Statistics II

Core topics in Mathematics

Lecture 24

Learning Outcomes

- Working with grouped data.
- Visualising grouped data using histograms in EXCEL.
- Fitting simple curves and trendlines using EXCEL.

Grouped Data

When there are many different measurements with few/no repetition or just a large number of data, then it is only possible to make any real sense of the data if they are grouped together in intervals.

In this case, the formulae for calculating values such as the mean are slightly different:

$$\bar{x} = \frac{\sum fx_i}{\sum f} \quad \text{(grouped data)}$$

Example 1: Grouped, continuous data

The heights (in cm) of 25 people of the same age were measured. The following table shows the data:

180.84	164.87	167.77	167.78	174.39
176.14	176.87	159.57	164.73	174.51
168.47	180.64	170.04	162.71	174.02
171.91	169.31	171.68	152.49	177.58
172.03	169.68	161.87	165.48	181.90

Summarise the data into a frequency table and find the mean.

Example 1: Grouped, continuous data

Group	Tally	Freq. f	Midpoint x_i	fx_i	C. f
150-155	I	1	152.5	152.5	1
155-160	I	1	157.5	157.5	2
160-165	IIII	4	162.5	650	6
165-170	𝍸 I	6	167.5	1005	12
170-175	𝍸 II	7	172.5	1207.5	19
175-180	III	3	177.5	532.5	22
180-185	III	3	182.5	547.5	25
		$\sum f = 25$		$\sum fx_i = 4257.5$	

$$\bar{x} = \frac{\sum fx_i}{\sum f} = \frac{4252.5}{25} = 170.1$$

The median is the 13^{th} value, thus 172.5 from the CF. The modal group is 170-175.

Graphing Data

Graphs can be used to quickly determine key characteristics of the data under analysis. Some possible graphs are:

- Bar charts and pie charts (discrete or qualitative data).
- Histograms/Frequency distributions (continuous, quantitative data).
- Frequency polygons.
- Cumulative frequency curves.

Creating a Histogram in EXCEL

1. Ensure that the Analysis Toolpak is enabled. In Windows 10, go to "File>Options>Add-ins" and ensure "Analysis Toolpak" is selected. For OSX, go to "Tools > Excel Add-ins . . . "
2. Choose a bin width suitable for your data, then create a column containing the upper limit of all of the bins.
3. Select the Data tab, then "Analysis" and choose "Histogram" from the list. A pop-up box will appear.
4. For "Input range", select *all* of the raw data points.
5. For "Bin range", select the cells containing your upper limits.
6. Click on "Ouput range" and choose an area of the worksheet that will not interfere with your raw data. Excel will create a frequency table (tally chart) here.
7. Make sure "Chart output" is ticked.

Groups/Bins and Histograms

If we are given a set of ungrouped data, how many bins should we group it into for calculating statistics or when creating a histogram?

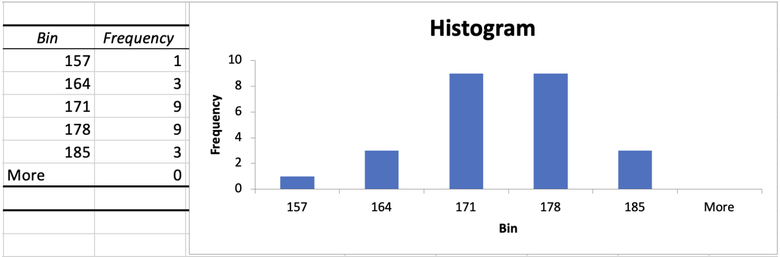
There are several rules that we can use to determine a sensible number of M bins for a set of N datapoints, such as:

- The square root rule: $M = \lceil \sqrt{N} \rceil$
- Sturge's formula: $M = \lceil \log_2(N) + 1 \rceil$

In each case, the "ceiling" function $\lceil \ \rceil$ indicates that the result should be rounded *up* to the nearest integer.

Example 2

Using the procedure outlined previously, we can produce a histogram for Example 1. There are $N = 25$ datapoints, so using the square root rule we group them in $M = 5$ bins.



Bin	*Frequency*
157	1
164	3
171	9
178	9
185	3
More	0

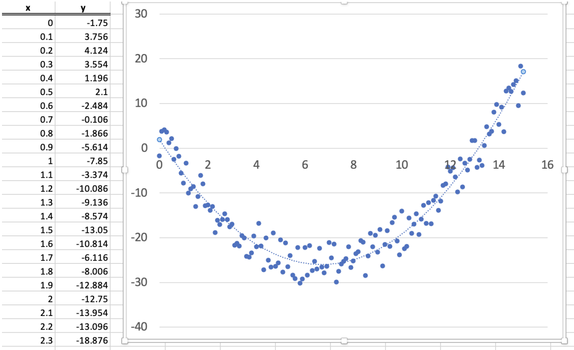
Curve fitting

Often we can obtain a set of experimental data, and hypothesise that the relationship between the independent variable (that we control) and the dependent variable (that we measure) is described by some function. If we could determine the exact relationship, we could make further predictions by extrapolating the fitted curve.

Once we have decided on a general form of the relationship between our variables (e.g. linear, quadratic, exponential, power law), curve fitting is the process of **finding the set of parameter values** that best fits the set of experimental data.

Example 3

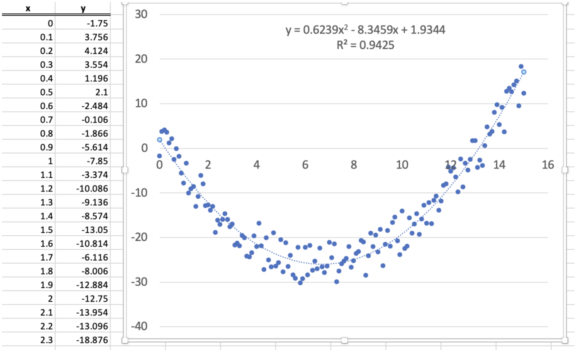
Suppose we have the set of data shown below:



x	y
0	-1.75
0.1	3.756
0.2	4.124
0.3	3.554
0.4	1.196
0.5	2.1
0.6	-2.484
0.7	-0.106
0.8	-1.866
0.9	-5.614
1	-7.85
1.1	-3.374
1.2	-10.086
1.3	-9.136
1.4	-8.574
1.5	-13.05
1.6	-10.814
1.7	-6.116
1.8	-8.006
1.9	-12.884
2	-12.75
2.1	-13.954
2.2	-13.096
2.3	-18.876

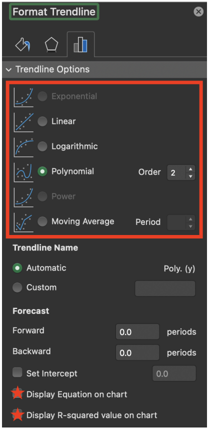
Perhaps a quadratic function $y = ax^2 + bx + c$ would fit. But what values of a, b and c would result in the *best* fit?

Example 3



x	y
0	-1.75
0.1	3.756
0.2	4.124
0.3	3.554
0.4	1.196
0.5	2.1
0.6	-2.484
0.7	-0.106
0.8	-1.866
0.9	-5.614
1	-7.85
1.1	-3.374
1.2	-10.086
1.3	-9.136
1.4	-8.574
1.5	-13.05
1.6	-10.814
1.7	-6.116
1.8	-8.006
1.9	-12.884
2	-12.75
2.1	-13.954
2.2	-13.096
2.3	-18.876

$y = 0.6239x^2 - 8.3459x + 1.9344$
$R^2 = 0.9425$



Format Trendline

Trendline Options

○ Exponential
○ Linear
○ Logarithmic
● Polynomial Order 2
○ Power
○ Moving Average Period

Trendline Name

● Automatic Poly. (y)
○ Custom

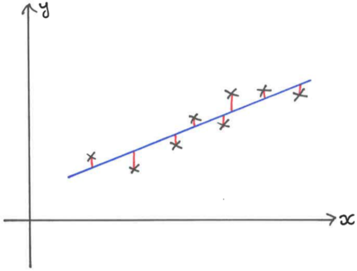
Forecast

Forward 0.0 periods
Backward 0.0 periods
□ Set Intercept 0.0
★ Display Equation on chart
★ Display R-squared value on chart

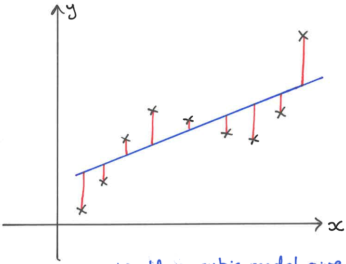
With the trendline tool we can specify fitting a 2nd-order polynomial (a quadratic function). The best such function is $y = 0.6239x^2 - 8.3459x + 1.9344$

Which model to choose?

In harder cases, we could choose several different models and fit the best parameter choices in each case.



$y = f(x) = mx + c$



Would a cubic model give a better fit?

How would we know which model described the data best by giving the closest fit?

R^2

To quantify the "goodness of fit" for each model, we can calculate the R^2 value, also called the coefficient of determination.

Calculating R^2

For a set of N data points (x_i, y_i), to which a model is fitted given by $y = f(x)$, we calculate R^2 using:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where: $SS_{res} = \sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2$ and $SS_{tot} = \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2$

Interpreting R^2

Interpreting R^2

If R^2 is equal to 1, it means that the curve fits the data perfectly.

A smaller value (nearer to zero) indicates a poorer fit.

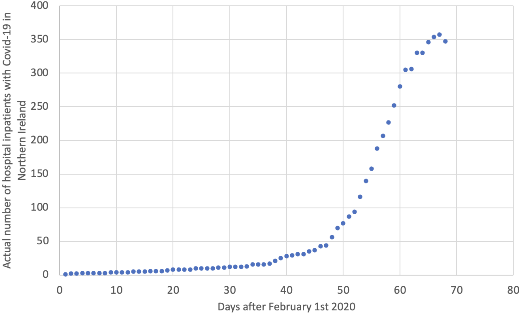
Given a data set, we can create a scatter plot, and undertake a curve-fitting procedure for each reasonable model to find the *best version of that model*. Then, compare the resulting R^2-values and determine which was the best overall best.

EXCEL's ability for curve fitting has limitations. Only certain simple functions can be fitted, and some cannot be fitted if there are zeros or negative values in the data.

Application: modelling the early spread of COVID-19

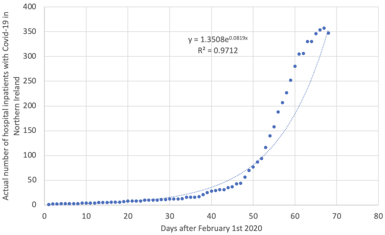
If we could use curve fitting to accurately fit a model to the data documenting the spread of coronavirus in the UK as it emerges, we may be able to estimate demand for healthcare and where and when to allocate resources.

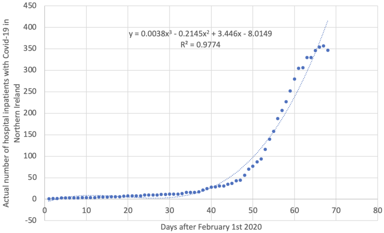
Consider this data, which shows the number of people in hospital in Northern Ireland with Covid-19 between February 1st and April 7th 2020.



Application: modelling the early spread of COVID-19

From the options available, the most reasonable (without using a very high-order polynomial) seem to be an exponential or a cubic function. They both fit the data very well ($R^2 \approx 0.97$).





Should we use these models to forecast hospitalisations in: (a) one week; (b) four months; (c) two years after the final datapoint?