Becoming confident about confidence intervals

Perry DC[(1)], Griffin XL[(2)], Dritsaki M[(2)], Costa ML[(2)] and Parsons NR[(3)]

Correspondence to D. Perry; email: danperry@liverpool.ac.uk

A clear understanding of variability is at the heart of many of the studies published in *The Bone & Joint Journal*. The terms variability, dispersion, spread, distribution and variance, amongst others, are used to refer to the range of values within a dataset. Statistics was founded as the study of variability, with variability often regarded as a nuisance.[1] Measures of variability can be difficult concepts to understand when reading papers. While the simplest representation of the variability of data is to show the raw data, this is only feasible for small datasets, and is often unethical if data refer to individual patients. Measures of variability vary, from the relatively simple, to the more complex.

Range

The simplest measure of variability is the range; the highest and lowest value for a variable. Table I and Figure 1 show raw data and the range for a fictional dataset, and graphically illustrate this using strip plots. The range is important but often fails to provide a complete description of variability if there are outliers (i.e. data points which are distant from other observations or lie outside expected or typical values).

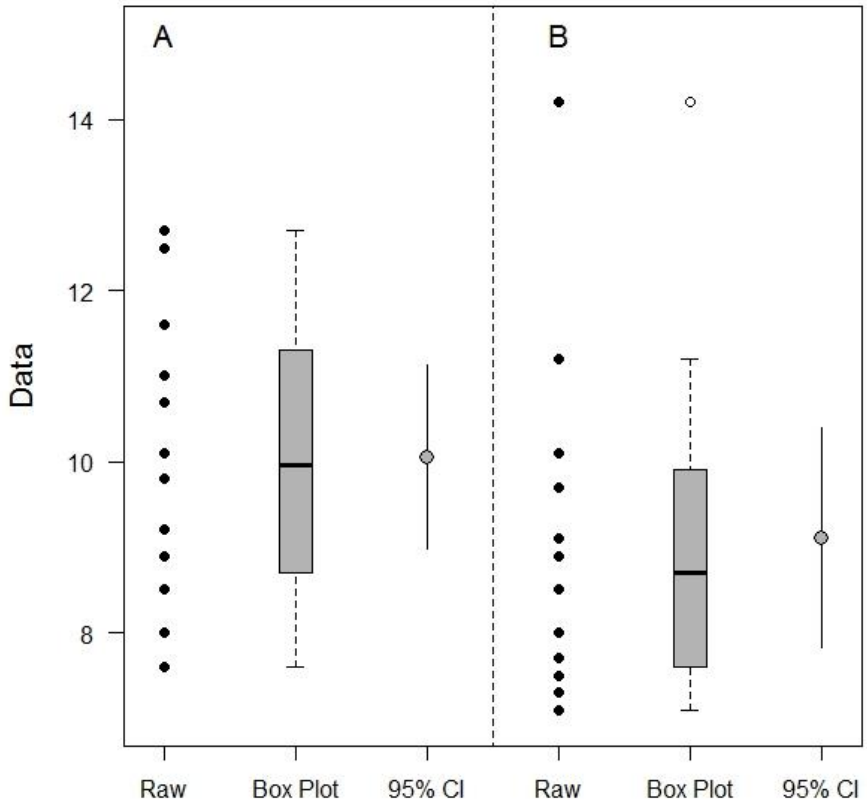
Table I. Table showing fictional datasets A & B

	Dataset A	**Dataset B**
Raw Data	7.6, 8.0, 8.5, 8.9, 9.2, 9.8, 10.1, 10.7, 11.0, 11.6, 12.5, 12.7	7.1, 7.3, 7.5, 7.7, 8.0, 8.5, 8.9, 9.1, 9.7, 10.1, 11.2, 14.2
Range	7.6 to 12.7	7.1 to 14.2
Interquartile range	8.7 to 11.3	7.6 to 9.9
Mean (standard deviation)	10.1 (1.7)	9.1 (2.0)
Mean (95% confidence intervals)	10.1 (9.0 to 11.1)	8.6 (7.8 to 10.4)

Quantiles

A more robust method of describing variability is to use quantile measures or percentiles. This methodology makes no assumptions about the distribution of the data, which more complex methods often do. The data are ordered from lowest to highest, and divided such that they are in equally sized groups. Data are often summarised using quartiles (dividing into four), which provides an interquartile range (IQR);[2,3] other divisions such as deciles (dividing into ten) are also widely reported.[4-6] A graphic summary, which is often used is the box and whisker plot (Fig. 1), which uses the IQR to form the box, with the median (the middle value of the data) as a central bar and whiskers out to 1.5 times the IQR. Values outside these whiskers are considered outliers.

Figure 1. Strip plots showing raw data, box plots and means with 95% confidence intervals (95% CI) for data samples A (a) and B (b).



Standard deviation

For data where the variability is approximately symmetrical about the mean, thus with no skew, a more informative and powerful summary measure than the IQR is the standard deviation (SD). A fundamental property of the normal distribution is that 95% of the data lie within approximately two SDs of the mean. The SD is determined from the position of each data point relative to the mean. Although it is associated with a normal distribution,[7] it is a perfectly valid measure of variability, regardless of the true distribution of the data.[8]

Confidence intervals

A confidence interval allows an estimate of certainty around the parameter of interest, often the mean. If we were to repeat a study or trial 100 times, there would undoubtedly be variation in the results reported (i.e. variations in the effect size of an intervention) creating uncertainty. We therefore need to be aware of the degree of certainty for the true value of the effect. A 95% confidence interval offers the range of values for which there is 95% certainty that the true value of the parameter lies within the confidence limits, i.e. the results of dataset

A suggests that the true value of the mean could be as low as 9.0 and as large as 11.1 but our best estimate is 10.1.

Confidence intervals are based on the standard error of the mean (SEM). This is calculated from the SD and sample size (n), and is another estimate of variability.[8] A 95% confidence interval is identified by calculating the values two (or more precisely 1.96) standard errors either side of the mean; just as it is well known that 95% of all data in the normal distribution lies within two standard deviations of the mean.

Applying confidence intervals and the relationship to p-values

The Distal Radius Acute Fracture Fixation Trial (DRAFFT) concerning the fixation of dorsally displaced distal radial fractures has been one of the most influential trials in orthopaedic surgery,[9-11] and has met with both controversy and nationwide praise and adoption.[12,13] The primary outcome in this trial was the 'Patient Rated Wrist Evaluation' (PRWE) at 12 months with the final result being a −1.3 point difference (95% CI −4.5 to 1.8) between the two groups, whereby zero was no effect. The best estimate of the treatment effect in the general population was hence -1.3, though 95% confidence intervals tell us that we are 95% certain the true effect would be as low as -4.5 and as large as 1.8. The confidence interval included the point of no effect (i.e. zero), which indicated that the result did not reach statistical significance; confirmed through significance testing (p = 0.40). Confidence intervals are generally more useful than a p-value alone, as they offer greater consideration of the size of the effect with particular consideration to the minimally clinically important difference (MCID). The MCID for PRWE is six points, reflecting the difference between turning a doorknob with mild pain, *versus* no pain. Even the most extreme values predicted by the DRAFFT study in the general population (−4.5 to 1.8) are less than the six-points considered necessary to be clinically significant. Thus, confidence intervals show us that even if the most extreme value of -4.5 points were the true effect in the general population, this would still be lower than the MCID and not be of clinical significance.

Variability is fundamental to the interpretation of data in clinical research. The presentation of variability must be considered when submitting to the journal. The p-value is a useful summary measure that is much loved by both orthopaedic surgeon and FRCS examiner, although in isolation it can often be uninformative and even misleading. These shortfalls can be overcome by using measures of spread such as confidence intervals. Research studies are much easier to understand for both reader and reviewer if good reporting principles are followed and adequate measures of variability are routinely presented.

References:

1. Fisher RA. Statistical methods for research workers. 13th ed. Edinburgh: Oliver and Boyd, 1958.
2. Metcalfe AJ, Clark DA, Kemp MA, Eldridge JD. Trochleoplasty with a flexible osteochondral flap results from an 11-year series of 214 cases. Bone Joint J 2017;99-B:344–350.
3. Handoll HH, Keding A, Corbacho B, et al. Five-year follow-up results of the PROFHER trial comparing operative and non-operative treatment of adults with a displaced fracture of the proximal Humerus. Bone Joint J 2017;99-B:383–392.
4. Middleton SD, McNiven N, Griffin EJ, Anakwe RE, Oliver CW. Long-term patient-reported outcomes following Bennett's fractures. Bone Joint J 2015;97-B:1004–1006.
5. Todd NV, Skinner D, Wilson-MacDonald J. Secondary neurological deterioration in traumatic spinal injury. Bone Joint J 2015;97-B:527–531.
6. Altman DG, Bland JM. Quartiles, quintiles, centiles and other quantiles. BMJ 1994;309:996.
7. Altman DG, Bland JM. Statistics notes: the normal distribution. BMJ 1995;310:298.
8. Altman DG, Bland JM. Standard deviations and standard errors. BMJ 2005;331:903.

9. Costa ML, Achten J, Parsons NR, et al. Percutaneous fixation with Kirschner wires versus volar locking plate fixation in adults with dorsally displaced fracture of distal radius: randomised controlled trial. BMJ 2014;349:4807.
10. Tubeuf S, Yu G, Achten J, et al. Cost effectiveness of treatment with percutaneous Kirschner wires versus volar locking plate for adult patients with a dorsally displaced fracture of the distal radius: analysis from the DRAFFT trial. Bone Joint J 2015;97-B:1082–1089.
11. Hobby JL. Cost effectiveness of percutaneous fixation with Kirschner wires versus locking plate for adult patients with a dorsally displaced fracture of the distal radius from the DRAFFT trial: a commentary. Bone Joint J 2015;97-B:1015–1016.
12. Fullilove S, Gozzard C. Dorsally displaced fractures of the distal radius: a critical appraisal of the DRAFFT (distal radius acute fracture fixation trial) study. Bone Joint J 2016;98-B:298–300.
13. Costa ML, Jameson SS, Reed MR. Do large pragmatic randomised trials change clinical practice?: assessing the impact of the Distal Radius Acute Fracture Fixation Trial (DRAFFT). Bone Joint J 2016;98-B:410–413.

Affiliations:
(1) University of Liverpool,
Institute of Translational
Medicine, Easton Rd, Liverpool
L12 2AP, UK.
(2) University of Oxford,
NDORMS, Oxford Trauma,
Kadoorie Centre, John Radcliffe
Hospital, Oxford, OX3 7UD, UK.
(3) University of Warwick,
Statistics and Epidemiology,
Warwick Medical School,
Coventry CV4 7AL, UK.