

STATISTICAL GUIDELINES

These guidelines are designed to help authors prepare statistical data for publication and are not a substitute for the detailed guidance required to design a study or perform a statistical analysis. Each section of a scientific paper is addressed separately.

Summary

The number and source of data must be stated and conclusions which have a statistical basis must be substantiated by inclusion of pertinent descriptive statistics [mean or median, standard deviation (SD) or interquartile range, percentage coefficient of variation (%CV), 95% confidence limits, regression equations, etc.].

Methods

Experimental design, subject selection and randomization procedures should be described and analytical precision quoted when appropriate. The hypotheses to be tested by a statistical procedure must be stated and where appropriate power calculations for the sample size used should be given (it is recommended that the power is <80%). In case-control studies, clearly define how cases and controls were selected and what matching has taken place.

Statistical tests should be described but need not be referenced unless they are unusual or are applied in a non-standard way. Computer software used should be referenced.

If the paper is reporting the results of a diagnostic trial read the STARD statement (1) and for a clinical trial read the CONSORT statement (2) to improve the quality of your report.

Results

Unnecessary precision, particularly in tables, should be avoided. Rounded figures are easier to compare and extra decimal places are rarely important. Descriptive statistics require an additional digit to those used for the raw data. Percentages should not be expressed to more than one decimal place and not be used at all for small samples.

Normally distributed data should be described using a mean, SD and/or %CV and expressed as »mean (SD)« not »mean \pm SD«. When data are not normally distributed, following demonstration by tests such as the Shapiro-Wilk test (3), then medians and interquartile ranges should be used in place of mean and SD. Skewed data can often be normalized by logarithmic transformation or a power transformation. The statistical analysis and calculation of summary statistics should be carried out on the transformed data and the summary statistics transformed back to the original scale for presentation. If a logarithmic scale is used, then graphs should display non-transformed data on a logarithmic scale.

Graphs showing data of comparable magnitude should be of similar size and design. All individual points should be displayed where possible by displacing overlapping points. Error bars showing the standard error of the mean (SEM) or interquartile range, as appropriate, can be used to aid the interpretation of data.

The results of significance tests such as Student's and chi-squared should be presented with descriptive statistics, degrees of freedom (if appropriate) and probability *P*. The validity of any assumptions should be checked (e.g. conventional *t*-tests assume a normal distribution and equal variance for each set of data). For 2×2 contingency table analysis by the chi-squared test the continuity correction must be applied, and for small expected frequencies Fisher's Exact Test used.

P values should be reported in full in 1 or 2 significant figures. Describing *P* values as > 0.05 or NS (not significant) should be avoided. If the results are highly significant and the calculated *P* value from the computer is e.g. 0.000, then the use of $P < 0.0005$ is acceptable. Confidence intervals should be stated, particularly for non-significant results.

The conventional use of statistical significance is $P \leq 0.005$. If a different significance level needs to be used, then the reasons for this must be clearly stated in the statistical method section.

Discussion

Statistical significance should not be equated to importance and *P* values should not be compared between different statistical tests. Association should not be interpreted as causation without additional evidence.

Problem Areas

Multiple comparisons can produce spurious and misleading significance values. The primary hypothesis should always be clearly stated, and associations detected by retrospective analysis should be interpreted with caution. Whenever possible a single overall statistical test should be applied first e.g. ANOVA. If this is not significant, then multiple comparisons must not be applied. If it is significant then some form of multiple range test can be applied. If a single overall test is not possible, then multiple comparisons must use a Bonferroni type significance level.

With paired data the differences between individual pairs of data and the variability of the differences are more important than the individual values. Graphical representation should also show the difference between individual pairs, e.g. by plotted lines joining the paired data points.

Standard regression analysis requires data points to be independent (repeated measurements are not independent). The independent variable should be measurements without significant error, e.g. age or time, and the points should be evenly distributed over the range and have no outliers (this can be easily examined with a scat-

ter plot). These requirements are rarely satisfied with biological data.

Method comparison using regression and correlation coefficients is inappropriate and should be performed using Altman and Bland difference plots (4). If a standard scatter plot and regression line are thought to be useful they can be given along with the Altman – Bland plot. Remember, if two methods are supposed to be measuring the same thing, then it is extremely likely they will be correlated so that a statistical tool correlation not tell you anything new.

If you are carrying out complicated statistical analyses, e.g. multivariate analysis, ROC analysis etc., then it is recommended that you seek advice from a statistician.

References

1. Bossuyt PM, Reitsma JB, Bruns DE, et al., for the STARD Group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Clin Biochem* 2003; 40: 357–63.
2. Moher D, Schultz KF, Altman DG, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomization trials. *Lancet* 2001; 357: 1191–4.
3. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall, 1991: 132–12.
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1 (8476): 307–10.