STATISTICS NOTES

Correlation in restricted ranges of data

J Martin Bland,[1] Douglas G Altman[2]

[1]Department of Health Sciences, University of York, York YO10 5DD
[2]Centre for Statistics in Medicine, University of Oxford, Oxford OX2 6UD

Correspondence to: Professor M Bland
martin.bland@york.ac.uk

Cite this as: *BMJ* 2011;342:d556
doi: 10.1136/bmj.d556

Competing interests: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

In a study of 150 adult diabetic patients there was a strong correlation between abdominal circumference and body mass index (BMI) (r = 0.85).[1] The authors went on to report that the correlation differed in different BMI categories as shown in the table.

Correlation between abdominal circumference and body mass index (BMI) in 1450 adult patients with diabetes

BMI group	r
<25	0.62
25 to 30	0.50
30 to 35	−0.09
>35	0.86
All patients	0.85

The authors' interpretation of these data was that in patients with low or high BMI values (BMI <25 kg/m^2 and BMI >35 kg/m^2) the correlation was strong, but in those with BMI values between 25 and 35 kg/m^2 the correlation was weak or missing. They concluded that measuring abdominal circumference is of particular importance in subjects with the most frequent BMI category (25 to 35 kg/m^2).

When we restrict the range of one of the variables, a correlation coefficient will be reduced. For example, fig 1 shows some BMI and abdominal circumference measurements from a different population. Although these people are from a rather thinner population, the correlation coefficient is very similar, $r = 0.82$ (P<0.0001). When we divide the sample into the same four restricted ranges of BMI at 20, 25, and 30 kg/m^2, the correlation coefficient in each interval is smaller than the correlation coefficient for the whole sample. This phenomenon is to be expected; it is a result of restricting the range of data, not any particular property of BMI and abdominal circumference.

One interpretation of the correlation coefficient r is that r^2 is the proportion of the variation in abdominal circumference explained or predicted by the variation in BMI. If we restrict the range of BMI values we reduce the variation in BMI, which will explain less variation in abdominal circumference, and r will fall. If we further reduce the variation in BMI until all remaining patients have the same BMI, then we cannot explain any variation in abdominal circumference and the correlation must be zero. (By contrast within any of the sections of fig 1 the fitted regression line would be the same, apart from random variation.)

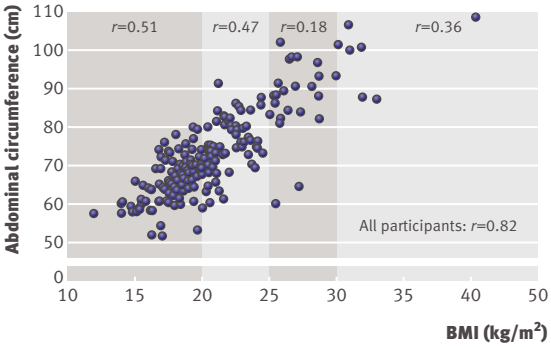

Fig 1 | BMI and abdominal circumference in 202 men and women, with correlation coefficients in four restricted ranges and overall

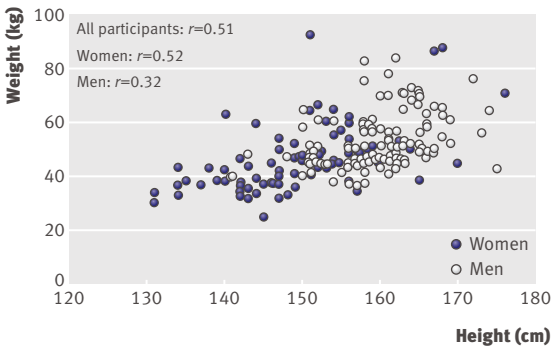

Fig 2 | Weight and height in 202 men and women, with correlation coefficients

For another example, fig 2 shows the weights and heights of the same sample, with different symbols for men and women. Clearly, the lower end of the height range for men is higher than the lower end of the range for women, but the upper ends of the ranges are very similar. The men's heights (SD 6.0 cm) are less variable than those of the women (SD 8.9 cm) or the heights of both sexes combined (also SD 8.9 cm). The correlation coefficients for women and for both men and women are very similar and considerably larger than that for men alone.

The same phenomenon can arise when the sample is restricted using another variable related to the ones being studies. For example, the correlation between weight and height of schoolchildren will increase as the age range is increased. But a spurious correlation may also be seen in such a situation, for example between shoe size and spelling ability.[2] Such an example illustrates the well worn phrase that an observed association does not imply causation.

Correlation coefficients are a property of the variables and also the population in which they are measured. If we look at a restricted population, we should not conclude that there is little or no relation between the variables because the correlation coefficient is small. But given a clear relation in the whole group, we see no point in looking within categories of one of the variables. In any case, regression is generally the preferred approach to considering the relation between two continuous variables.

The data are taken from a student elective project by Dr Malcolm Savage.
Contributors: JMB and DGA jointly wrote and agreed the text, JMB did the statistical analysis.

1 Nádas J, Putz Z, Kolev G, Nagy S, Jermendy G. Intraobserver and interobserver variability of measuring waist circumference. *Med Sci Monit* 2008;14:CR15-8.
2 Goodwin LD, Leech NL. Understanding correlation: factors that affect the size of r. *J Exp Educ* 2006;74:251-66.

Brackets (parentheses) in formulas

Douglas G Altman,[1] J Martin Bland[2]

[1]Centre for Statistics in Medicine, University of Oxford, Oxford OX2 6UD
[2]Department of Health Sciences, University of York, York YO10 5DD
Correspondence to: DG Altman
doug.altman@csm.ox.ac.uk

Cite this as: *BMJ* 2011;343:d570
doi: 10.1136/bmj.d570

Each year, new health sciences postgraduate students in York are given a simple maths test. Each year the majority of them fail to calculate $20 - 3 \times 5$ correctly. According to the conventional rules of arithmetic, division and multiplication are done before addition and subtraction, so $20 - 3 \times 5 = 20 - 15 = 5$. Many students work from left to right and calculate $20 - 3 \times 5$ as $17 \times 5 = 85$. If that was what was actually meant, we would need to use brackets: $(20 - 3) \times 5 = 17 \times 5 = 85$. Brackets tell us that the enclosed part must be evaluated first. That convention is part of various mnemonic acronyms that indicate the order of operations, such as BODMAS (Brackets, Of (that is, power of), Divide, Multiply, Add, Subtract) and PEMDAS (Parentheses, Exponentiation, Multiplication, Division, Addition, Subtraction).[1]

Schoolchildren learn the basic rules about how to construct and interpret mathematical formulas.[1] The conventions exist to ensure that there is absolutely no ambiguity, as mathematics (unlike prose) has no redundancy, so any mistake may have serious consequences. Our experience is that mistakes are quite common when formulas are presented in medical journal articles. A particular concern is that brackets are often omitted or misused. The following examples are typical, and we mean nothing personal by choosing them.

Example 1

In a discussion of methods for analysing diagnostic test accuracy, Collinson [2] wrote:

$$\text{Sensitivity} = \text{TP/TP} + \text{FN}$$

where TP = true positive and FN = false negative. The formula should, of course, be:

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}).$$

Example 2

For a non-statistical example, Leyland [3] wrote that the total optical power of the cornea is:

$$P = P_1 + P_2 - t/n_2(P_1P_2)$$

where $P_1 = n_2 - n_1/r_1$ and $P_2 = n_3 - n_2/r_2$. Here n_1, n_2, and n_3 are refractive indices, r_1, r_2, and t are distances in metres, and P, P_1, and P_2, are powers in dioptres. But he should have written $P_1 = (n_2 - n_1)/r_1$ and $P_2 = (n_3 - n_2)/r_2$. $P_1 = n_2 - n_1/r_1$ is clearly wrong dimensionally, as P_1 is dioptres, 1/metre, n_2 and n_1 are ratios and so pure numbers, and r_1 is in metres. Also, it is not clear whether $t/n_2\,(P_1P_2)$ means $(t/n_2)\,P_1P_2$, which it does, or $t/(n_2P_1P_2)$.

Do such errors matter? Certainly. In our experience the calculations are usually correct in the paper, but anyone using the published formula would go wrong. Sometimes, however, the incorrect formula was used, as in the following case.

Example 3

In their otherwise exemplary evaluation of the chronic ankle instability scale, Eechaute et al[4] made a mistake in their formula for the minimal detectable change (MDC) or repeatability coefficient,[5] writing: $\text{MDC} = 2.04 \times \sqrt{(2 \times \text{SEM})}$. Here SEM is the standard error of a measurement or within subject standard deviation.[5] This formula uses 2.04 where 2 or 1.96 is more usual,[6] but, much more seriously, the SEM should not be included within the square root, as the brackets indicate. This might be dismissed as a simple typographical error, but the authors actually used this incorrect formula. Their value of SEM was 2.7 points, so they calculated the minimal detectable change as $2.04 \times \sqrt{(2 \times 2.7)} = 4.7$. They should have calculated $2.04 \times \sqrt{(2)} \times 2.7 = 7.8$. Their erroneous formula makes the scale appear considerably more reliable than it actually is.[6] The formula is also wrong in terms of dimensions, because the minimum clinical difference should be in the same units as the measurement, not in square root units.

Some mistakes in formulas may be present in a submitted manuscript, but others might be introduced in the publication process. For example, problems sometimes arise when a displayed formula is converted to an "in-text" formula as part of the editing, and the implications are not realised or not noticed by either editing staff or authors. Often it is necessary to insert brackets when reformatting a formula. So the simple formula:

$$\frac{p}{1-p}$$

should be changed to $p/(1 - p)$ if moved to the text.

Formulas in published articles may be used by others, so mistakes may lead to substantive errors in research. It is essential that authors and editors check all formulas carefully.

Acknowledgements: We are very grateful to Phil McShane for pointing out a mistake in an earlier version of this statistics note.

Contributors: DGA and JMB jointly wrote and agreed the text.

Competing interests: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

1 Wikipedia. Order of operations. [cited 2010 Nov 23]. http://en.wikipedia.org/wiki/Order_of_operations.
2 Collinson P. Of bombers, radiologists, and cardiologists: time to ROC. *Heart* 1998;80:215-7.
3 Leyland M. Validation of Orbscan II posterior corneal curvature measurement for intraocular lens power calculation. *Eye* 2004;18:357-60.
4 Eechaute C, Vaes P, Duquet W. The chronic ankle instability scale: clinimetric properties of a multidimensional, patient-assessed instrument. *Phys Ther Sport* 2008;9:57-66.
5 Bland JM, Altman DG. Statistics notes. Measurement error. *BMJ* 1996;312:744.
6 Bland JM. Minimal detectable change. *Phys Ther Sport* 2009;10:39.

bmj.com

Previous articles in this series

- Choosing target conditions for test accuracy studies that are relevant to clinical practice (*BMJ* 2011;343:d4684)
- Brackets (parentheses) in formulas (*BMJ* 2011;343:d570)
- How to obtain the confidence interval from a P value (*BMJ* 2011;343:d2090)
- How to obtain the P value from a confidence interval (*BMJ* 2011;343:d2304)
- Verification problems in diagnostic accuracy studies: consequences and solution (*BMJ* 2011;343:d4770)