Using Large Data Sets Workbook – MEI version

This booklet uses Excel and Desmos

This workbook explores the different types of activities that students and teachers might undertake with a Large Data Set so that it can be used effectively to support the learning of statistical concepts. You will need the appropriate MEI dataset.

Key Skills

- Understand the dataset and its context
- Clean a dataset and know how to deal with outliers
- Sort and Filter the dataset
- Produce summary statistics
- Draw frequency charts and box plots for a set of data
- Draw graphs of several datasets side by side for comparison
- Draw scatterplots and plot lines and curves of best fit
- Use technology to calculate correlation coefficients and equations of regression lines
- Take a random sample from a dataset

Software Used

- A spreadsheet (in this case Excel)
- Graphing and statistical software (in this case Desmos).

Becoming familiar with the dataset

Open the excel file which contains the dataset. The first tab gives the source of the data and contains a glossary of terms. Students are required to understand the context of the data so that it important that they read the glossary whilst looking through the dataset. Some questions you might like to consider are:

- *What is the source of the data and how up to date is it?*
- *Who collected it and how was it collected?*
- *What does #N/A mean and why is it used? How should we treat these items when analysing the data? Would we treat some fields differently?*

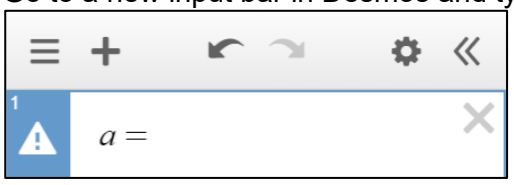
Students need to understand each of the fields and how they are determined. Some of them warrant considerable discussion such as the different Systolic and Diastolic pressures and body mass index. Students should be encouraged to research further so that they fully understand the concepts by looking at the website: *http://www.cdc.gov/nchs/nhanes/about_nhanes.ht*

1 Producing summary statistics

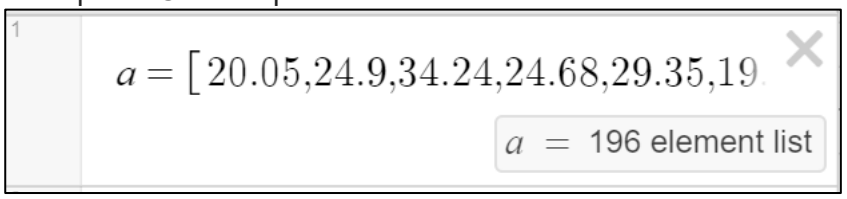
Filter out the #N/A entries from the BMI column and then highlight and copy (Ctrl-C) the remaining data items:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Sex	Age	Marital	Weight	Height	BMI	ThighLen	UpperArm Length	Waist	Food30	Arm	Pulse
43	Female	59	Divorced	106.1	161.7	40.58	37.2	35.2	127.8	Yes	Right	72
44	Male	56	Married	99.2	188.4	27.95	49.6	39.7	110.2	No	Right	74
45	Male	32	Married	84	172.9	28.1	41.7	39.9	94	No	Right	68
46	Male	70	Married	90.8	179.6	28.15	45.9	41.8	104.1	#N/A	#N/A	#N/A
47	Female	60	Married	73.2	177.1	23.34	46	37.1	95.9	Yes	Right	62
48	Male	37	Married	80.3	166.5	28.97	39.1	37.4	101.2	No	Right	50
49	Male	61	Divorced	81.5	170.4	28.07	39.7	37.7	101.1	Yes	Right	70
50	Male	41	Married	92.8	177.2	29.55	45	38.7	109	No	Right	52
51	Male	62	Married	88.5	173.1	29.54	40	40.5	107.8	Yes	Right	60
52	Male	51	Married	91.8	174.8	30.04	41.4	#N/A	106.3	No	Right	62
53	Female	64	Widowed	73.7	159.5	28.97	36.4	36.2	100.3	Yes	Right	76
54	Male	72	Married	91.4	173.2	30.47	40.8	40	109.4	No	Right	64
55	Male	42	Married	82	163.6	30.64	34.5	38.3	106.7	No	Right	50
56	Male	65	Married	86	167.4	30.69	44.1	42.8	106.6	#N/A	#N/A	#N/A

Go to a new input bar in Desmos and type:



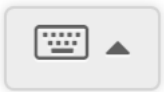
Then press Ctrl-V to paste the data:



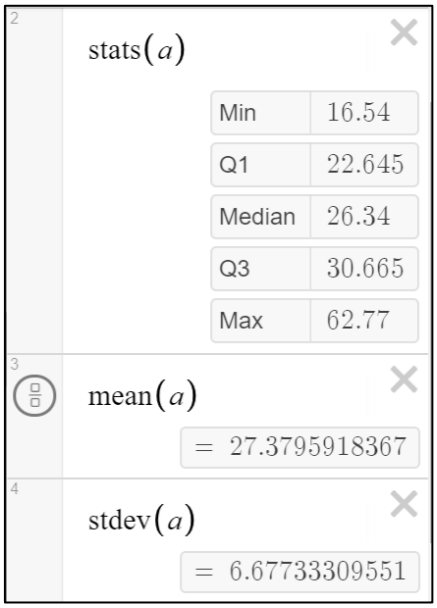
This should create the list (called a).

You can now refer to the list in other commands.

The following commands can be found in functions > Stats from the onscreen keypad:



- stats(a)
- mean(a)
- stdev(a)



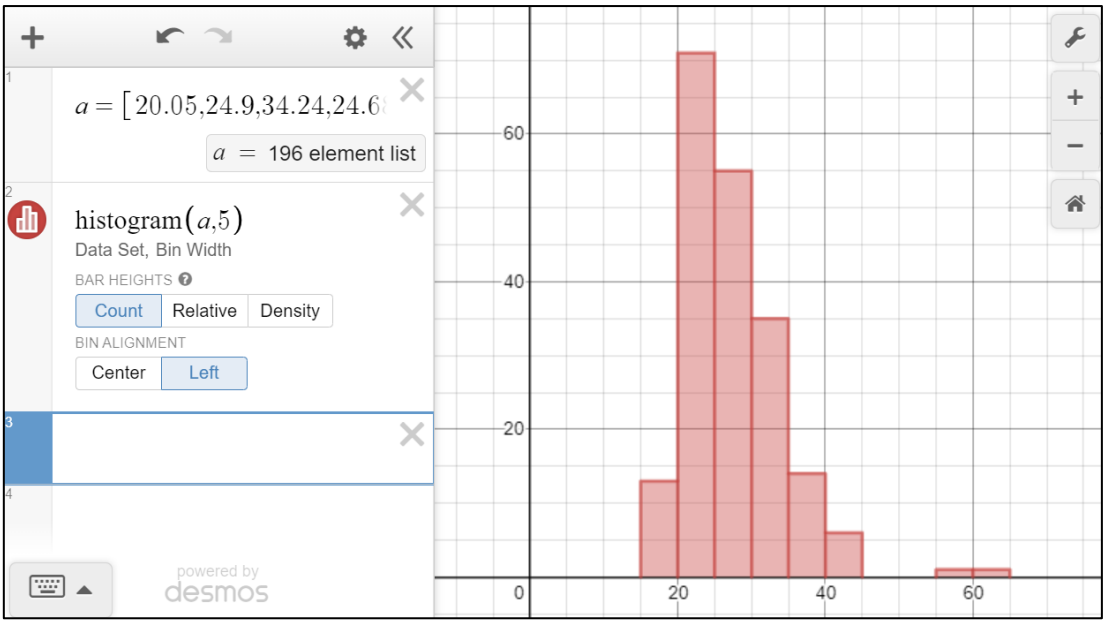
Exercise: Produce these statistics separately for the BMI for Males and the BMI for Females. What similarities or differences do these statistics show?

2 Drawing frequency charts and box plots for a set of data

Desmos can display a range of graphs and charts. You can use the previous steps for copying the data into Desmos and then select a visualization from: functions > Dist

Desmos includes both histograms and boxplots.

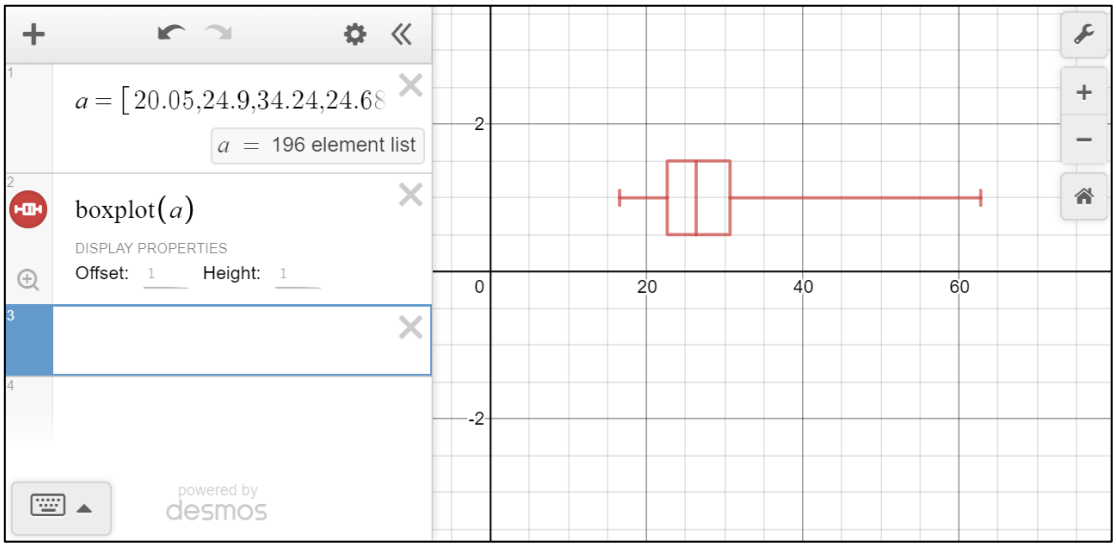
For a histogram you should enter the data set and the bin width.



The magnifying glass icon in the input row can be used to auto-scale. Setting the bin alignment to Left is often more useful. For bins of width 5 the first bin will contain values of x where $0 \leq x < 5$.

Desmos uses a definition of histogram that has frequency on the vertical axis and equal interval widths on the horizontal axis.

For a boxplot enter the data set.



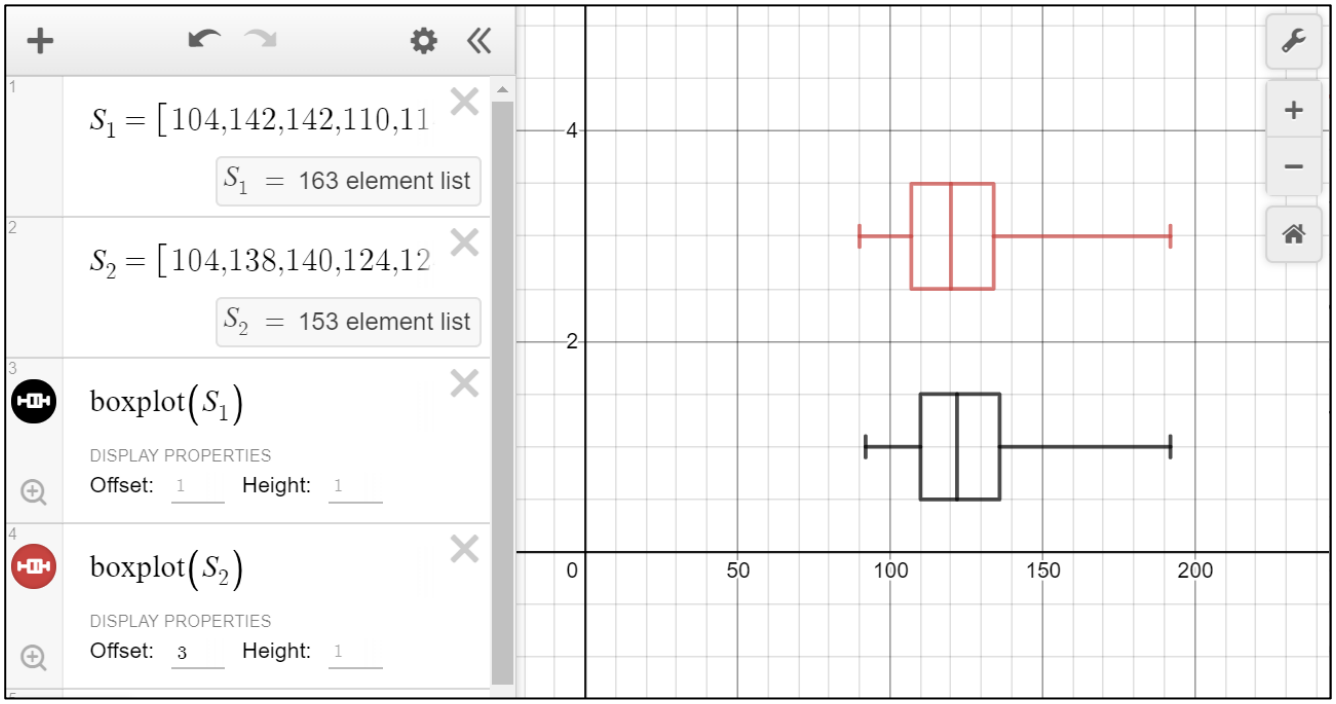
What information does each of these charts show you about the dataset? When might you use a histogram and when might you use a boxplot?

3 Drawing graphs side by side for comparison

The following example compares the first reading with the second reading for Systolic blood pressure.

Each set of numbers will need to be copied as a new list into Desmos.

Filter each set to remove the #N/A values and copy into a new list in Desmos (NB just type S1 to obtain the variable name with the subscript):



Exercise: Draw the boxplot for the third reading. What conclusions can be reached by comparing these plots?

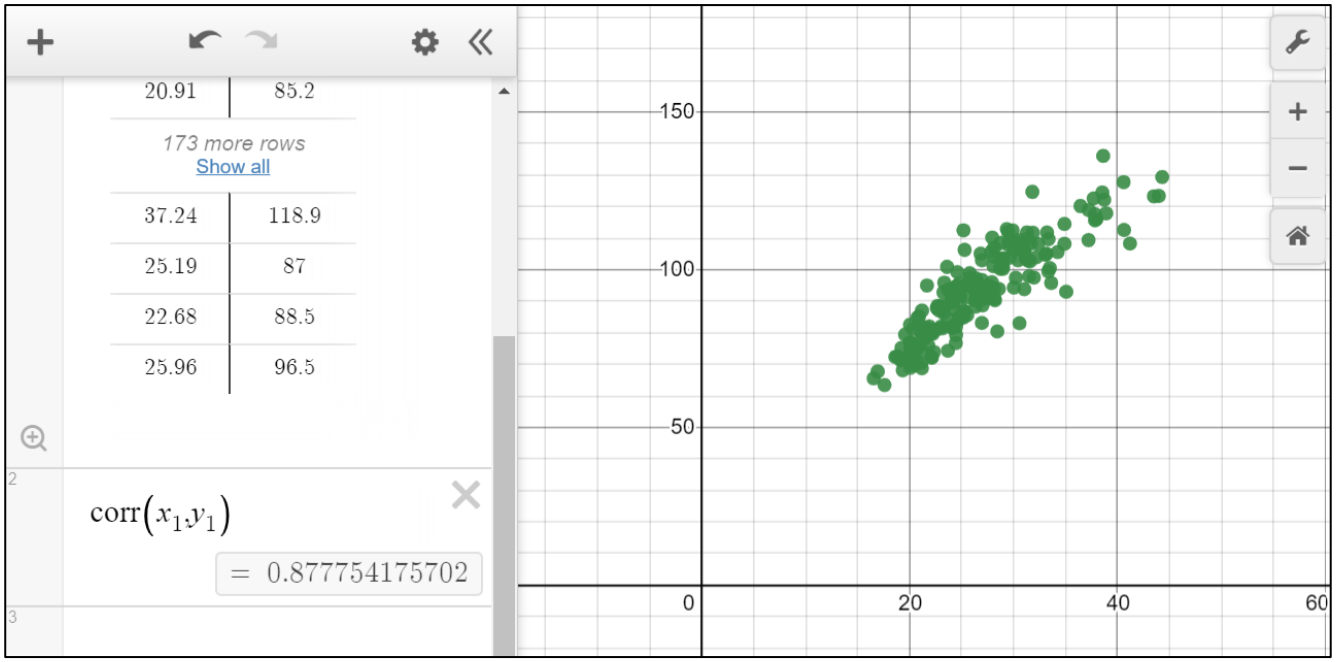
4 Drawing Scatterplots

You can investigate whether there is a relationship between two variables, regarding the data as bivariate data, by copying and pasting two columns of data from Excel into Desmos.

In the following example BMI will be compared against waist size. These two columns have been filtered to remove any #N/A values from either column and then copied by selecting each one separately and copying into a new blank spreadsheet. The pair of adjacent columns in the new spreadsheet can then be copied and pasted into Desmos.

	A	B	C	D	E	F	G	H	I	
1	Sex	Age	Marital	Weight	Height	BMI	ThighLen	UpperArm Length	Waist	F
2	Female	34	Married	60.3	173.4	20.05	41	37.7	82.5	Ye
4	Female	48	Divorced	100.6	171.4	34.24	40.4	39.5	105.6	No
5	Male	61	Married	70.9	169.5	24.68	36.1	36	92.2	No
6	Male	68	Divorced	96.8	181.6	29.35	43.7	41.2	112.9	No
7	Female	28	Never married	50.2	158.5	19.98	37	36.5	71.6	Ye
8	Male	37	Living with partner	115.9	172.6	38.9	41.5	38.1	117.9	No
9	Female	36	Living with partner	54.4	161.3	20.91	42.9	35.8	85.2	No
10	Male	42	Married	71.5	172.3	24.08	41.6	38.8	90.3	No
11	Male	37	Living with partner	86.8	183.5	25.78	44	39.7	98.9	No
13	Male	58	Never married	84.2	169	29.48	38	37.7	112	No
14	Female	54	Divorced	69.5	168.6	24.45	42.6	34.9	76.7	Ye
15	Female	47	Divorced	59.6	156.6	24.3	34.1	32.5	87.4	No
17	Female	77	Never married	96.6	166.4	34.89	#N/A	37	114.5	No
19	Female	64	Widowed	73.7	159.5	28.97	36.4	36.2	100.3	Ye

	A	B
1	BMI	Waist
2	20.05	82.5
3	34.24	105.6
4	24.68	92.2
5	29.35	112.9
6	19.98	71.6
7	38.9	117.9
8	20.91	85.2
9	24.08	90.3
10	25.78	98.9
11	29.48	112
12	24.45	76.7
13	24.3	87.4
14	34.89	114.5
15	28.97	100.3



The Scatterplot shows some positive correlation between the two variables. This can be confirmed by calculating: $\text{corr}(x_1, y_1)$

NB for subscripts typing $x1$ will automatically be updated to x_1.

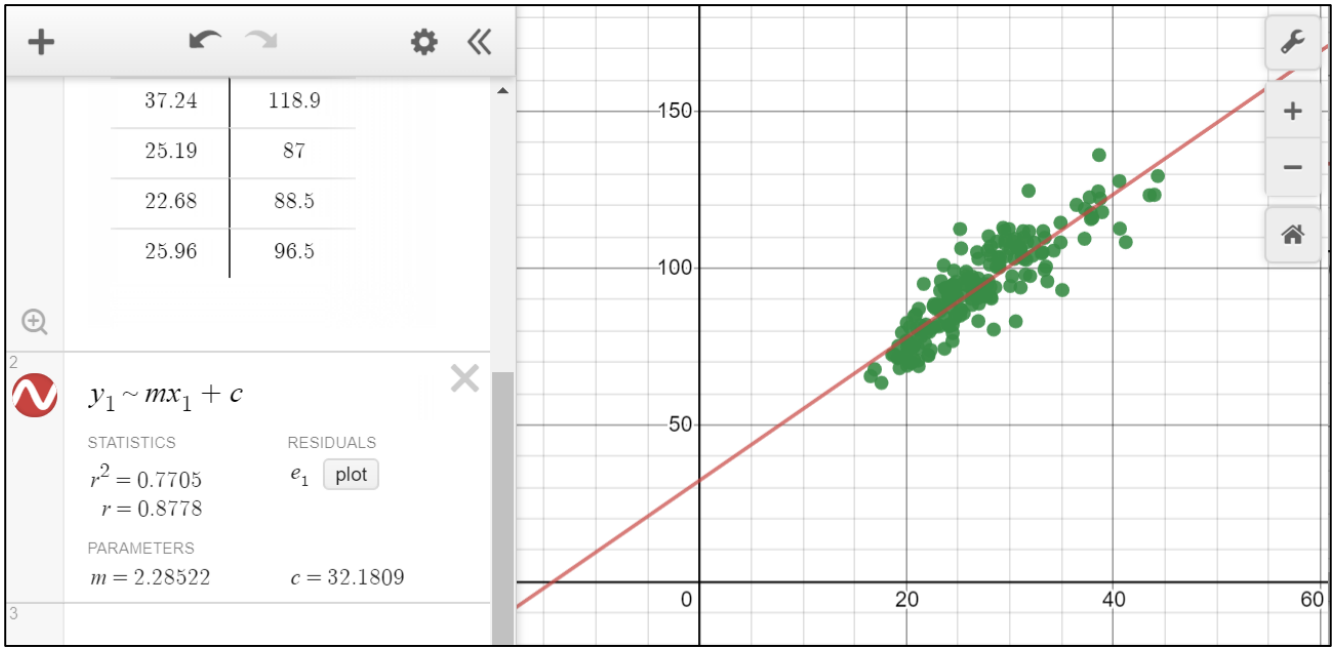
5 Regression models

Desmos allows for any regression model that can be defined as a generalised function. In practice students will mainly use a linear model: $y = mx + c$.

Bivariate data can be copied into Desmos as described in section 6.

The regression model is defined using the tilde symbol, ~, e.g. $y_1 \sim mx_1 + c$.

The example below shows this for the data used in the previous section: BMI will be compared against waist size.



Students are expected to use (but not derive) non-linear models for data. For example you might consider a log model instead.

The residual sum of squares (RSS), also known as the sum of squared errors of prediction (SSE) gives a guide to how good a fit the model will be. The residuals can be plotted using the plot button.

Exercise: Explore the correlation between BMI and waist size separately for males and females. Is the amount of correlation similar?

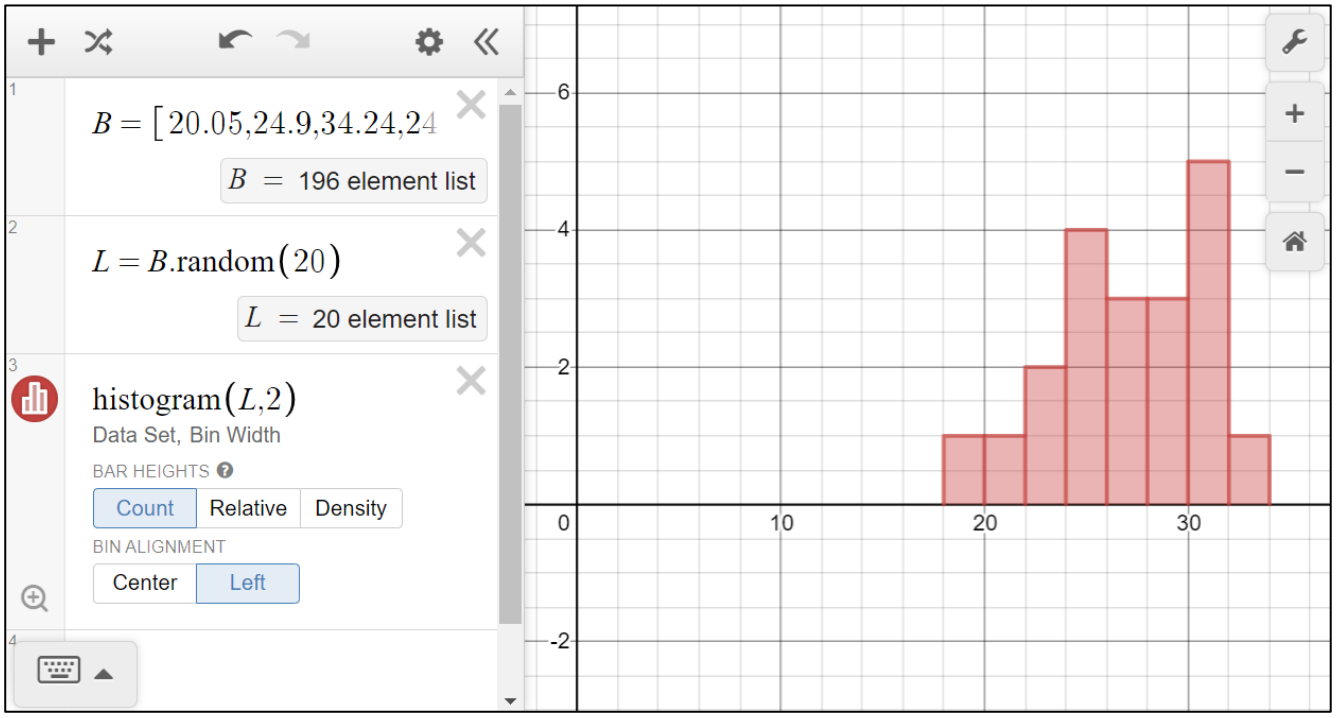
6 Random Sampling

You can use the random function to select a random sample from a list of values. In the following example a random sample of size 20 will be selected from BMI data.

Enter a new variable of B and copy the full list into Desmos (with the #NA values filtered out).

To create a random sample of size 20 type: $L = B.\text{random}(20)$

A histogram for this sample can be plotted using $\text{histogram}(L, 2)$.



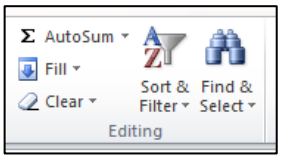
To generate different samples click on  (second icon top-left).

One application of this method is in selecting several samples of the same size and comparing statistics such as the mean or the standard deviation with their true values in the whole dataset. This illustrates the idea of statistical variation. The sample size can then be changed to see how that affects the variation.

Appendix 1: Sorting and filtering the dataset in Excel

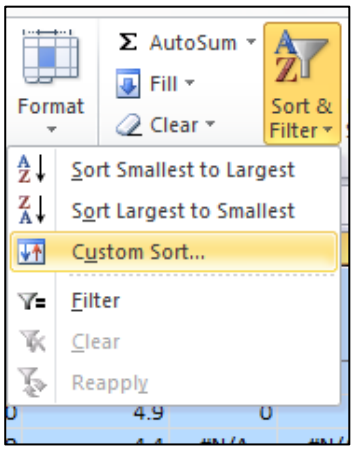
Further familiarity with the dataset can be gained by sorting and filtering the data within Excel. This can help identify any possible outliers or rogue values.

These functions can be found at the far end of the top toolbar:

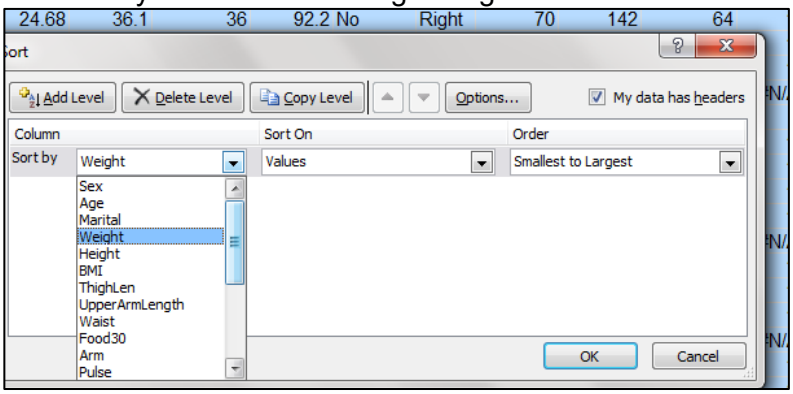


Suppose we want to sort the data according to weight. Use Crtl-A to select all the data.

Select the custom sort option:



When the dialogue box appears select the field that you want to sort on and specify the order, smallest to largest. Also make sure that the 'My data has headers' box is checked otherwise your column headings will get sorted as well.

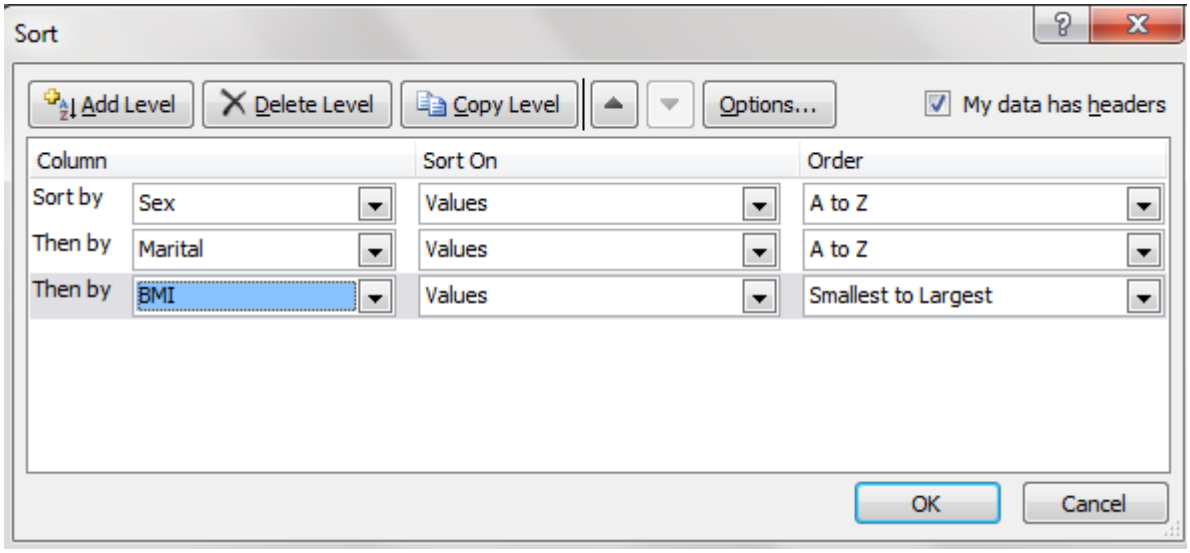


The data is now sorted in order of weight and the #N/A values appear at the bottom.

Sex	Marital	Weight	Height	BMI	ThighLen	UpperArm Length
Female	Married	121.7	173.1	40.62	46	42
Female	Married	122.5	166.3	44.29	40	40
Female	Never married	128.5	171.9	43.49	37.4	43.2
Female	Separated	132.2	153.7	55.96	#N/A	35.5
Male	Divorced	139	193.2	37.24	49.5	41
Male	Married	193.1	175.4	62.77		
Female	Never married	#N/A	#N/A	#N/A	#N/A	#N/A
Male	Never married	#N/A	#N/A	#N/A	#N/A	#N/A
Female	Married	#N/A	#N/A	#N/A	#N/A	#N/A
Male	Married	#N/A	#N/A	#N/A	#N/A	#N/A

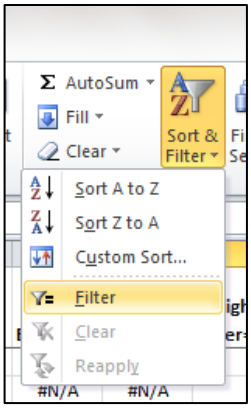
The last entry (before those with #N/A) stands out and warrants further discussion. Is it an outlier?

It is possible to sort the data using several fields using the 'Add Level' button:

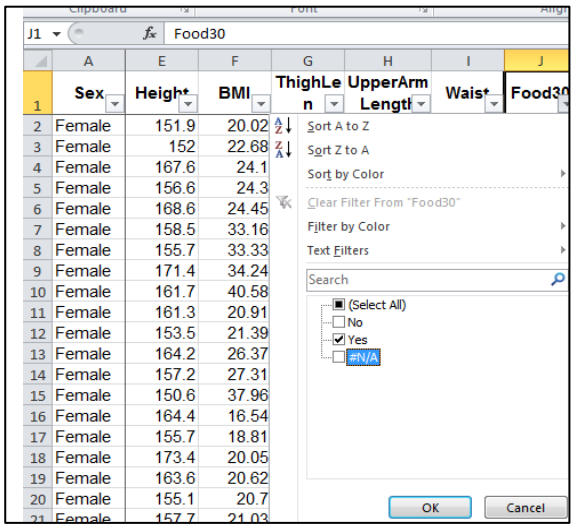


Try the above sort (remember to select all the data first using Crtl-A). Sorting like this might be useful if we wanted to compare the BMI of different genders or marital statuses.

An alternative would be to apply a filter Click on filter and an arrow should appear next to each heading:



Click on the arrow next to Food 30 and then scroll down and uncheck the box next to #N/A and No:



Now we have filtered out some of the records and we are just left with those people who had food 30 minutes prior to the measurements being taken.

A	E	F	G	H	I	J
Sex	Height	BMI	ThighLen	UpperArmLength	Waist	Food30
Female	152	22.68	35	31.7	88.5	Yes
Female	167.6	24.1	39.6	35.5	82	Yes
Female	168.6	24.45	42.6	34.9	76.7	Yes
Female	155.7	33.33	39	35.2	99.5	Yes
Female	161.7	40.58	37.2	35.2	127.8	Yes
Female	155.7	18.81	33.7	31.8	72.5	Yes
Female	173.4	20.05	41	37.7	82.5	Yes
Female	163.6	20.62	37.3	34.2	74.6	Yes
Female	165.6	21.19	39.1	37.2	87	Yes
Female	154.3	21.92	#N/A	#N/A	#N/A	Yes

To turn the filters off click on the filter button again.

Exercise: Sort the data by age. This data is a random sample of 200 from a larger sample of 5000. Is the distribution of ages as you would expect? What method of sampling might have been used?

Appendix 2: Suggested Large Data Set Investigations

Large Data Set 7

- Which regions have the highest life expectancy?
- Is there a correlation between life expectancy and GDP?
- Is there a correlation between land area and population?
- Is there a correlation between birth rate and GDP?
- Do countries that spend more of their GDP on health have higher birth rates/ life expectancy?
- Does a higher physician density have a positive impact on countries?
- Does a high unemployment rate have a negative impact on countries?
- Are there regional differences in any of the categories?

Large Data Set 5

- For the most recent years how does male and female life expectancy for the inner and outer London boroughs compare to other regions in the country?
- Is there a correlation between median house price and life expectancy?
- Is there a correlation between the percentage of students achieving 5+ GCSEs at grades A*-C and the employment rate?
- How has the recycling rate for inner and outer London boroughs changed over time? How does this compare to other regions?
- Which boroughs have shown the most change in median house price between 2004 and 2015?
- Is life expectancy increasing?
- Which London boroughs are most similar to the UK national average for these data?
- Are there regional differences in the change in any of the categories over time?

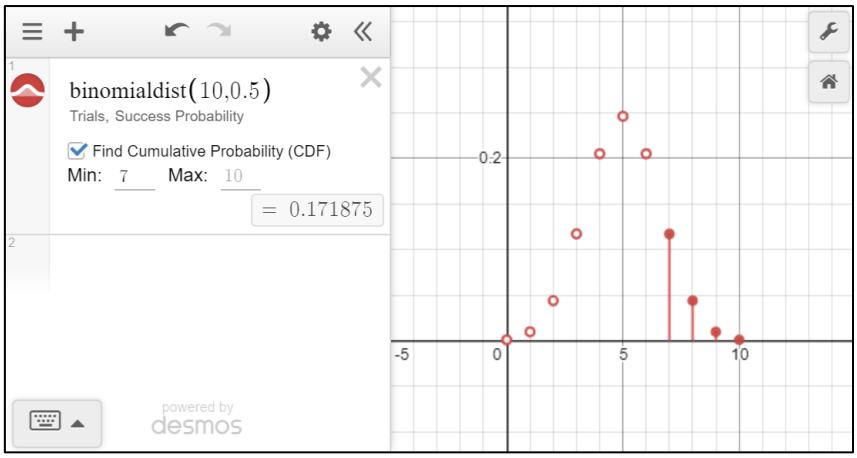
Large Data Set 6

- Does marital status have an impact on BMI?
- Is there a correlation between thigh length and arm length?
- Is there a correlation between waist size and blood pressure?
- Is there a correlation between BMI and pulse rate?
- Do people who have eaten in the last 30 minutes have a higher pulse rate/blood pressure?
- Does marital status have an impact on health?
- Are there gender differences in any of the categories?
- Are there age-related differences in any of the categories?

Appendix 3: Using Desmos for distributions

Binomial distribution

- Select: functions > Dist > binomialdist
- Enter the number of trials and probability of success.
- To calculate the probability of a range select CDF and set the limits.



Normal distribution

- Select: functions > Dist > normaldist
- Enter the mean and standard deviation.
- To calculate the probability of a range select CDF and set the limits.