ACTIVITY 2 ANSCOMBE'S QUARTET

Anscombe's quartet is a collection of four bivariate data sets which have interesting statistical properties.

It was first described in 1973 by the English statistician **Francis Anscombe** (1918 - 2001). At the time, computers were becoming increasingly popular in statistics, as they allowed for more large scale and complex computations to be done within a reasonable amount of time. However, many common statistical packages primarily performed numerical calculations rather than produce graphs. Such output was often limited to those with advanced programming skills.

Francis Anscombe
Photo courtesy of Yale University.

In his 1973 article, Anscombe stressed that:

> "*A computer should make* both *calculations* and *graphs. Both sorts of output should be studied; each will contribute to understanding.*"

The data values for Anscombe's quartet are given in the tables below:

Data set A:

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Data set B:

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.1	6.13	3.1	9.13	7.26	4.74

Data set C:

x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Data set D:

x	8	8	8	8	8	8	8	19	8	8	8
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.5	5.56	7.91	6.89

Enter the data into your **graphics calculator** or click on the icon to access the data in the **statistics package**.

What to do:

1 For each data set, use technology to calculate:

a the mean of each variable **b** the population variance of each variable.

Comment on your answers.

2 Find the regression line for each data set. What do you notice?

3 Construct a scatter diagram for each data set, and plot the corresponding regression line on the same set of axes.

4 How do your calculations in **1** and **2** compare to your graphs in **3**? Is a linear model necessarily appropriate for each data set?

5 Why is it important to consider both graphs *and* descriptive statistics when analysing data?