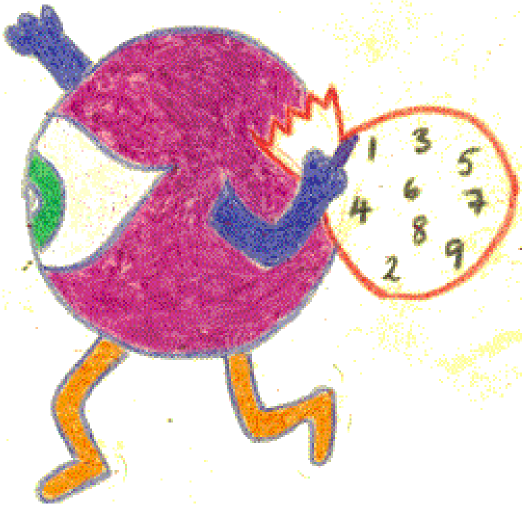
Relevant & Engaging Statistics & Data Handling

Using Random Samples of Real Data



ROYAL STATISTICAL SOCIETY

FOUNDED 1834

Centre for Statistical Education

Chapter 3

Second Edition: published in the UK in 2009 by

The Royal Statistical Society Centre for Statistical Education
The University of Plymouth
Plymouth
PL4 8AA
UK

The advice and information in this booklet are believed to be true and accurate at the date of printing, but neither the authors, nor the publisher can accept any legal responsibility or liability for errors or omissions.

We would like to thank the many contributors to this book. These include: Doreen Connor, Neville Davies, John Marriott, Alan Catley (Chapter 4) and Mark Crowley (Chapter 6) and those who reviewed and helped in it's production including Alison Davies, Claire Webster, Peter Holmes & Elizabeth Gibson.

This booklet is aimed at all secondary level teachers: there are hints and tips that we hope will be useful to support the teaching and learning of statistics and data handling. We hope that you find the material useful. Please email or write to us with suggestions for improvements. We will try to respond to all communications.

Doreen Connor
The Royal Statistical Society Centre for Statistical Education
The University of Plymouth
2009

email info@censusatschool.org.uk

www.censusatschool.org.uk

www.rsscse.org.uk

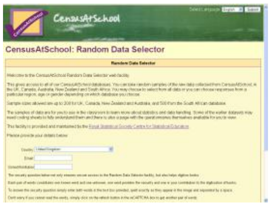
Chapter 3

Using Random Samples of Real Data

Once you have your own learners' data you may well want to compare this to other comparable data in order for your learners to be able to compare themselves with others. In this chapter we go through the steps needed to both take and use random samples of real data from the CensusAtSchool website. In addition we offer some ideas to allow learners to use samples of real data in their data handling and statistics lessons.

The process of obtaining and using the samples can be done equally well at home or in school. All you need is a computer and access to a spreadsheet.

Step 1 *Access the random data selector page*



CensusAtSchool: Random Data Selector

Random Data Selector

Welcome to the CensusAtSchool Random Data Selector web facility.

This gives access to all of our CensusAtSchool databases. You can take random samples of the raw data collected from CensusAtSchool in the UK, Canada, Australia, New Zealand and South Africa. You may choose to select from all data or you can choose responses from a particular region, age or gender depending on which database you choose.

Sample sizes allowed are up to 200 for UK, Canada, New Zealand and Australia, and 500 from the South African database.

[illegible]

This facility is provided and maintained by the Royal Statistical Society Centre for Statistical Education

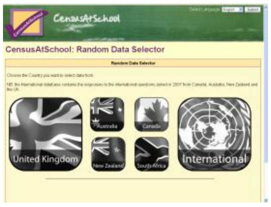
Please provide your details below

Country: United Kingdom
Email:
School/Institution:

[illegible]

Go to the *CensusAtSchool* random data selector (RDS) web page (currently http://rds.censusatschool.org.uk/ and fill in your email address, name, institution and complete the security question. The RDS can also be found by linking through the 'get data' section of CensusAtSchool.

Step 2 *Choose your sample – country, phase, size*



CensusAtSchool: Random Data Selector

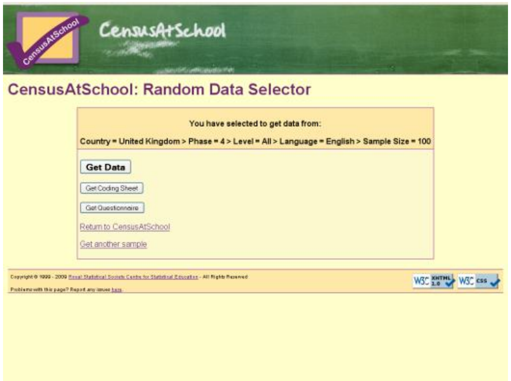
Random Data Selector

Choose the Country you want to select data from

NB: the International database contains the responses to the international questions asked in 2007 from Canada, Australia, New Zealand and the UK.

Begin by clicking the flag to select the country you are interested in. The international button provides you with answers to the 11 questions asked in 2007/8 which were the same across NZ, Australia, the UK and Canada. After this you need to select the correct phase and the size of sample you require. This is usually up to 200.

Step 3 *Get supporting documents*



CensusAtSchool: Random Data Selector

You have selected to get data from:

Country = United Kingdom > Phase = 4 > Level = All > Language = English > Sample Size = 100

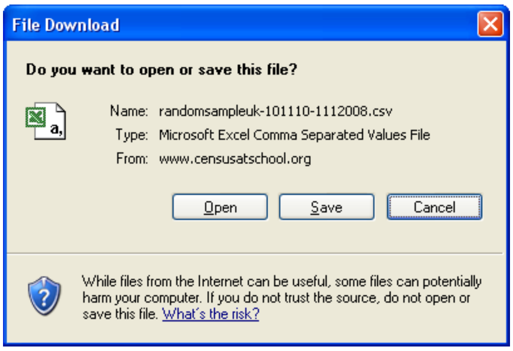
Copyright © 1999 - 2009 Royal Statistical Society Centre for Statistical Education - All Rights Reserved

Problems with this page? Report any issues here.

Before viewing your sample you will have the opportunity to download the appropriate coding sheet and questionnaire for the data you have selected. This is because abbreviations are sometimes used in the spreadsheets so will help you digest the numbers easily.

Step 4 Get data

By clicking the 'get data' button the sampled data will download immediately (NB: many computers will have security which will mean you need to click on the top information bar to allow the computer to download the file).



Do you want to open or save this file?

Name: randomsampleuk-101110-1112008.csv

Type: Microsoft Excel Comma Separated Values File

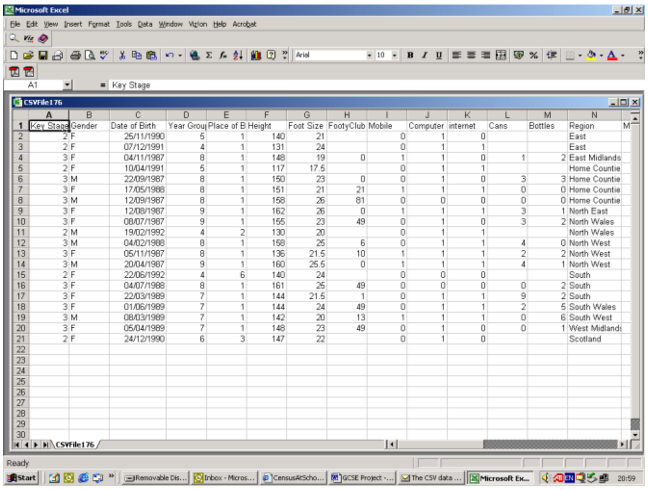
From: www.censusatschool.org

While files from the Internet can be useful, some files can potentially harm your computer. If you do not trust the source, do not open or save this file. What's the risk?

The standard format is a comma separated variable (CSV) file. This will take place immediately and there are more options to download located further down the web page if you run into any problems or wish to directly download an Excel version of the file.

Step 5 *Store the returned data file*

You will need to save the file to a convenient folder on your computer. This can be done by right clicking on the .CSV file and saving to a folder in the usual way. Alternatively, as long as you have Excel installed, by double left-clicking on the .CSV file, the file will open in it and will appear similar to the following image.


Microsoft Excel
File Edit View Insert Format Tools Data Window Vizion Help Acrobat
Arial
10
CSVFile176
Ready

A1 = Key Stage

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Key Stage	Gender	Date of Birth	Year Grou	Place of B	Height	Foot Size	FootyClub	Mobile	Computer	internet	Cans	Bottles	Region
2	2	F	25/11/1990	5	1	140	21		0	1	0			East
3	2	F	07/12/1991	4	1	131	24		0	1	1			East
4	3	F	04/11/1987	8	1	148	19	0	1	1	0	1	2	East Midlands
5	2	F	10/04/1991	5	1	117	17.5		0	1	1			Home Countie
6	3	M	22/09/1987	8	1	150	23	0	0	1	0	3	3	Home Countie
7	3	F	17/05/1988	8	1	151	21	21	1	1	1	0	0	Home Countie
8	3	M	12/09/1987	8	1	158	26	81	0	0	0	0	0	Home Countie
9	3	F	12/08/1987	9	1	162	26	0	1	1	1	3	1	North East
10	3	F	08/07/1987	9	1	155	23	49	0	1	0	3	2	North Wales
11	2	M	19/02/1992	4	2	130	20		0	1	1			North Wales
12	3	M	04/02/1988	8	1	158	25	6	0	1	1	4	0	North West
13	3	F	05/11/1987	8	1	136	21.5	10	1	1	1	2	2	North West
14	3	M	20/04/1987	9	1	160	25.5	0	1	1	1	4	1	North West
15	2	F	22/06/1992	4	6	140	24		0	0	0			South
16	3	F	04/07/1988	8	1	161	25	49	0	0	0	0	2	South
17	3	F	22/03/1989	7	1	144	21.5	1	0	1	1	9	2	South
18	3	F	01/06/1989	7	1	144	24	49	0	1	1	2	5	South Wales
19	3	M	08/03/1989	7	1	142	20	13	1	1	1	0	6	South West
20	3	F	05/04/1989	7	1	148	23	49	0	1	0	0	1	West Midlands
21	2	F	24/12/1990	6	3	147	22		0	1	0			Scotland

The column containing the ***Date of Birth*** variable may have to be widened to be able to view its contents (one or more of the symbols #### may appear in it as default). Save the file as a .XLS extension type so that it will automatically re-open in Excel and be amenable to all Excel facilities.

Discussion

There is some evidence that, although the UK schools that take part in the *CensusAtSchool* project *volunteer* to do so and may therefore produce self-selection bias, the responses collected from the children *are* representative of all UK school children. You could discuss this with your learners:

- What do you think are the limitations, or otherwise, of using a sample of 200 in your statistical investigation? What could you do to overcome any problems you identify with your sample?
- If you needed to take a random sample of 400 from the UK database, discuss why combining 2 separate samples of size 200 may not be a trustworthy thing to do. If you *did* create a sample of 400 in this way, what precautions would you take before proceeding with analysis?

- Try taking a random sample of 200 responses from the Queensland database and find the most popular jobs done at home by male and female learners.

Taking random samples from within *CensusAtSchool* data files

Even though the file downloaded *already* contains a random sample from *our* databases, it is a useful exercise to learn how to take random samples from *within* that file.

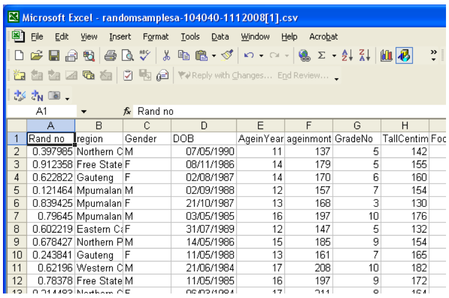
In the following steps we use a file of 500 random responses from the South African database. Our objective will be to take a random sample of the ***height (cm)*** and ***foot size*** variables.

Step 1 *Open the sample file of data and generate some random numbers*

There are several ways of taking a random sample. You can number the whole list and then use random numbers to select specific samples, or you can allocate each entry a random number, which is the method we use. Then sorting within a group allows you to take the first *n* as a random sample of size *n*. Download a file of 500 random responses from the South African country database choosing all variables. Open the data file within Excel and create a set of *reference random numbers* in a new column by highlighting the first column A, ***Region,*** then **Insert > Columns.**

A new empty column **A** will be added to the worksheet.
Click on cell **A1** and type in ***Rand no*** to label the column.
Click on cell **A2** and type **=rand()** followed by the **Enter** key**.**
Copy the formula in cell **A2** in the usual way all the way down to row 501.

Column **A** now contains 500 random numbers between 0 and 1. A fresh set of random numbers can be created in this column by pressing the **F9** key. More importantly, because **rand()** is a function it will recalculate anytime the spreadsheet does any calculation. The following image shows part of a spreadsheet with the generated random numbers in column **A**.



Microsoft Excel - randomsamplesa-104040-1112008[1].csv

A1 fx Rand no

	A	B	C	D	E	F	G	H	
1	Rand no	region	Gender	DOB	AgeinYear	ageinmont	GradeNo	TallCentim	Foo
2	0.397985	Northern C	M	07/05/1990	11	137	5	142	
3	0.912358	Free State	F	08/11/1986	14	179	5	155	
4	0.622822	Gauteng	F	02/08/1987	14	170	6	160	
5	0.121464	Mpumalan	M	02/09/1988	12	157	7	154	
6	0.839425	Mpumalan	F	21/10/1987	13	168	3	130	
7	0.79645	Mpumalan	M	03/05/1985	16	197	10	176	
8	0.602219	Eastern Ca	F	31/07/1989	12	147	5	132	
9	0.678427	Northern P	M	14/05/1986	15	185	9	154	
10	0.243841	Gauteng	F	11/05/1988	13	161	7	165	
11	0.62196	Western C	M	21/06/1984	17	208	10	182	
12	0.78378	Free State	M	11/05/1985	16	197	9	172	

Step 2 *Generate a random sample of 50 height (cm) and foot size responses*

You will first need to freeze the generated random numbers, so that they do not change every time you change something in the worksheet. There are two ways to do this:
Either
Highlight all entries in column **A** from row 2 to row 501.
Click **Edit > Copy > Edit > Paste Special > Values > OK**.
Press the **Esc** key to remove the active dots round column **A**
Or
Tools > Options click on the **Calculation** tab and choose **Manual.**

> The 500 random numbers in column **A** are now fixed.
>
> Click anywhere in the table, then **Data > Sort** and in the **Sort by** box ensure the variable ***Rand no*** is selected and the radio button against *Ascending* is activated. Then click **OK**.

Discussion

The ***Rand no*** column will be sorted from lowest to highest and the corresponding rows in the other columns will be automatically moved to match. Rows 2 to 51, or indeed any consecutive 50 rows of all the variables, including ***height (cm)*** and ***foot size,*** will comprise a random sample of 50 responses. It is a good idea to copy these into a new worksheet for separate analysis. This may be done as follows.

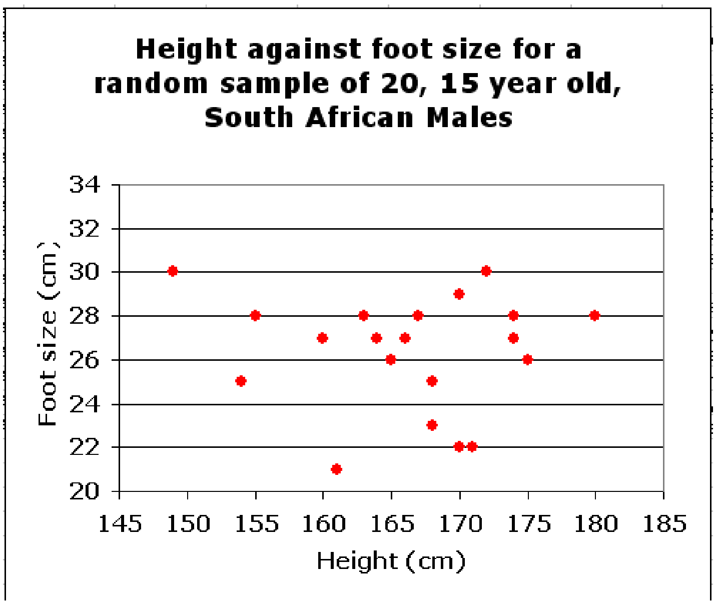
> Right click on the label tab **CSVFile***** at the bottom of the worksheet, where *** will be a number, such as 187.
>
> Click **Insert > OK** and a blank worksheet, **Sheet1**, will be inserted in the file before the **CSVFile*** worksheet**.
>
> Now highlight rows 1 to 51 of the two variables you want, in this case ***height (cm)*** and ***Foot Size***, in the **CSVFile***** worksheet and copy and paste them into columns **A, B,…** of the new **Sheet1** worksheet. If the two columns you want are not next to each other, press **Ctrl** whilst highlighting the second column. You can rename **Sheet1** by right clicking on the name, choosing *Rename* and typing in the new name.

More complex samples can be generated by using Excel's **Sort** facility. For example we can generate a random sample of heights of 20 male and 20 female students from each year group. This can be achieved by using the sequential sorting facilities within the **Sort** function in Excel.

> Click **Data > Sort** and in the **Sort by** box select ***Grade no***, in the first **Then by** box select ***Gender,*** and in the second **Then by** box select ***Rand no,*** and click **OK**.

The data now come in the order grade 3 females, grade 3 males, grade 4 females and so on. The final ***Rand no*** sorting variable has randomised all the responses in each of the previous two categories. So the first 25 responses in the variable ***height (cm)*** matching to the variable ***Gender*** having a cell entry of **F** and the variable ***Grade no*** having a cell entry **3** will be a random sample of female heights from children in grade 3.

Can you plot a scatter graph of height (cm) against foot size for a random sample of 20 South African 15 year olds?



Activities to do with your class: Using Random Samples.

One of the easiest ways to enable your class to learn a lot about variation and statistics is to allow each learner to download their own random sample from the same database. You can give them all the same tasks and activities to do on their random sample but you will need to ensure they share their results with each other either by working in pairs or groups or by asking individuals what results they got in a class discussion. Because they are each working with a random sample from a large database their statistics will show the natural variation that exists between samples. This can very easily lead to thinking through the concepts of the central limit theorem and confidence intervals although this terminology is best not used until the learners really understand what is happening.

Another important question to get your learners to discuss is regarding the size of the sample that is needed in order to make any conclusions or findings reliable. Many learners put far too great a level of importance on a finding or conclusion reached from either a single sample or one that is not large enough.

Example Activities

For the following activities we will investigate the average height of a year 7 (aged 11/12) from the CensusAtSchool phase 7 database. As we do have access to the full database we do actually know in this case what the averages from the entire database of 3576 yr 7 learners are

Mean = **152.14 cm**
Median = **152 cm**
Mode = **150 cm**

Investigation 1

Investigate what effect the size of a single sample has:

Take a number of single random samples of varying sizes and compare the findings.

This is the results of what we found when doing this investigation: (all figures in cm's)

SIZE of SAMPLE	MEAN	MEDIAN	MODE
5	153.4	157	none
10	146.3	151	157
20	150.9	148.5	150
50	150.1	149.5	150
100	152.1	152	150
200	153.2	154	150
500	152.3	152.5	150
1000	151.7	152	150
Full database	152.1	152	150

Investigation 2

Investigate one single sample of 200 compared to 10 repeat samples of 20 and 10 repeat samples of 200

Will a single sample of 200, 10 repeated samples of 20 or 10 repeated samples of 200 give the most accurate and reliable results? What are the advantages and disadvantages of each method?

This can be illustrated very simply using random samples from our databases.

Single sample of 200
The average height in cm of a year 7 learner was:

One sample taken of size 200	Averages found:
	Mean = **152.8377 cm**
	Median = **153 cm**
	Mode = **154 cm**

Ten samples of 20 (all figures in cm's)

Sample	1	2	3	4	5	
Mean	154.775	158.075	153.2	152.212	150.9	
Median	153	157.75	151.5	153	149.5	
Mode	152	169	150	154	146	
Sample	6	7	8	9	10	**Average of averages**
Mean	150.124	153.575	151.4737	155.3	152.265	**153.1**
Median	151.5	155	150	155.5	149.5	**151.5**
Mode	#N/A	154	145	157	143	**154**

Ten samples of 200 (all figures in cm’s)

Sample	1	2	3	4	5	
Mean	152.9	153	153.3	151.9	152.2	
Median	153	152.8	153	152	152	
Mode	150	150	152	152	162	
Sample	6	7	8	9	10	**Average of averages**
Mean	153.3	153.1	152.2	153.6	150.9	**152.64**
Median	154	154	152	154	151	**152.9**
Mode	154	150	152	152	152	**152**

Why is there an N/A for the mode in sample 6 of the second table?

Can you explain how we worked out the average of average figures?

Would taking even more samples help or not?

Discuss the advantages and disadvantages of each method

- which seems to give the most accurate results?
- which is the most work to do?