful to visualise the shape of a distribution.

Example

Grades on a Science Test

Stem	Leaf
7	2 2 4 5 6 9
8	1 4 5 7 7 9
9	0 1 3 5 8
10	0 0

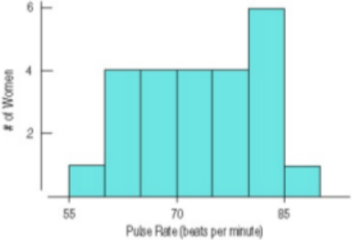
Key: 7 / 2 means 72 percent

Stem and Leaf Displays

Exercise

Stem-and-Leaf Example

- Compare the histogram and stem-and-leaf display for the pulse rates of 24 women at a health clinic. Which graphical display do *you* prefer?



8	8
8	000044
7	6666
7	2222
6	8888
6	0444
5	6

Slide 4- 5

Constructing a Stem and Leaf Display:

- Cut each data value into leading digits (stems) and trailing digits (leaves).
- Use only one digit for each leaf.
- Arrange leaves in ascending order.

Exercises

1. The scores of ten exam students are

$$77, 88, 92, 57, 55, 73, 64, 89, 44, 76.$$

   Display these data using a stem and leaf plot.

2. Subjects in a psychological study were timed while completing a certain task. Set up a stem and leaf plot for the following list of times.

$$5.8, 5.9, 6.1, 6.2, 6.8, 7.3, 7.6, 7.7, 8.1, 8.1, 8.2, 8.8, 9.2$$

Exercises

(1) The scores of ten exam students are

$$77, 88, 92, 57, 55, 73, 64, 89, 44, 76.$$

Display these data using a stem and leaf plot.

- 7|7 8|8 9|2 5|7 7|3 6|4 8|9 4|4 7|6
- Use one digit for each leaf.

4	4
5	7 5
6	4
7	7 3 6
8	8 9
9	2

in ascending order →

4	4
5	5 7
6	4
7	3 6 7
8	8 9
9	2

Exercises

(2) Subjects in a psychological study were timed while completing a certain task. Set up a stem and leaf plot for the following list of times.

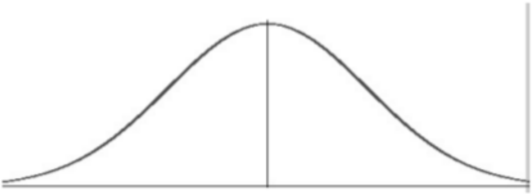
5.8, 5.9, 6.1, 6.2, 6.8, 7.3, 7.6, 7.7, 8.1, 8.1, 8.2, 8.8, 9.2

- Cut stems and leaves.

5|8 5|9 6|1 6|2 6|8 7|3 7|6 7|7 8|1 8|1
8|2 8|8 9|2

5	8 9
6	1 2 8
7	3 6 7
8	1 1 2 8
9	2

We will now consider some **numerical** methods to describe and analyse data sets.
There are many summary values that can be attached to a data set. They are used to describe things like centre and spread.
For an unimodal symmetric distribution it is easy to find the centre - it's just the centre of the symmetry.



Midrange

As another measurement of the centre we could use the **midrange**:

Midrange

$$\text{midrange} = \frac{\text{minimum}+\text{maximum}}{2}$$

However, this is very sensitive to skewed distributions and outliers.

Example

Consider the midrange of the following list of exam results.

$$0, 37, 38, 39, 38, 41, 45, 36, 39, 43, 40$$

Is the midrange a good measure of centre for this data set?

Example

Consider the midrange of the following list of exam results.

$$0, 37, 38, 39, 38, 41, 45, 36, 39, 43, 40$$

Is the midrange a good measure of centre for this data set?

$$\text{midrange} = \frac{0+45}{2} = 22.5$$

not ideal ...

Median

Another measure of the centre is the median. The median $\tilde{x}$ is the middle value, i.e. the data value with half of the data above it and half of the data below it, when our data is arranged in order.

Median

For n data values, the median $\tilde{x}$ is the $\frac{n+1}{2}$ largest observation. If $\frac{n+1}{2}$ is not a whole number, the median is the average of the two data values on either side.

Example

Find the median values of the following data sets.

1. $\{7, 5, 2, 11, 13, 4, 3, 2, 9, 7, 12\}$
2. $\{1, 2, 2, 3, 4, 5, 7, 8, 8, 9, 11, 12, 13, 13, 14, 15\}$

Median

Example

Find the median values of the following data sets.

1. $\{7, 5, 2, 11, 13, 4, 3, 2, 9, 7, 12\}$

   · Order : 2, 2, 3, 4, 5, 7, 7, 9, 11, 12, 13

   $(n = 11)$ $\frac{11+1}{2} = 6$ 6th   median $\tilde{x} = 7$

2. $\{1, 2, 2, 3, 4, 5, 7, 8, 8, 9, 11, 12, 13, 13, 14, 15\}$

   $(n = 16)$ $\frac{16+1}{2} = 8\frac{1}{2}$

   $$\tilde{x} = \frac{8^{th} + 9^{th}}{2} = \frac{8+8}{2} = 8$$

Later in this lecture we will introduce more measure of centre.

For describing a distribution numerically we need a measure of its spread along with its centre.
The range of the data is the difference between the maximum and the minimum values.

Range

range = maximum - minimum.

One disadvantage of using the range as a measure of spread is that a single extreme value can make the range very large and thus is not representative of the bulk of the data.

Example

Calculate the range of the following data sets.

1. $\{25, 30, 31, 33, 33, 42, 3, 99\}$
2. $\{25, 30, 31, 33, 33, 42\}$

Range

Example

Calculate the range of the following data sets.

1. $\{25, 30, 31, 33, 33, 42, 3, 99\}$

   range $= 99 - 3 = 96$

2. $\{25, 30, 31, 33, 33, 42\}$

   range $= 42 - 25 = 17$

Interquartile Range

The interquartile range (IQR) allows us to ignore extreme data values and concentrate on the middle values of the data.
To find the IQR we first need to define quartiles.
The lower quartile Q_1 divides the bottom half of the data into two.

Lower Quartile

Q_1 = median of data below the median

The upper quartile Q_3 divides the upper half of the data into two.

Upper Quartile

Q_3 = median of data above the median

Note: The lower and upper quartile are sometimes referred to as the 25^{th} and 75^{th} percentile of the data.

Interquartile Range

The difference between the quartiles range is the IQR.

Interquartile Range (IQR)

$$IQR = Q_3 - Q_1 .$$

Note: The interquartile range contains the middle 50% of the data.

Example

The number of moons of the eight planets in the solar system are:

$$\{0, 0, 1, 2, 63, 61, 27, 13\}$$

Calculate the IQR of this data set.

Interquartile Range

Example

The number of moons of the eight planets in the solar system are:

$$\{0, 0, 1, 2, 63, 61, 27, 13\}$$

Calculate the IQR of this data set.

Order: $\{0, 0, 1, 2, 13, 27, 61, 63\}$

min $= 0$, max $= 63$

median $\tilde{x} = \frac{2+13}{2} = 7.5$

$Q_1 = \frac{0+1}{2} = 0.5$

$Q_3 = \frac{27+61}{2} = 44$

$IQR = Q_3 - Q_1 = 44 - 0.5 = \underline{\underline{43.5}}$

Five Number Summary

The five number summary provides a concise summary of the distribution.

Five number summary

The five number summary is defined as

$$\{min, Q_1, median, Q_3, max\}$$

and it is a useful set of summary statistics for a data set.

Example

Calculate the five number summary for number of moons of the eight planets in the solar system: $\{0, 0, 1, 2, 63, 61, 27, 13\}$.

Five Number Summary

Example

Calculate the five number summary for number of moons of the eight planets in the solar system: $\{0, 0, 1, 2, 63, 61, 27, 13\}$.

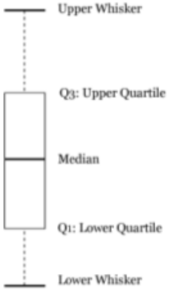
Five number summary:

$$\{\min, Q_1, \text{median}, Q_3, \max\}$$

$$= \{0, 0.5, 7.5, 44, 63\}$$

Boxplot

The five number summary can be represented graphically using a **box plot**.



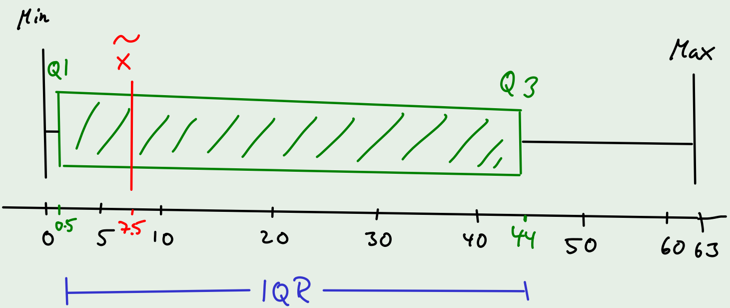
Example

Draw a boxplot representing the number of moons of the eight planets in the solar system: $\{0, 0, 1, 2, 63, 61, 27, 13\}$.

Example

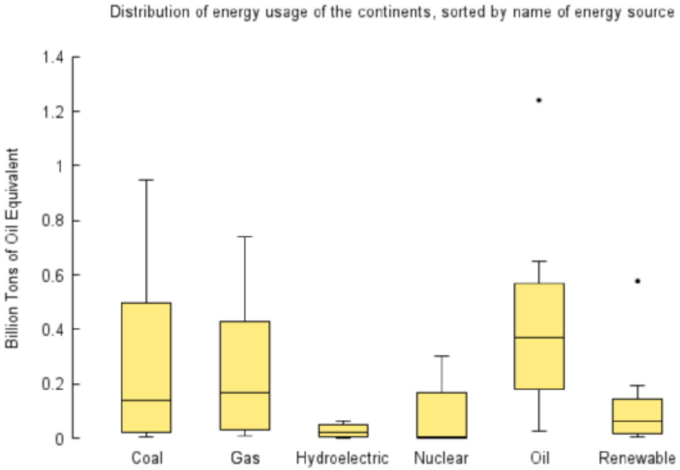
Draw a boxplot representing the number of moons of the eight planets in the solar system: $\{0, 0, 1, 2, 63, 61, 27, 13\}$.

Five number summary = { 0, 0.5, 7.5, 44, 63 }



Boxplot

Boxplots are useful for quickly comparing data from different groups.



Boxplot

To turn a five number summary into a box plot we need to decide on the length of the whiskers. These are lines that extend from either end of the box. The rule of thumb for whiskers is that they should extend a maximum distance of $1.5 \times IQR$, but not pass the smallest/largest value.

Exercise

An insurance company has collected the following data on the number of car thefts per day in a large city for a period of 21 days.

52	61	44	64	55	58	76
97	53	65	59	57	33	48
54	80	57	57	69	70	59

Construct a box plot for this data set.

Boxplot

Solution

- Order data set. 33 44 48 52 53 54 55 57 57 57
  58 59 59 61 64 65 69 70 76 80 97
- Min $= 33$, Max $= 97$

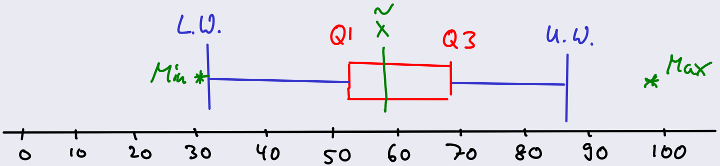
  $\tilde{x} = \left(\frac{21+1}{2}\right)^{th} = 11^{th}$ position $= 58$

  $Q_1 = \frac{53+54}{2} = 53.5$, $Q_3 = \frac{65+69}{2} = 67$

  $IQR = 67 - 53.5 = 13.5$

  Upper Whisker $= Q_3 + 1.5\,IQR = 87.25$

  L.W. $= Q_1 - 1.5\,IQR = 33.25$



We have two outliers.