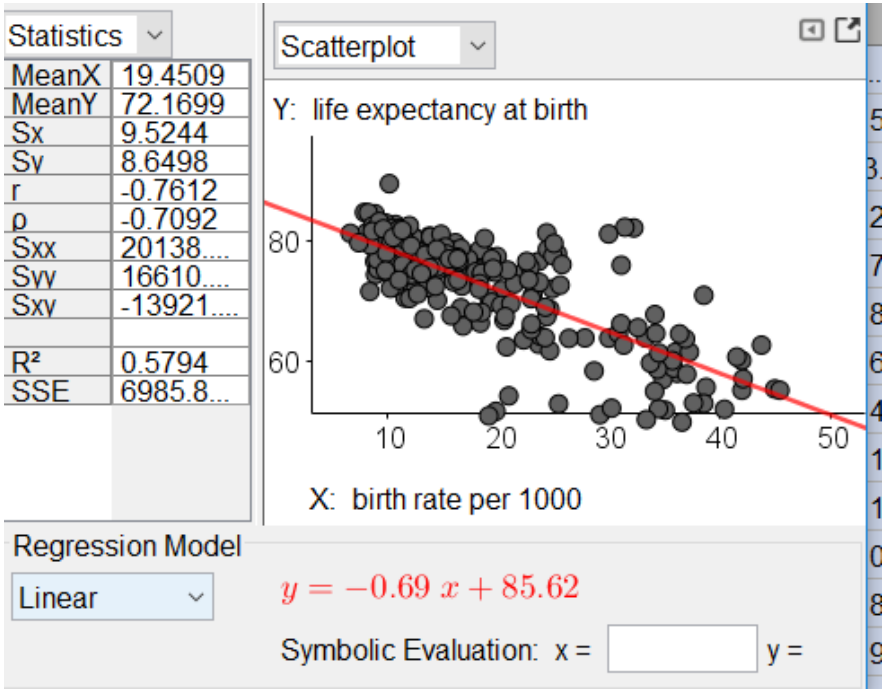
Sponsored by

CASIO

@MEIConference #MEIConf2018

Using the MEI datasets
Paul Chillingworth



If you don't have the MEI datasets please download them at:

http://bit.ly/FMSPLDS

Data Sets: Key Points

- Each awarding body will supply one or more datasets; questions will be set which assume familiarity with the dataset.
- Students must be familiar with both the data and its context; they need to know the origin of the data and how it might have been collected.
- Technology must be used to explore the data set.
- Increased emphasis on interpretation over calculation and being able to select the correct representation or model

In the exam:

- Students should be expected to interpret output from a spreadsheet or statistical software.
- Students should be expected to interpret and explain terminology which has been introduced via the data set.
- Students should be asked to explain what effect missing data would have on a model that has been derived.

An opportunity

- Using the large data sets is an excellent opportunity to teach statistical concepts in context and to use real-life data
- Their use provides an opportunity for students to analyse data using technology and to interpret the results; a skill that is important in further study and the workplace

The MEI datasets

Why three?

Whilst students will only need to be ‘familiar’ with one dataset for their examination, three have been provided to allow a wider scope of statistical exploration in the classroom.

It is unlikely that a single data set would lend itself to the full range of analysis that was originally suggested by Ofqual.

How might students use technology?

- Sorting and searching; identifying outliers;
- Producing and using summary statistics to make conclusions about the data;
- Producing graphs such as histograms and box-plots that allow comparisons to be made;
- Producing scattergraphs and modelling using trend lines or curves;
- Selecting random samples and comparing them to the population to illustrate variation
- Checking to see if data fits a particular model or distribution; hypothesis testing.

Data Set 1

- Data about countries from the CIA World Factbook.
- This is a collection of data for US policymakers. The latest version can be found at https://www.cia.gov/library/publications/the-world-factbook/
- The LDS is a subset of the data available in the CIA World Factbook. All countries for which the CIA has population data have been included; some of them with small populations are not independent countries.

Data Set 1

The data is largely ‘demographic’ or geographical.

Country, Sub region, population, birth rate, death rate, median age, life expectancy at birth, labour force, unemployment, GDP per capita (US$), physician density, health expenditure, total area

land area, water area, land borders, dependency status

Familiarity with the dataset

- The first task when working with any dataset is to fully understand what the data means and how it might be collected/calculated/recorded.
- There is a glossary to help with this but further research might be needed.
- e.g. what does ‘Life expectancy at birth’ mean and how is it worked out?

https://www.gapminder.org/

Data Set 2

- Data about boroughs in London together with some comparative data for other areas in the UK.
- 32 boroughs together with the City of London make up London.
- Further data available through the borough profiles on the London Datastore https://data.london.gov.uk/dataset/london-borough-profiles

Data Set 2

Area code, Area, Inner or Outer London

Male Life Expectancy at birth

Female Life Expectancy at birth

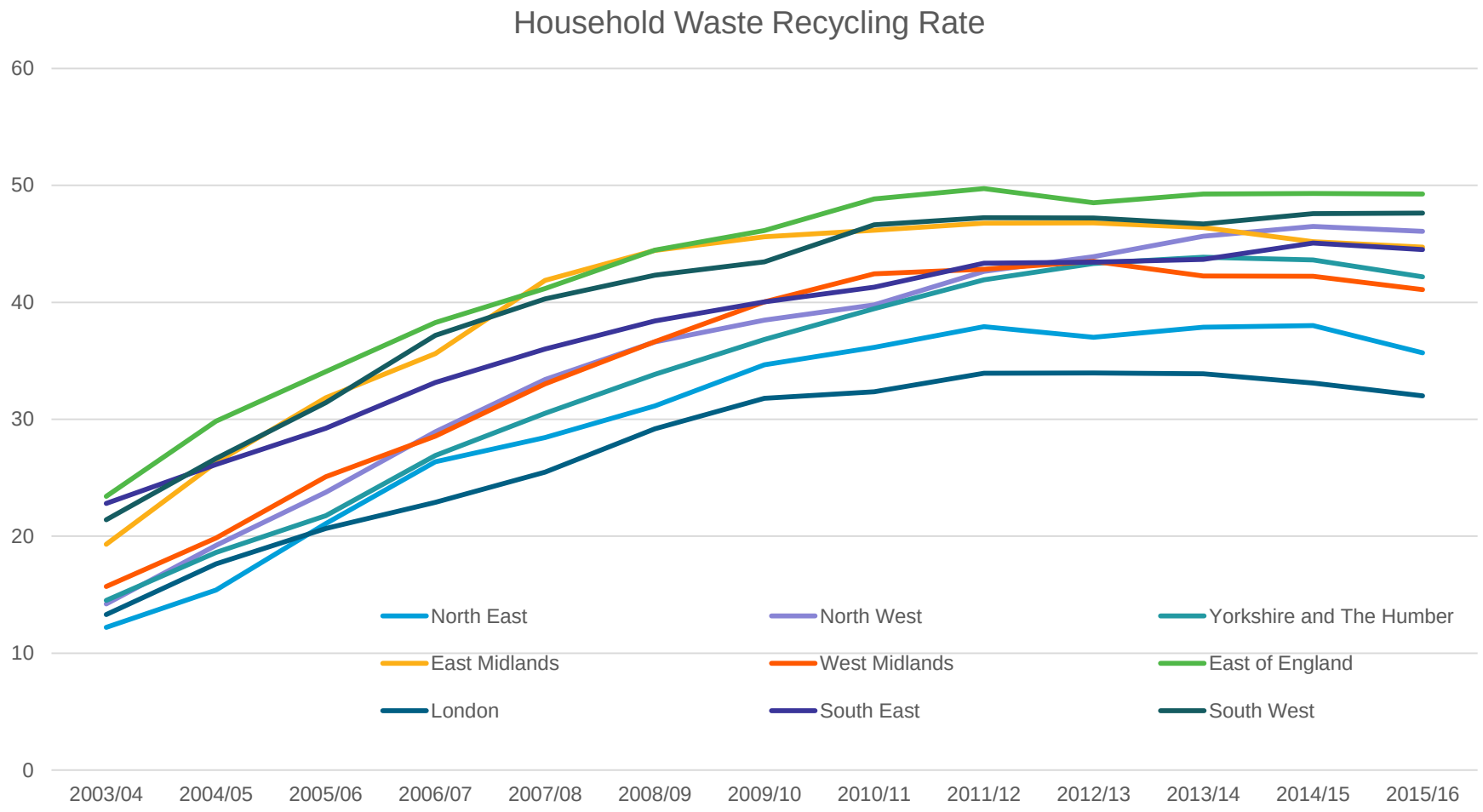
Employment Rate (age 16 to 64)

Median House Price (£)

All pupils at end of KS4 achieving 5+ A-C including English and Maths*

Household Waste Recycling rate

Across a number of years, so time series


Household Waste Recycling Rate
60
50
40
30
20
10
0
2003/04
2004/05
2005/06
2006/07
2007/08
2008/09
2009/10
2010/11
2011/12
2012/13
2013/14
2014/15
2015/16
North East
North West
Yorkshire and The Humber
East Midlands
West Midlands
East of England
London
South East
South West

Data Set 3

- Data about individuals who took part in the American National Health and Nutrition Examination Survey (NHANES) in 2003-4.
- http://www.eeps.com/zoo/nhanes/source/choose.php
- Students can compare results from this sample with other samples from the same source.

Data Set 3

The National Health and Nutrition Examination Survey has been taking place since the 1960s in the US. It combines interview and physical examination to assess the health and nutritional status of adults and children in the United States.
http://www.cdc.gov/nchs/nhanes/about_nhanes.htm .

The survey examines a representative sample of 5 000 people each year.
The data in the Large Data Set is a subset of the information collected for a random sample of 200.

Data Set 3

Sex, Age, Marital Status

Weight, Height, BMI

ThighLen, UpperArmLength, Waist

Food30

Arm, Pulse

Systolic

Diastolic

Use of Technology

- Excel
- GeoGebra
- R

Most analysis can be done using these.

Any others?

Activities to try

Datasets can be accessed from:

http://bit.ly/FMSPLDS

Survey

- Message from Jennie Golding at UCL:-
- Are you teaching, or preparing to teach, A/AS Mathematics? Ofqual and DfE are interested in your reflections on the introduction of the large data set, including the use of appropriate technology.
- Go to https://goo.gl/forms/wQfjImeIqk5NSL3F3 for a short (10-15 mins) anonymous survey that will be actively harnessed to influence future support for teachers and students, and future policy developments.
- Any queries to Jennie Golding (j.golding@ucl.ac.uk).

About MEI

- Registered charity committed to improving mathematics education
- Independent UK curriculum development body
- We offer continuing professional development courses, provide specialist tuition for students and work with employers to enhance mathematical skills in the workplace
- We also pioneer the development of innovative teaching and learning resources

MEI Conference 2018

Sponsored by CASIO

Activities with the MEI Datasets

Paul Chillingworth

paul.chillingworth@mei.org.uk

1. Measures (using dataset 1)

High birth rates and poverty undermine a generation of African children – report

See https://www.theguardian.com/global-development/2016/aug/25/high-birth-rates-poverty-undermine-generation-african-children-odi-report

Investigate the issues raised in the report by considering the birth rate, death rate, median age and life expectancy at birth. Compare these statistics for African countries to those of the rest of the world.

How are these four statistics calculated?

The populations of these countries are also important factor to consider. Why?

Are there any countries in Africa that have different trends to the rest of the continent? Why might this be?

Follow up ideas:

How has life-expectancy changed over time?

What factors influence the life-expectancy of a country?

2. Correlation (using dataset 1)

What do you think the following scatter diagrams would look like?

Land area v population

Unemployment v GDP

Life expectancy v GDP

Birth rate v life expectancy

Land area v median age

For each one, estimate then calculate the correlation coefficient. Investigate any interesting features of the diagram

3. Time series; median (using dataset 2)

Draw time series graphs for median house prices in England and for London on the same axes.

If you do not live in London, include your region as well.

What do you notice about how median house prices have changed?

Draw the time series graphs for median house prices for London as a whole on the same axes as two individual areas in London. Choose one expensive area and one cheaper area. What do you notice about how median house prices have changed?

What information does the median house price give? What else would it be useful to know when thinking about whether someone could afford to buy their first home? Can you find any further information online?

4. Hypothesis Testing (using dataset 3)

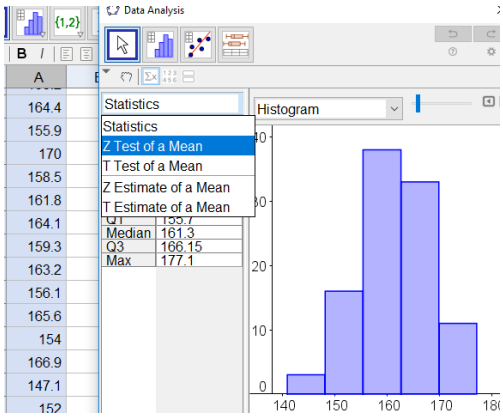
The 3rd MEI Dataset gives data from the American National Health and Nutrition Examination Survey in 2003-04.

Filter the spreadsheet to focus on the 101 women and consider their heights.

In 2010, the average height of US women was 162.05cm with SD of 8.9cm.

Test the hypothesis that your sample is drawn from a population with mean of 162.05cm against the alternative that the mean is less than this. Test at the 5% level. Assume that the data is normally distributed and that the SD is 8.9cm.

GeoGebra can be used for this test:



5. Repeated sampling (using dataset 3)

If each data item in the population is numbered, then we can use these numbers to select a random sample. If the population has not been numbered then the first step is to add numbers.

If you need to number the sampling frame, insert a blank column in Column A. Enter *1* into A2 and then *=A2+1* into A3 and *copy* this down to the bottom of the sheet.

	A	B	C	D	E	F
1		Sex	Age	Marital	Weight	Height
2	1	Female	34	Married	60.3	173.4
3	2	Female	85	Widowed	64.7	161.2
4	3	Female	48	Divorced	100.6	171.4
5	4	Male	61	Married	70.9	169.5
6	5	Male	68	Divorced	96.8	181.6
7	6	Female	28	Never married	50.2	158.5
8	7	Male	37	Living with partner	115.9	172.6
9	8	Female	36	Living with partner	54.4	161.3
10	9	Male	42	Married	71.5	172.3

To take a sample in a new sheet:

Create a new sheet in which to place the sample. Suppose we want a sample of size 20, drawn from a sampling frame of 200 items.

Enter this formula into A2 and copy down to A21

```
=RANDBETWEEN(1,200)
```

This has generated 20 random numbers between 1 and 200. If you edit the sheet or press F9 these numbers will change.

Now we need to look up the relevant data from the datasheet using these numbers as a reference. As our data is in columns, we can use the vlookup function to do this.

```
=VLOOKUP(A2,Data!$A$1:$W$201,2,FALSE)
```

The first argument *A2* specifies where to find the value to lookup.

The second argument specifies where to find the data. 'Data!' is the name of the sheet in the spreadsheet and $A$1:$W$201 is the range of cells within that sheet that contain the data. It is necessary to use a $ sign to stop this range changing when we copy it down or across.

The third argument gives the column in the sheet that we are interested in. Here we will display the data from column 2 (the gender).

The final argument should be set to false; this concerns how the data is displayed.

This formula can be entered into B2 and copied down to B21.

It can also be copied across and edited slightly to produce further data:

```
=VLOOKUP(A2,Data!$A$1:$W$201,5,FALSE)
```
Height

```
=VLOOKUP(A2,Data!$A$1:$W$201,6,FALSE)
```
Weight

So my sample looks like:

	A	B	C	D
1	Rand No	Gender	Height	Weight
2	173	Female	62.9	160.3
3	89	Female	87	164
4	4	Male	70.9	169.5
5	50	Male	70.4	159.4
6	101	Female	100.1	172.9
7	69	Female	71.1	164.2
8	88	Male	78	177.3
9	170	Female	76	158.6
10	171	Male	67.6	173
11	86	Female	52.2	154.3
12	162	Male	90.7	172.8
13	28	Male	61.7	177.5
14	100	Male	85	177.7
15	157	Female	132.2	153.7
16	53	Male	87	179
17	64	Female	52.3	157.7
18	109	Male	91.4	173.2
19	61	Male	120.3	176.5
20	146	Female	88.8	152.7
21	32	Female	88.2	166.9

Pressing F9 will refresh the random numbers and change the sample.

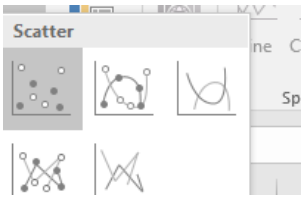
This is a great opportunity to see how the sample characteristics change.

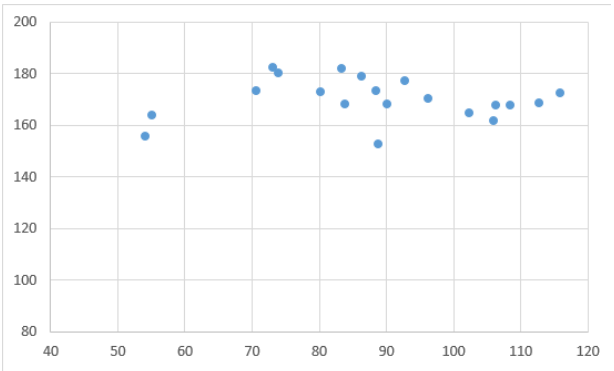
A good example of this might be the mean and SD of each variable or the correlation between them. You could draw charts, such as a scatter graph.

For example, to get the PMCC use:

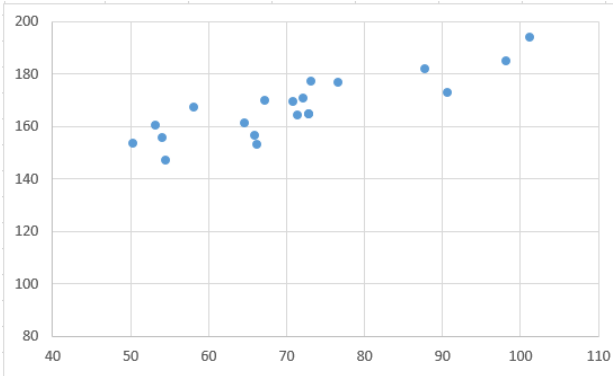
```
=CORREL(C2:C21,D2:D21)
```

To draw a scatter graph, highlight the data and click on insert





Correlation	-0.0103



Correlation	0.86837

Here we have two very different pictures. Try changing the sample sizes and see what happens.

Compare the sample values to the true value from the whole population as a whole.

Should men and women be considered separately? How could you do this?

Activity 3 based on an activity available in Integral: https://integralmaths.org/