statstutor

community project

encouraging academics to share statistics support resources

Statistical Methods
7. Introduction to SPSS

Based on materials provided by Coventry University and Loughborough University under a National HE STEM Programme Practice Transfer Adopters grant

www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Summary

- What is SPSS?
- Opening and closing SPSS
- The SPSS windows
- Uploading a data set from Excel into SPSS
- Creating boxplots
- Creating histograms with fitted normal curves:
  - Skewness and kurtosis
  - Commenting on the normality of the distribution
  - Splitting the data sets

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

What does SPSS stand for?

- Originally:

  Statistical Package for the Social Sciences

- Then:

  Statistical Product and Service Solutions

- Then (version 18, under IBM):

  Predictive Analytics SoftWare (PASW)

- Now (version 19 onwards, still under IBM):

  Back to SPSS

SPSS is a widely used statistical package in UK Higher Education

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

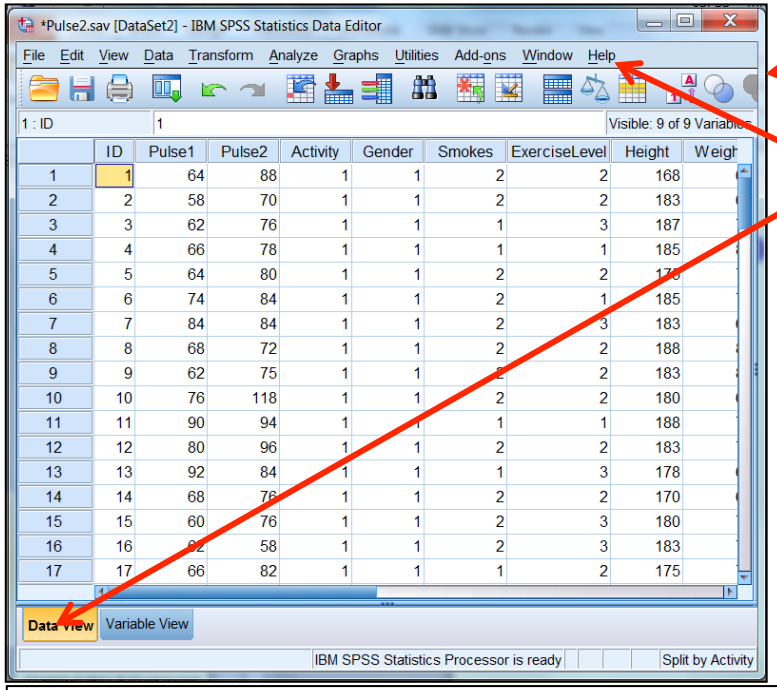
What is SPSS?

- Highly sophisticated statistical software package
- These slides relate to version 19
- Main elements:
  - Data Editor windows – Data and Variable Views
  - Output window
  - Menu-based interface on both windows – but sometimes need to use a specific data window
  - Graphics window appears when editing charts

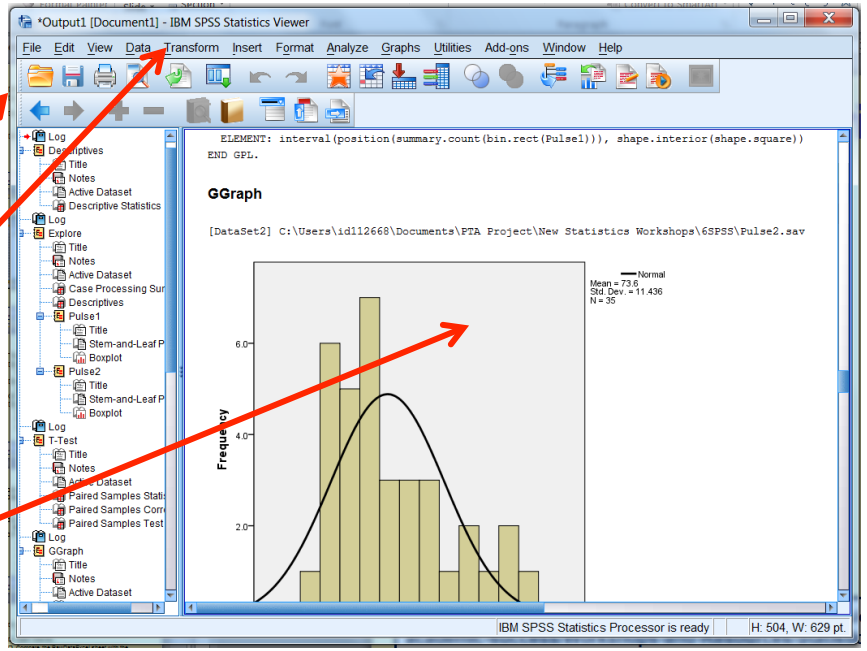
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Main SPSS windows



- Data Editor window
- Currently in *Data View*
- Menu interface
- You can have several data windows open at once



- Output window (only one)
- Also has a menu interface (though better to use the Data Editor window one)
- You can copy and paste output objects into Word

www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Starting SPSS

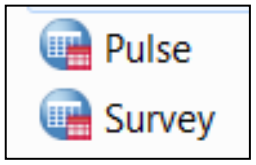
To start SPSS, either:

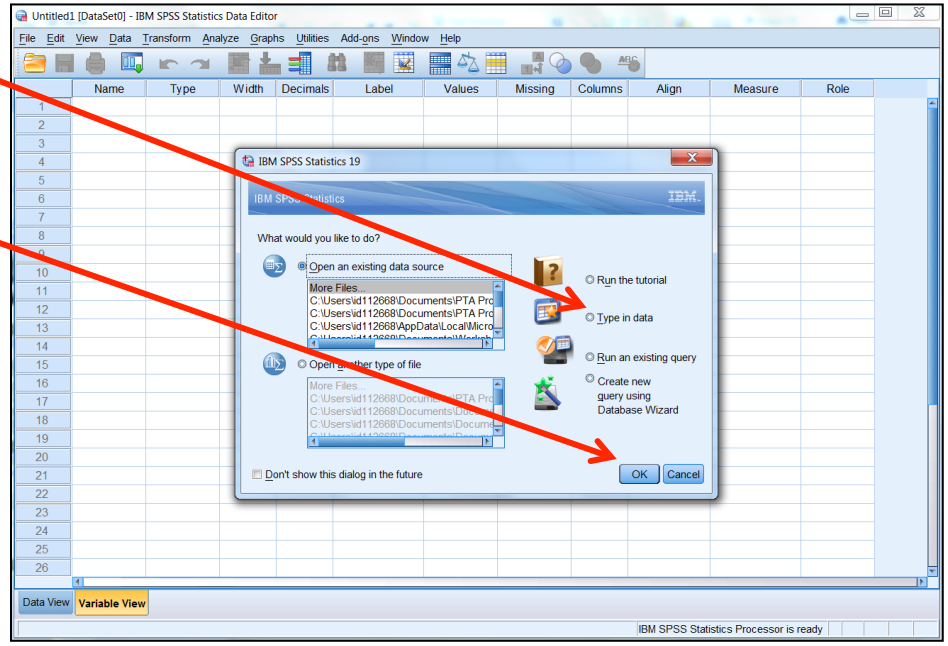
- Select Start – All Programs – IBM SPSS Statistics (or similar) – IBM SPSS Statistics 19

- Then Select *Type in Data* and *OK*

Or:

- Double click on an SPSS data file – these have icons that look like this:





www.**statstutor**.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Pulse data set

- 91 students were asked to take their pulse for a minute
- Each student was then asked to toss a coin:
  - If it came up heads they ran on the spot for a minute
  - If it came up tails they sat for a minute
- At the end of the minute they all took their pulse rates again for a minute
- They also supplied additional personal information:
  - Gender
  - Smoking habits
  - Normal exercise level
  - Height
  - Weight

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Excel file of data set

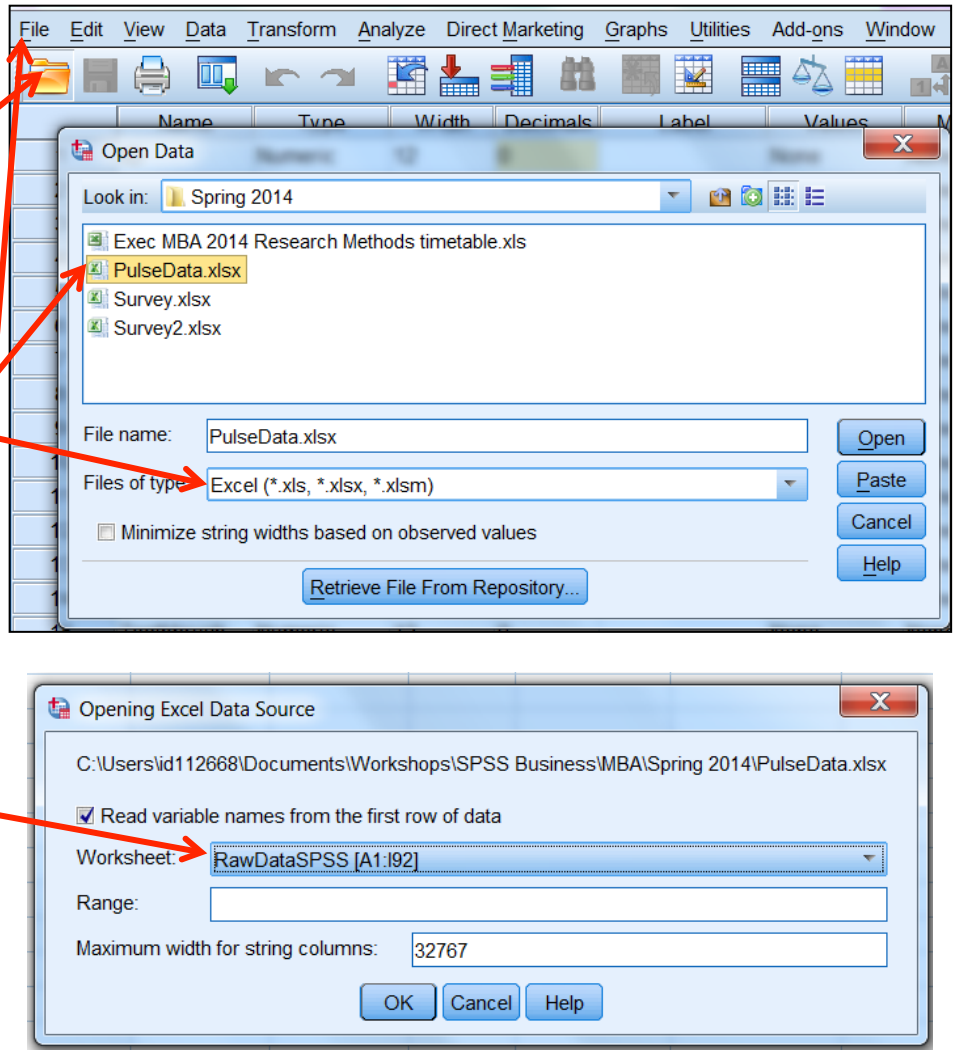
- Save a copy of the file PulseData.xlsx associated with this presentation
- Open this file
- Compare the RawDataExcel sheet with the RawDataSPSS sheet – what are the differences?
- We advise using two data sheets for statistical analysis:
  - Excel is generally better for simple descriptive statistics
  - Words work better with Excel data – automatically generates pivot table labels for use in charts
  - SPSS is generally better for more complex descriptives and statistical testing
  - SPSS prefers number-based data

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Set up the data file in SPSS

- Close the Excel file PulseData
- In the SPSS data window, select the open button
- Change the file type to Excel
- Locate the file PulseData and select Open
- Select the sheet RawDataSPSS and click on OK
- Save the data file as Pulse1.sav using File – Save As

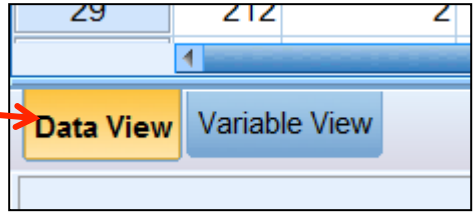


www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

In the Data View:



- Resize the columns by placing the cursor between the columns and clicking and dragging

	ID	Pulse1	Pulse2	Activity	Gender	Smokes
1	1	64	88	1	1	2
2	2	58	70	1	1	2
3	3	62	76	1	1	1
4	4	66	78	1	1	1
5	5	64	80	1	1	2

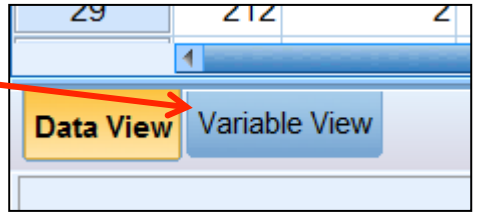
- The sheet should then look like this:

	ID	Pulse1	Pulse2	Activity	Gender	Smokes	ExerciseLevel	Height	Weight
1	1	64	88	1	1	2	2	168	63
2	2	58	70	1	1	2	2	183	66
3	3	62	76	1	1	1	3	187	73
4	4	66	78	1	1	1	1	185	86
5	5	64	80	1	1	2	2	175	70
6	6	74	84	1	1	2	1	185	75
7	7	84	84	1	1	2	3	183	68
8	8	68	72	1	1	2	2	188	86

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

In the Variable view:



- Most of the settings SPSS has assumed are OK:

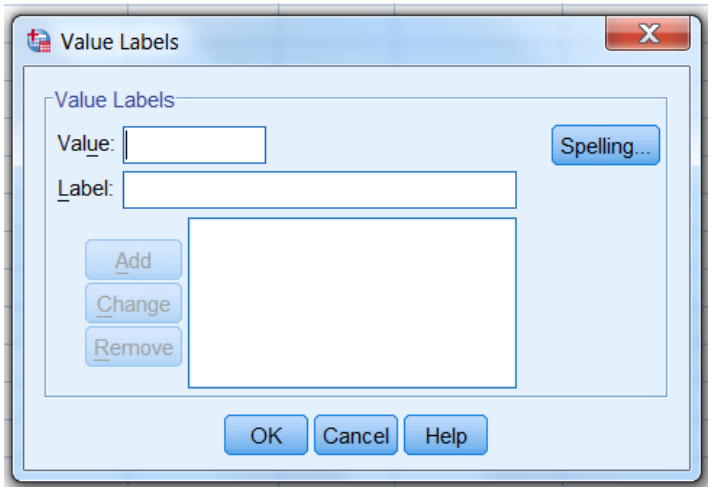
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	ID	Numeric	12	0		None	None	3	Right	Scale	Input
2	Pulse1	Numeric	12	0		None	None	6	Right	Scale	Input
3	Pulse2	Numeric	12	0		None	None	6	Right	Scale	Input
4	Activity	Numeric	12	0		None	None	5	Right	Nominal	Input
5	Gender	Numeric	12	0		None	None	6	Right	Nominal	Input
6	Smokes	Numeric	12	0		None	None	7	Right	Nominal	Input
7	ExerciseLe...	Numeric	12	0		None	None	10	Right	Nominal	Input
8	Height	Numeric	12	0		None	None	5	Right	Scale	Input
9	Weight	Numeric	12	0		None	None	6	Right	Scale	Input

- Widen the column Name as before
- Change the Measure for *ID* to ordinal
- Select the Values field for *Activity* then click on the "…" button

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

- This dialogue box should appear:



- Enter the Value *1* and the Label *Run* then click on Add
- Enter the Value *2* and the Label *Sit* then click on Add
- Click on OK
- Repeat for *Gender* with *1* for *Male* and *2* for *Female*
- Repeat for *Smokes* with *1* for *No/Not Regular* and *2* for *Regular*
- Repeat for *ExerciseLevel* with *1* for *Slight*, *2* for *Moderate* and *3* for *Heavy*

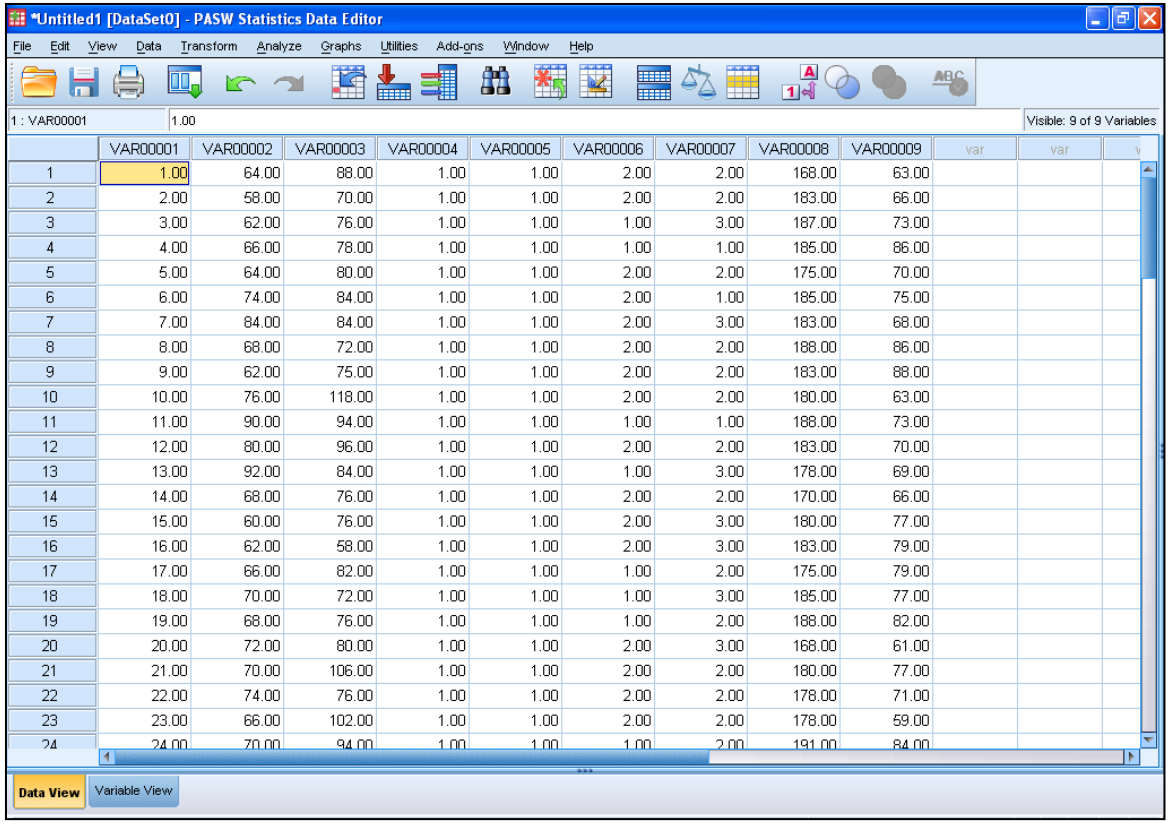
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Copying raw data into SPSS

- Go to the RawDataSPSS sheet of the PulseData Excel file
- Select the cells from A3 to L93 and press copy
- Go to SPSS and Select the Data View
- Select the cell in the top left corner
- Select Edit – Paste from the SPSS window

The Data View window should then look like this:



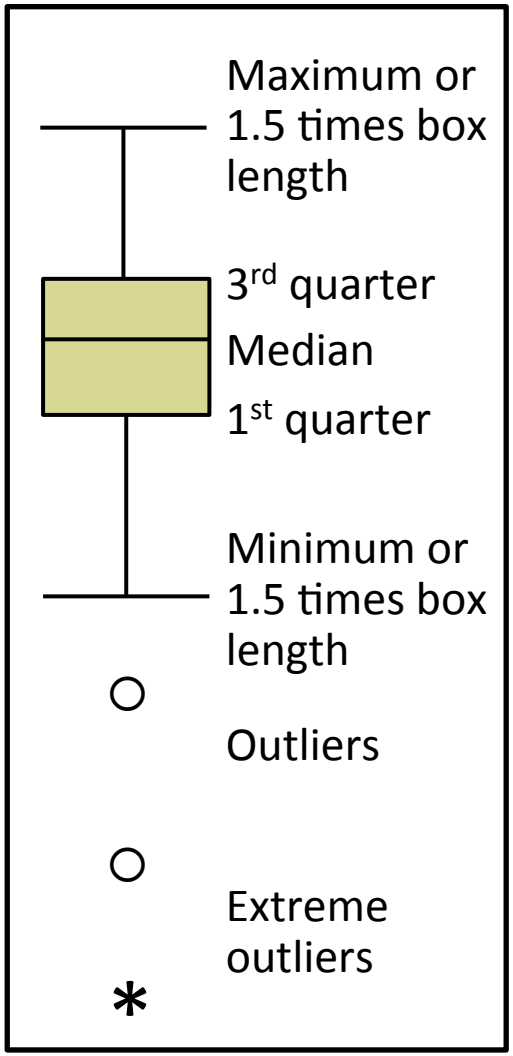
	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007	VAR00008	VAR00009
1	1.00	64.00	88.00	1.00	1.00	2.00	2.00	168.00	63.00
2	2.00	58.00	70.00	1.00	1.00	2.00	2.00	183.00	66.00
3	3.00	62.00	76.00	1.00	1.00	1.00	3.00	187.00	73.00
4	4.00	66.00	78.00	1.00	1.00	1.00	1.00	185.00	86.00
5	5.00	64.00	80.00	1.00	1.00	2.00	2.00	175.00	70.00
6	6.00	74.00	84.00	1.00	1.00	2.00	1.00	185.00	75.00
7	7.00	84.00	84.00	1.00	1.00	2.00	3.00	183.00	68.00
8	8.00	68.00	72.00	1.00	1.00	2.00	2.00	188.00	86.00
9	9.00	62.00	75.00	1.00	1.00	2.00	2.00	183.00	88.00
10	10.00	76.00	118.00	1.00	1.00	2.00	2.00	180.00	63.00
11	11.00	90.00	94.00	1.00	1.00	1.00	1.00	188.00	73.00
12	12.00	80.00	96.00	1.00	1.00	2.00	2.00	183.00	70.00
13	13.00	92.00	84.00	1.00	1.00	1.00	3.00	178.00	69.00
14	14.00	68.00	76.00	1.00	1.00	2.00	2.00	170.00	66.00
15	15.00	60.00	76.00	1.00	1.00	2.00	3.00	180.00	77.00
16	16.00	62.00	58.00	1.00	1.00	2.00	3.00	183.00	79.00
17	17.00	66.00	82.00	1.00	1.00	1.00	2.00	175.00	79.00
18	18.00	70.00	72.00	1.00	1.00	1.00	3.00	185.00	77.00
19	19.00	68.00	76.00	1.00	1.00	1.00	2.00	188.00	82.00
20	20.00	72.00	80.00	1.00	1.00	2.00	3.00	168.00	61.00
21	21.00	70.00	106.00	1.00	1.00	2.00	2.00	180.00	77.00
22	22.00	74.00	76.00	1.00	1.00	2.00	2.00	178.00	71.00
23	23.00	66.00	102.00	1.00	1.00	2.00	2.00	178.00	59.00
24	24.00	70.00	94.00	1.00	1.00	1.00	2.00	191.00	84.00

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Box and whisker diagrams (boxplots)

- Give a good 'picture' of scale data
- Box ends are the lower and upper quarters
- Box middle line is the median
- Whiskers are the minimum/maximum value of the data or 1.5 times the box height from the end of the box, whichever is less
- Circles represent outliers (up to 3 times the box height from the end of the box)
- Asterisks represent extreme outliers (>3 times box height from the end of the box)



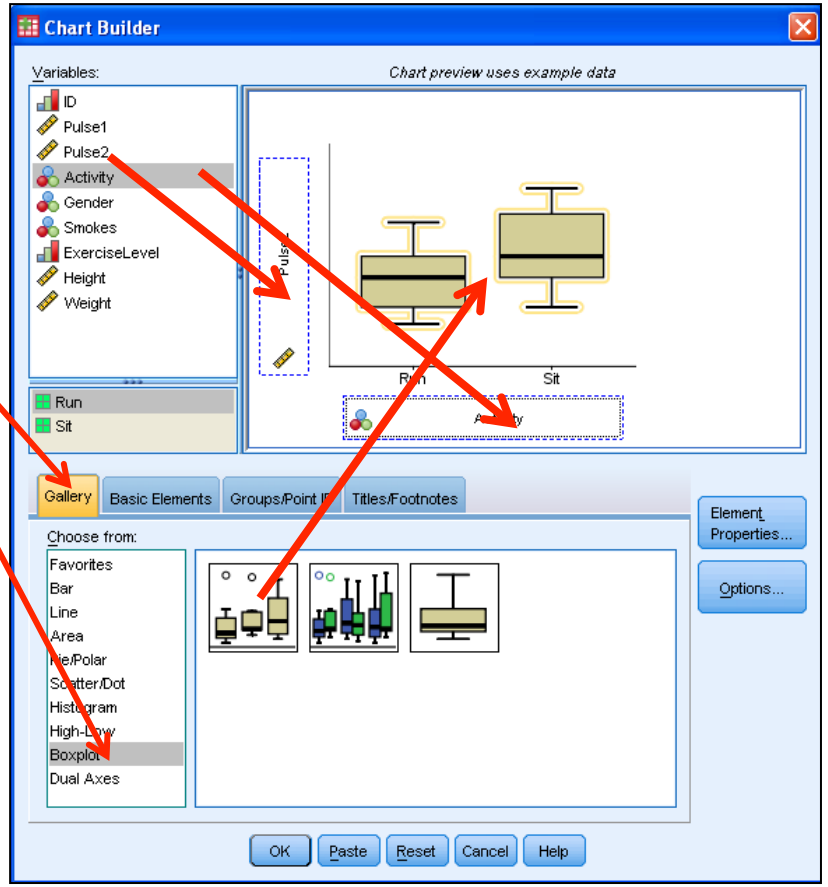
www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Creating boxplots in SPSS

- From the menus choose: Graphs > Chart Builder...
- Select the *Gallery* tab
- Select *Boxplot* in the *Choose from:* list
- Drag and drop the *Simple Boxplot* icon into the canvas area of the Chart Builder
- Drag and drop the variable *Activity* onto the X-Axis
- Drag and drop the variable *Pulse2* onto the Y-Axis
- Repeat with *Gender*, *Smokes* and *ExerciseLevel* on the X-Axis

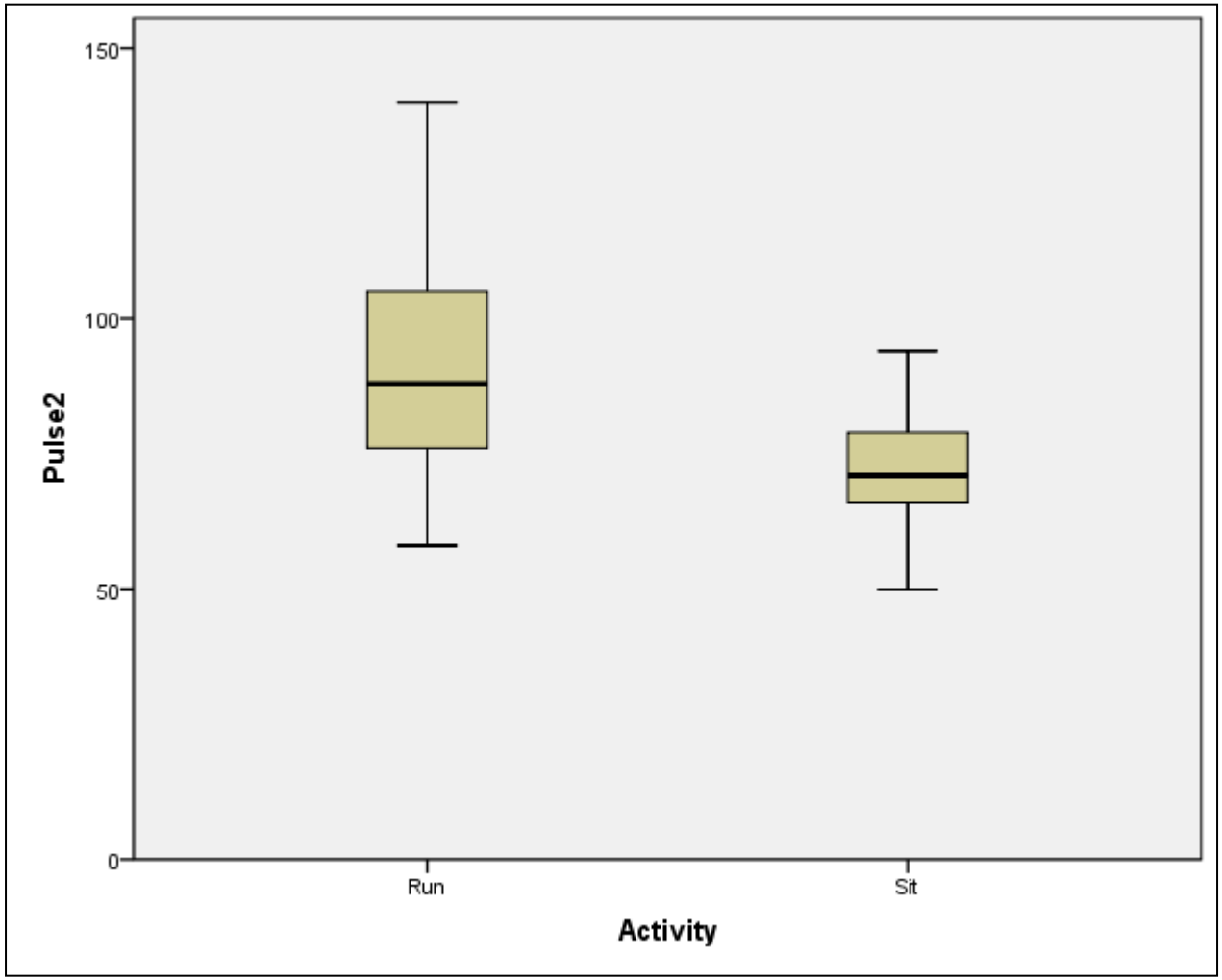


www.statstutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Pulse2 v. *Activity*



Labels appear on the X-axis

Running seems to have increased the pulse rate, as expected.

The spread of running data is also wider.

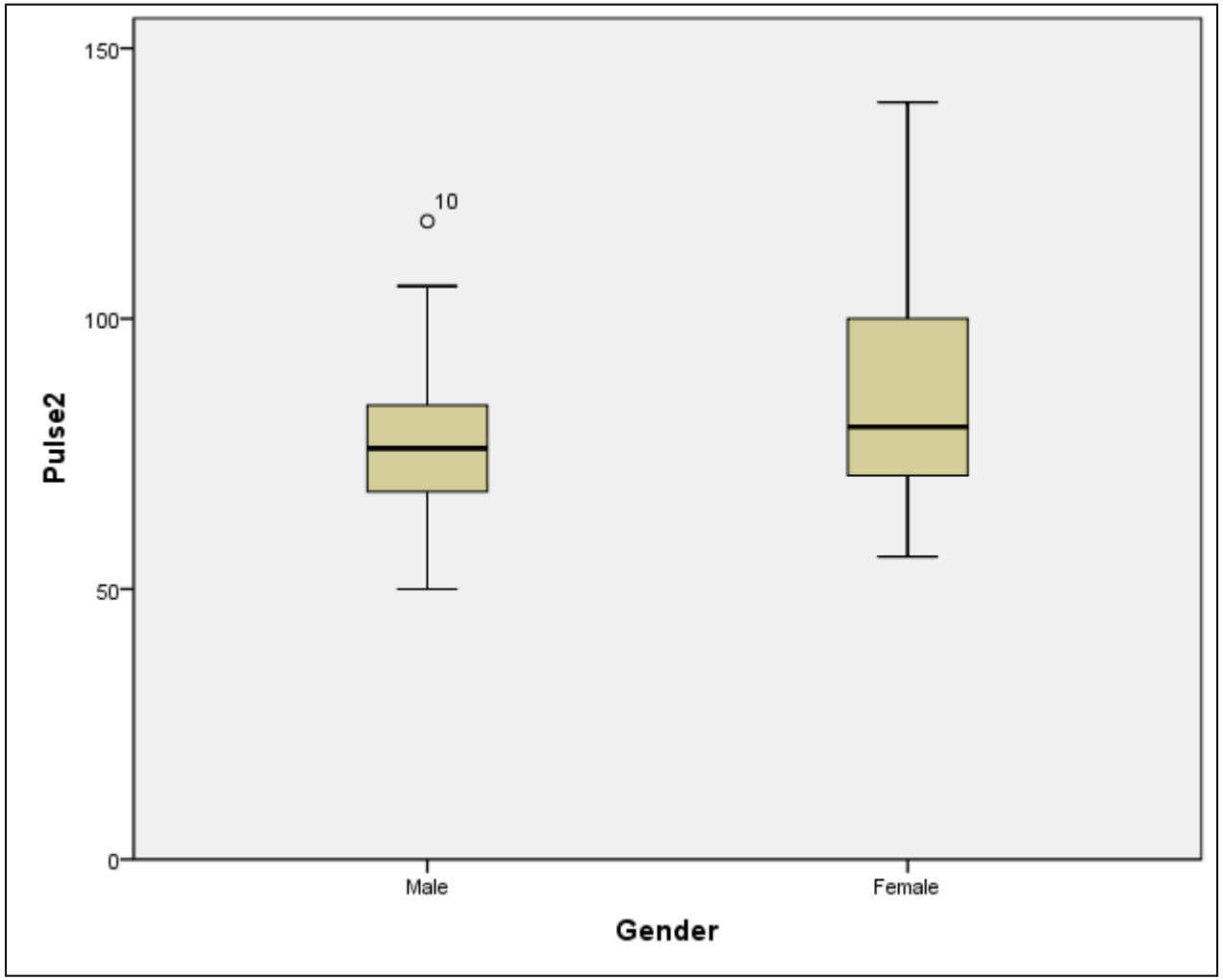
Both boxes are wider above the median than below (skewed).

www.**stats**tutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Pulse2 v. *Gender*



3rd quarter and upper whisker for Female clearly higher than for Male.

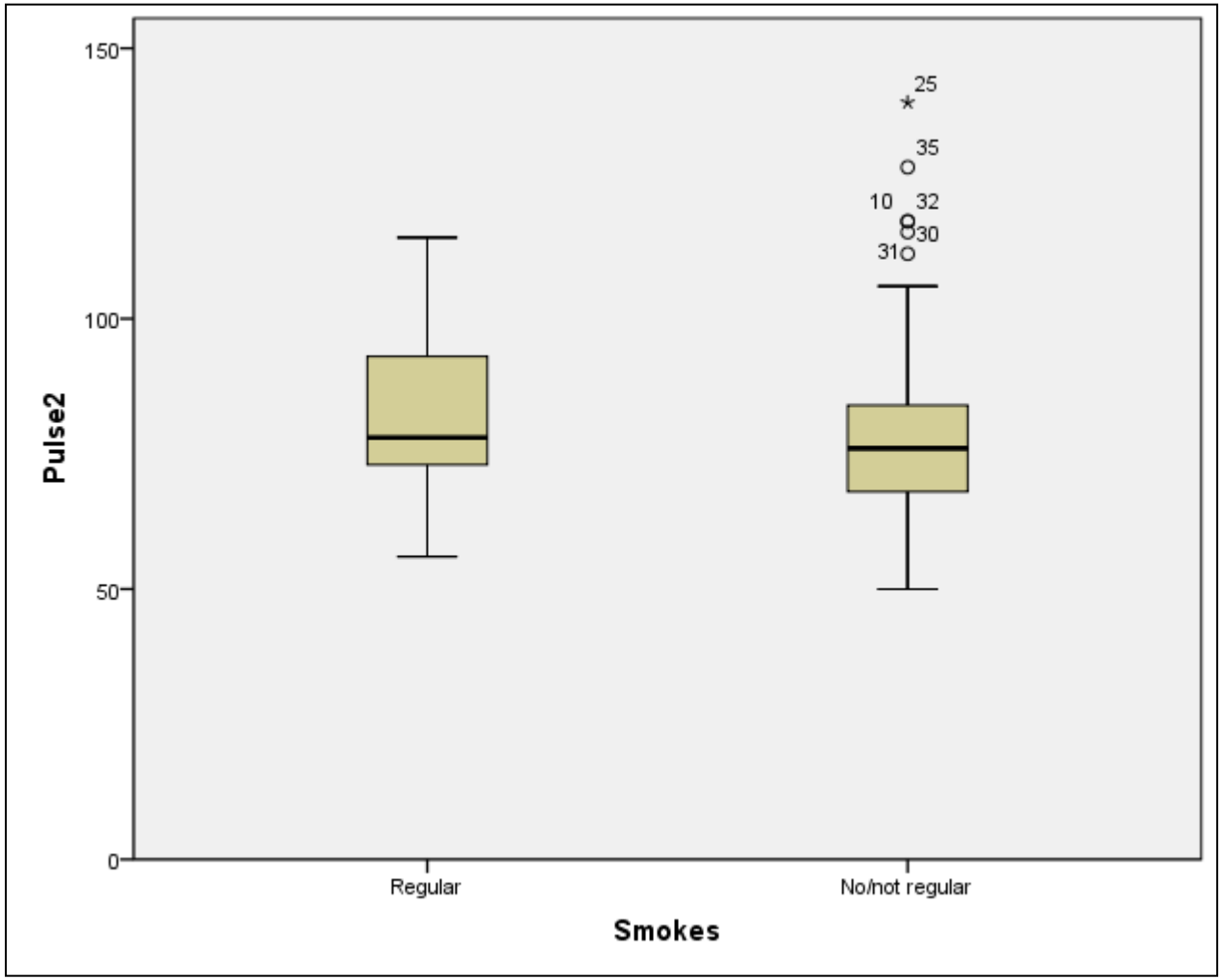
Male distribution much more symmetrical than the Female one.

www.**stats**tutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Pulse2 v. *Smokes*



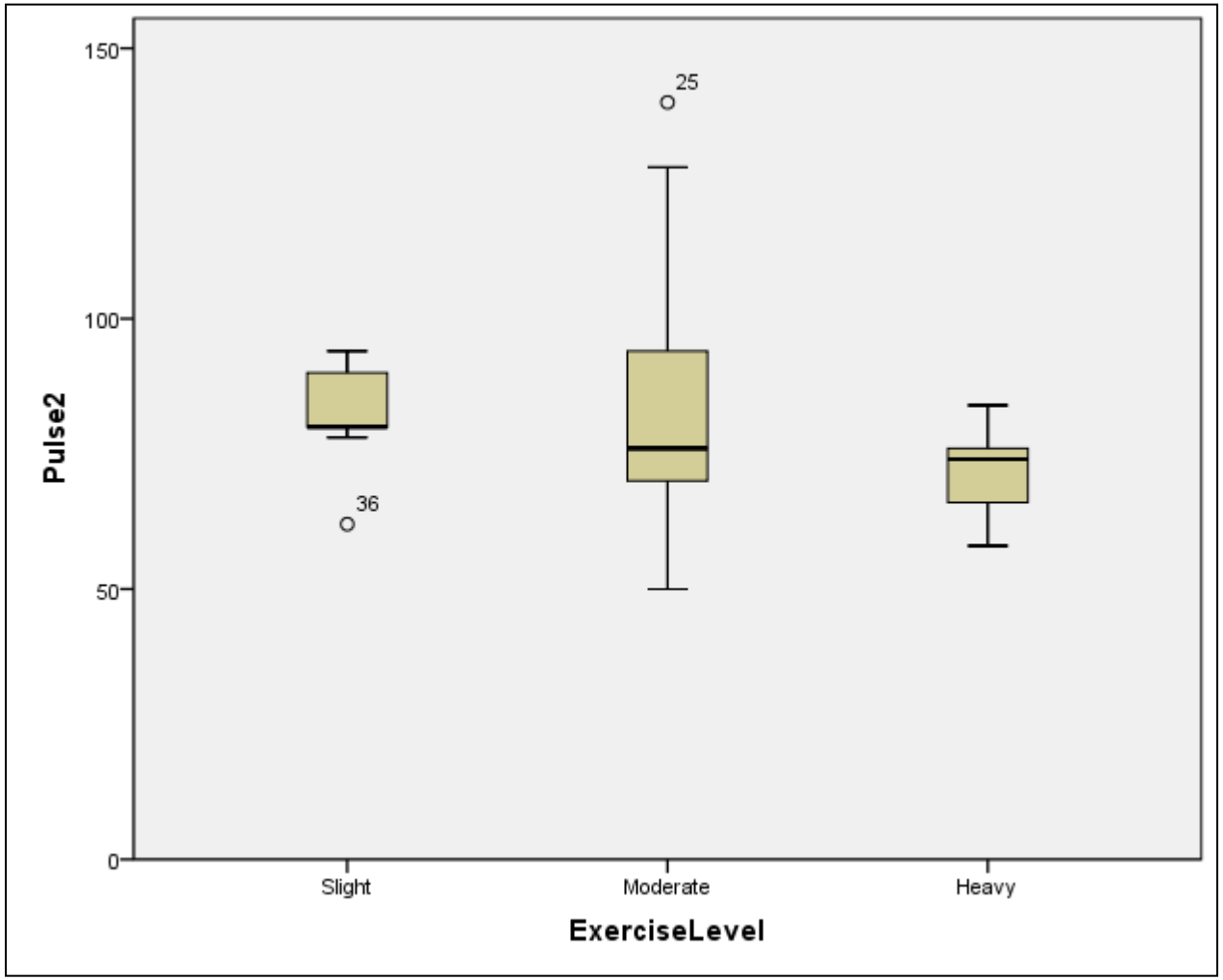
Regular smokers seem to have a slightly higher pulse rate, although there are a lot of outliers with high pulse rates for the No/not regular smokers: interpretation?

www.**stats**tutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Pulse2 v. *ExerciseLevel*



Medians go down with increase in *ExerciseLevel*, as expected.

Much wider spread for Moderate *ExerciseLevel*: interpretation?

www.**stats**tutor.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Exercise

- Create boxplots for *Pulse1*, *Height* and *Weight* against *Activity*, *Gender*, *Smokes* and *ExerciseLevel* (i.e. 12 boxplot charts altogether)
- Comment informally on differences between groups in these plots
- Which differences might have been expected?
- Which differences are surprising?

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

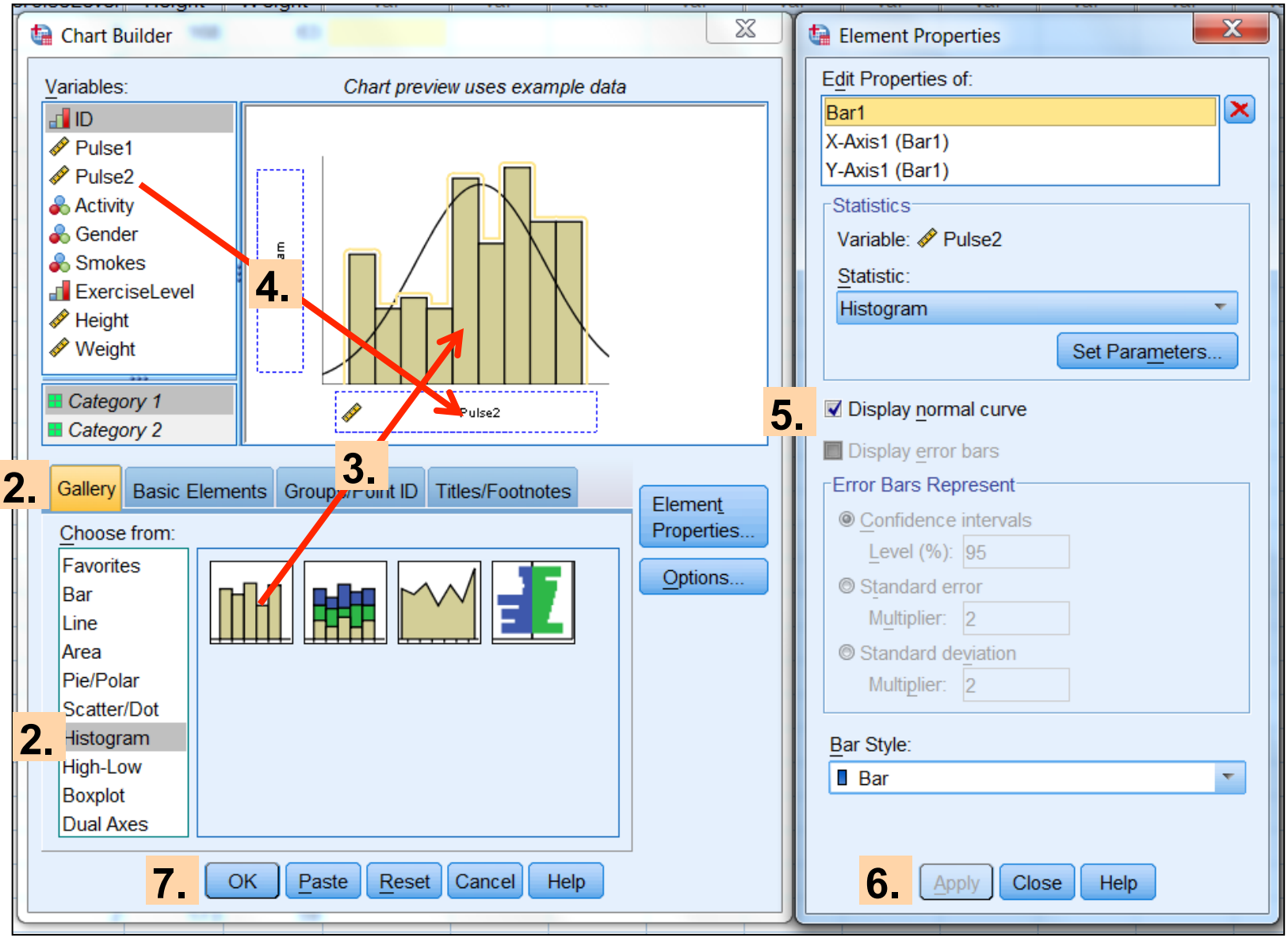
Histograms with fitted normal curves

- SPSS provides a facility to fit a normal curve estimate to a histogram
- This can be very useful when checking to see if a data set is normally distributed

1. Select *Graphs – Chart Builder…*
2. Select the *Gallery* tab and select *Histogram* from the *Choose from* list
3. Click and drag the first picture into the chart preview area
4. Click and drag the variable *Pulse2* onto the X-axis
5. In the Element Properties window on the right select *Display normal curve*
6. Click on *Apply*
7. Click on *OK*

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield



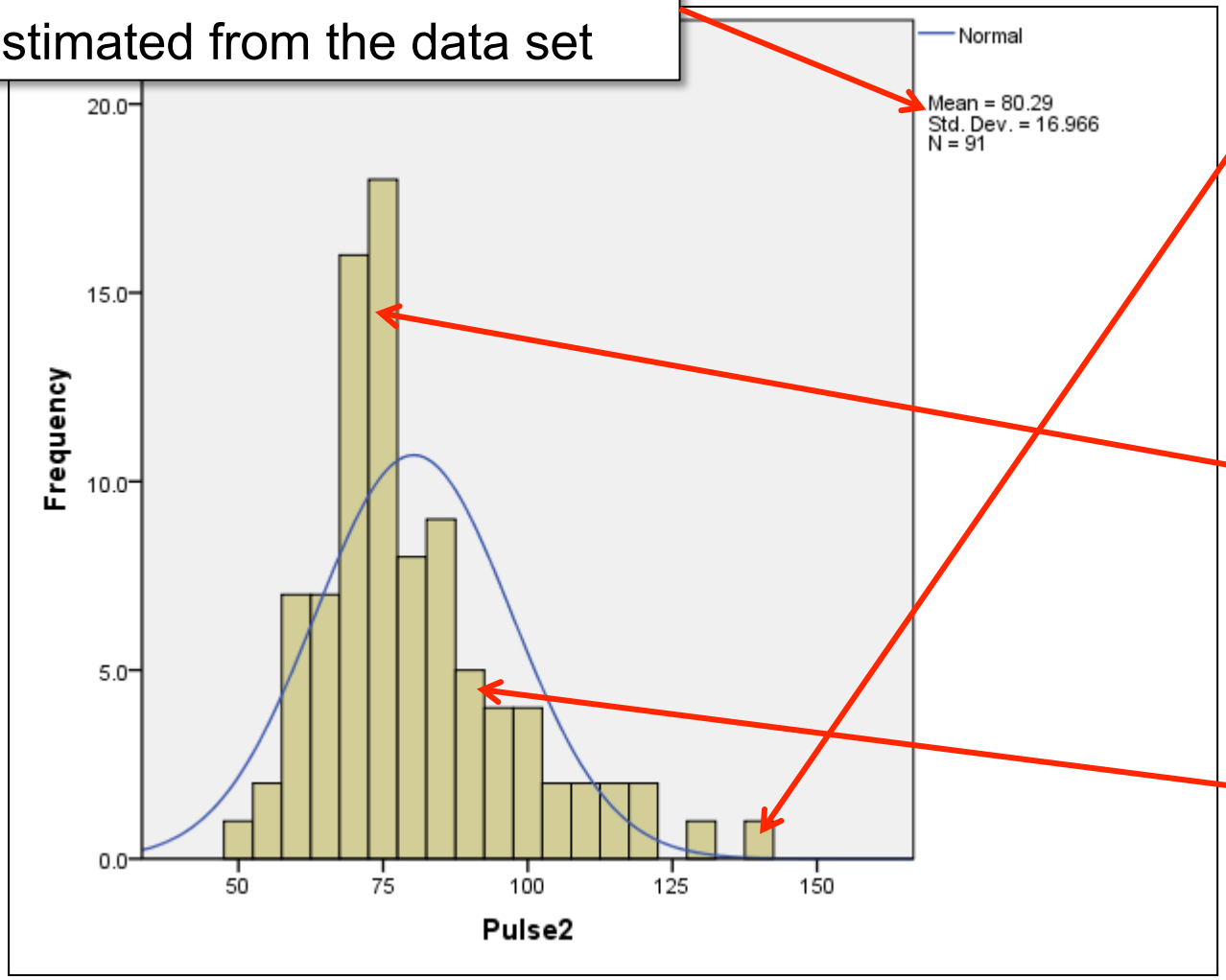
www.**statstutor**.ac.uk

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

This should create a histogram that looks like this:

Mean and standard deviation estimated from the data set



Frequencies calculated over intervals

Frequencies below the estimated mean are clearly higher than the normal curve whilst those for intervals above the mean are lower

This is known as a **positive skew**

www.**stats**tutor.ac.uk

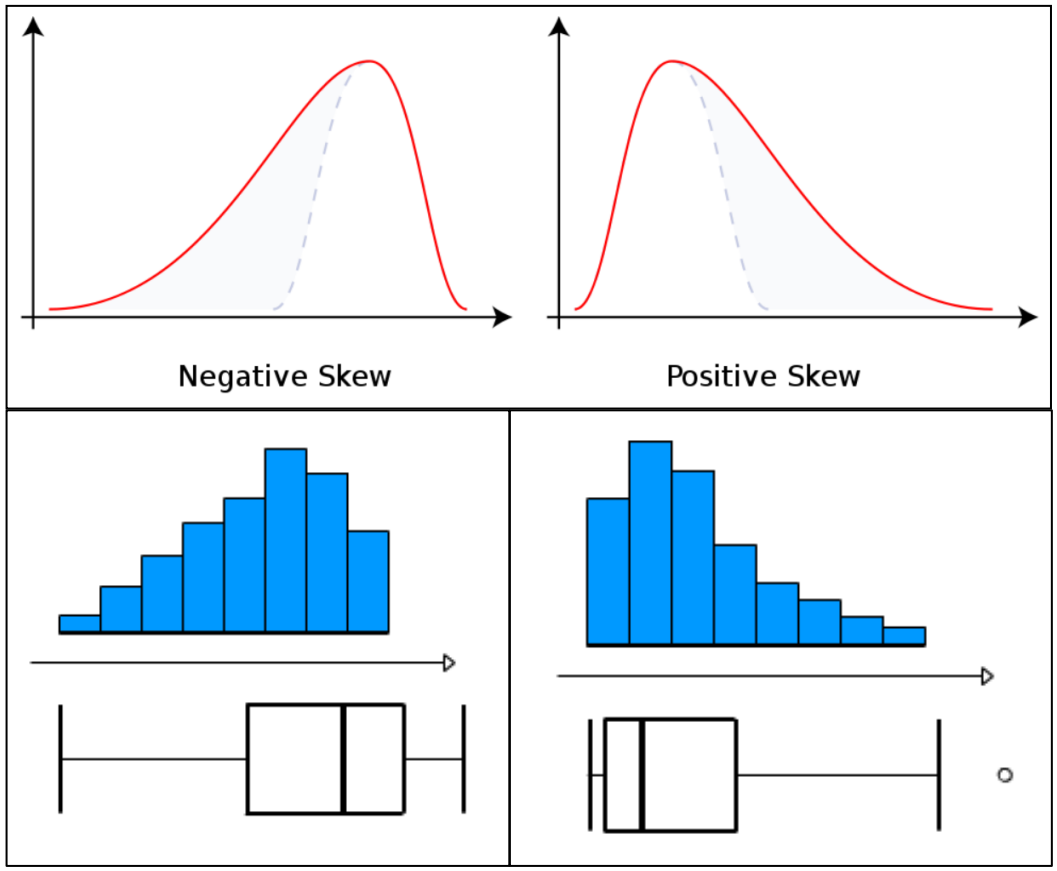
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Skewness

Skewness is a measure of **asymmetry**:

- The normal distribution has zero skewness



www.**stats**tutor.ac.uk

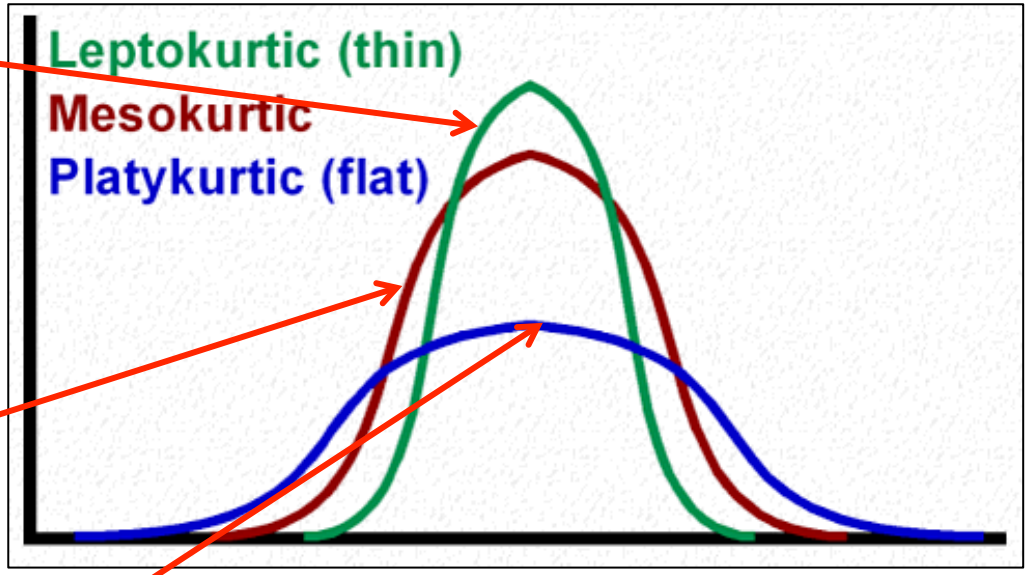
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Kurtosis

Kurtosis is a measure of **peakness**.

- Positive kurtosis data has a smaller standard deviation than a normal distribution
- Zero kurtosis (normal distribution)
- Negative kurtosis data has a larger standard deviation than a normal distribution



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Exercise

- Create boxplots with normal curves for the other scale variables:
  - *Pulse1*
  - *Height*
  - *Weight*
- Comment on the comparison of each histogram with the normal curve in terms of its skewness and kurtosis

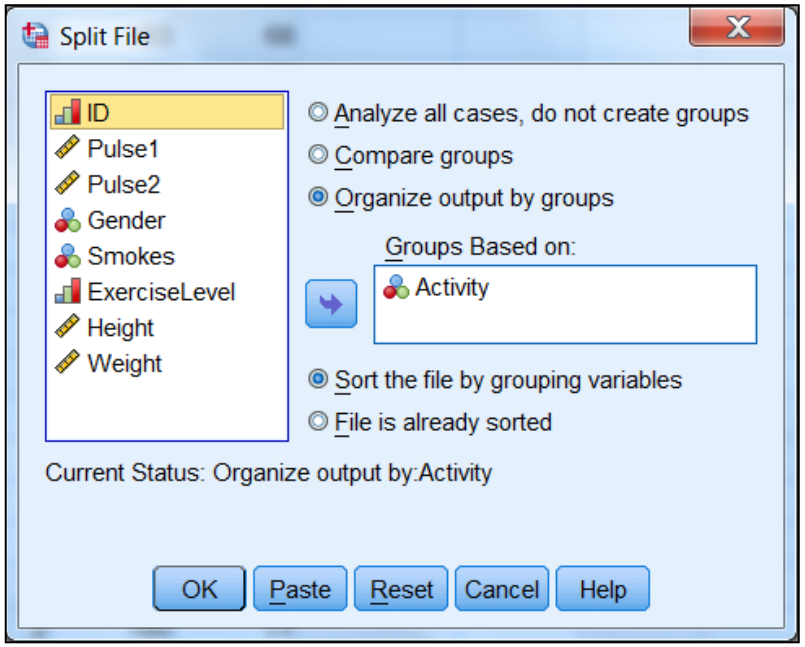
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Comparing several groups

To compare several groups as we did with the boxplots, e.g. by *Activity*, we first need to split the file:

- Select *Data – Split File…*
- Select *Organise output by groups* and select the variable *Activity*
- Repeat the *Chart Builder* as before
- **Alternatively**, you can either use the Panel by option with Graphs – Legacy Dialogs – Histogram **or** the Factors option with Explore – Analyze – Descriptive Statistics, select Histogram under Plots… and then add a normal curve using the Chart Editor

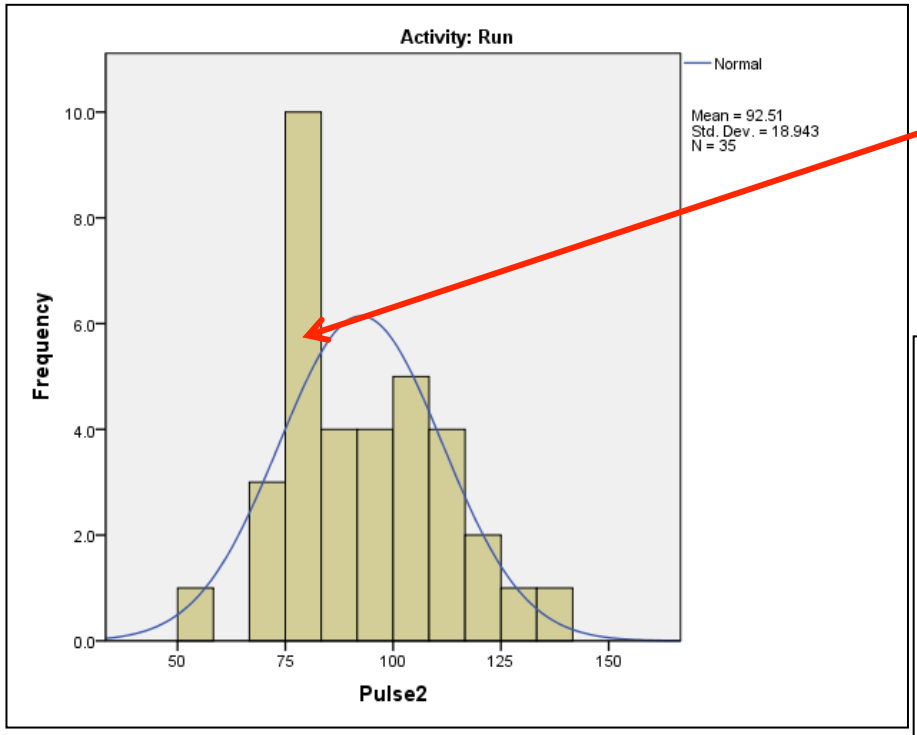


www.statstutor.ac.uk

Peter Samuels
Birmingham City University

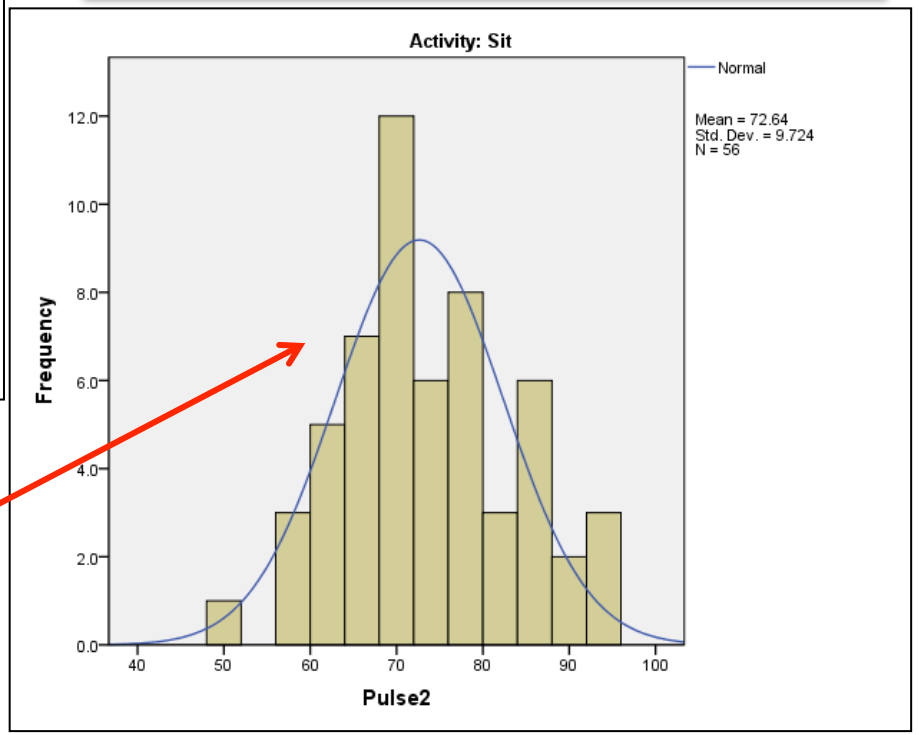
Reviewer: Ellen Marshall
University of Sheffield

This should create two graphs which look like this:



Run distribution clearly positively skewed – as also seen on boxplot

Also has positive kurtosis



Skewness and kurtosis of *Sit* data not so obvious

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Exercise

- Split the *Pulse2* variable by the other 3 categorical variables:
  - *Gender*
  - *Smokes*
  - *ActivityLevel*
- Comment on the comparison of each histogram with the normal curve in terms of skewness and kurtosis

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Closing SPSS

- Re-save your data file Pulse.sav
- Save your output file as Pulse.spv
- Close the output file
- Close the data file – this closes SPSS
- Copy Pulse.sav and Pulse.spv onto a memory stick
- Also close the Excel data file and copy it onto a memory stick

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Recap

- What is SPSS?
- Opening and closing SPSS
- The SPSS windows
- Uploading a data set from Excel into SPSS
- Creating boxplots
- Creating histograms with fitted normal curves:
  - Skewness and kurtosis
  - Commenting on the normality of the distribution
  - Splitting the data sets

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield