Pearson's correlation coefficient

G Hall

`http://www.hep.ph.ic.ac.uk/~hallg/UG_2015/Pearsons.pdf`

February 17, 2015

Abstract

An explanation of Pearson's correlation coefficient is given and its suitability for evaluating curve fits to data in the third year lab is discussed.[1]

1 Introduction

The Pearson's r coefficient (or r^2), also known as the Pearson Product-Moment Correlation, is often used in modern software packages available for data display and curve fitting. What is the meaning of this variable and how should it be used in fitting physical data?

It is defined as the ratio of the covariance of two variables representing a set of numerical data, normalised to the square root of their variances, i.e.:

$$r = \frac{C_{xy}}{\sqrt{C_{xx}C_{yy}}} = \frac{C_{xy}}{\sigma_x \sigma_y} \tag{1}$$

or, in more detail, for a set of N two-dimensional data points $[x_1, x_2, \ldots, x_N]$ and $[y_1, y_2, \ldots, y_N]$, we have:

$$\bar{x} = \frac{1}{N}\sum_i x_i \qquad \bar{y} = \frac{1}{N}\sum_i y_i \quad \text{and}$$

$$C_{xy} = \frac{1}{N-1}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \tag{2}$$

$$C_{xx} = \sigma_x^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2 \tag{3}$$

$$C_{yy} = \sigma_y^2 = \frac{1}{N-1}\sum_i (y_i - \bar{y})^2 \tag{4}$$

[1] The document has been formatted using a simple Latex template in a style which might act as a guide for producing the lab reports.

It is primarily a statistical quantity, not devised to be used in physics data analysis (which does not necessarily mean it should not be), but with the availability of statistical software packages which are useful for curve fitting, it often seems to be taken to be a measure of the quality of a fit to physics data. In most cases, this is probably a dangerous, and often incorrect, assumption.

2 The properties of the Pearson coefficient

Two important properties [1] can immediately be noted:

- r depends only on the data values and spread, not on any hypothesised relationship between them
- r does not depend on errors on the measured quantities

This immediately tells us that r is not evaluating anything other than some internal relation between the data values; it is not testing a hypothesis for the origin of the data, nor is it giving more weight to some data points, for example because they are measured with greater precision, than others.

In fact, another assumption of the Pearson statistic [1] is that the relationship to be tested is a linear one. In this case the outcome is easy to derive. If

$$
\begin{aligned}
y_i &= Ax_i + B \quad \text{then} \qquad \bar{y} = A\bar{x} + B && (5)\\
C_{xy} &= \frac{1}{N-1}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N-1}\sum_i (x_i - \bar{x})(Ax_i + B - A\bar{x} - B)\\
&= \frac{A}{N-1}\sum_i (x_i - \bar{x})^2 && (6)\\
C_{xx} &= \sigma_x^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2 && (7)\\
C_{yy} &= \sigma_y^2 = \frac{A^2}{N-1}\sum_i (x_i - \bar{x})^2 && (8)\\
r &= \frac{C_{xy}}{\sqrt{C_{xx}C_{yy}}} = \frac{A}{|A|} = \pm 1 && (9)
\end{aligned}
$$

In other words, if y and x are *exactly* linearly related, $r = \pm 1$, depending on whether the slope is positive or negative (correlation or anticorrelation). More likely, with real data of any kind, there will be a spread in the values of x and y, in which case the correlation will be less than maximal, i.e. $|r| < 1$.

Fig.1 shows some examples of simulated data with random gaussian fluctuations of different magnitude applied, and the resulting r values. In this case, since the data were generated from a linear relationship, shown in Fig.1(a), lower values of r should be understood as greater fluctuations about the trend, not as absence of correlation or worse fits to the data values.

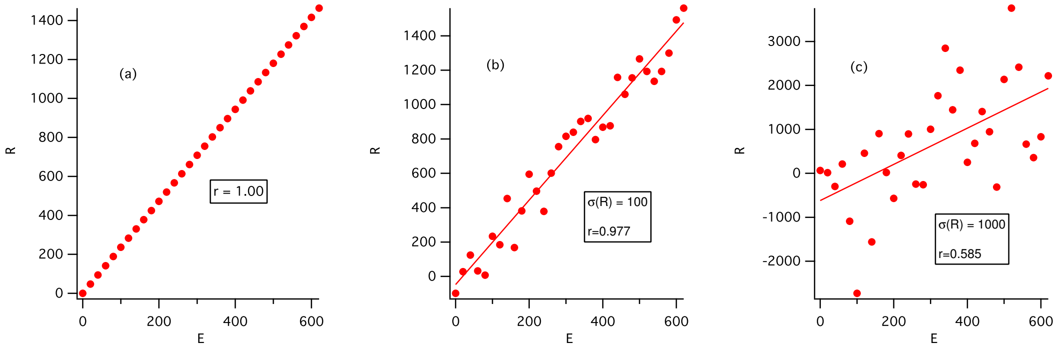

Figure 1: Simulated data following a linear relationship between R and E with random gaussian fluctuations added to the R values. (a) no fluctuations, resulting in $r = 1$. (b) $\sigma(R) = 100$ resulting in $r = 0.977$. (c) $\sigma(R) = 1000$, and $r = 0.585$. Note the change of vertical scale.

In view of the fact that r does not include experimental errors on the data points, which are often not identical for each data point, it is unlikely that the computed fit parameters or their errors based on the Pearson coefficient will be reliable, except in the fortuitous case that the spread of the data points coincides closely with the experimental fluctuations.

3 Relevance to physics data

Although software like Origin [2] produces plots and cites the r value of the data, is this the correct way to test the behaviour of experimental physics data? Certainly many hypothesised relationships can be linearised by taking logarithms, so evaluation using r is possible. However, measurements in physics are invariably associated with errors whose influence must be taken into account when establishing the fit parameters, such as A and B in equation (5), and the Pearson's coefficient does not permit this. The most common way of including experimental errors is by using the χ^2 function, described in many textbooks [3, 4, 5, 6]

$$\chi^2 = \sum_i \frac{(y_i - f(x_i))^2}{\sigma_i^2} \tag{10}$$

where $f(x_i)$ is the predicted value of y_i based on the chosen parameterisation and σ_i is the (assumed gaussian) error[2]on the measured value y_i. In this case a fit to the data would be considered good if the χ^2 per degree of freedom N_{DF} has a value $\chi^2/N_{DF} \sim 1$.[3] If $\chi^2/N_{DF} >> 1$ then most likely either the chosen function does not describe the data well or the errors are significantly underestimated. Conversely if $\chi^2/N_{DF} << 1$ then either the errors are greatly overestimated or the data are suspect, e.g. faked.

[2] A more sophisticated version of the formula is applicable if there are correlations between the data points, which is sometimes the case.

[3] Normally N_{DF} = number of data points - number of free parameters = $N - 2$ for a linear parameterisation.

Origin does permit the use of the χ^2 function and this is recommended in fitting data. If it is, and the $\chi^2/N_{DF} \sim 1$ then the errors reported by the program for the fit parameters are likely to be realistic.

4 Conclusions

Pearson's coefficient may be a useful statistical tool but it should not generally be used in evaluating the quality of fits to physics data. Instead the χ^2 function is available in the Origin software and is a better choice.

References

[1] For other properties and assumptions, see for example https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

[2] http://www.originlab.com/index.aspx?go=Support/DocumentationAndHelpCenter

[3] L. Lyons. *A Practical Guide to Data Analysis for Physical Science Students.* Cambridge University Press (1998)

[4] L. Lyons. *Statistics for Nuclear and Particle Physicists.* Cambridge University Press (1986)

[5] F. James. *Statistical Methods In Experimental Physics.* World Scientific (2006)

[6] G. Cowan. *Statistical Data Analysis.* Oxford University Press (1998)