Thorax 1991;46:391–393

Statistics in respiratory medicine · 1

Ranges, confidence intervals, and related quantities: what they are and when to use them

Susan Chinn

Statistics is now taught to all medical students in the United Kingdom, but at an elementary level, usually in the preclinical course. There is a gap therefore, not only in time between when the subject is learnt and when it is needed but also between the content of what can be taught and what is subsequently needed. This is inevitable, as requirements beyond the basic course of medical practitioners in different specialties vary. This series of three articles seeks to complement other publications by concentrating on concepts that seem to give particular difficulty to contributors to *Thorax*, with illustrations taken from respiratory medicine research.

This first article deals with ranges and confidence intervals, confusion between which is by no means confined to respiratory physicians. The second will look at repeatability and method comparison, including that of measurements made on different scales, an analysis that causes particular difficulty. The third article will cover choice of scale of measurement, transformations, and related issues. Although these last are the first considerations at the start of data analysis, they are also the ones that the reader may find the most difficult of the issues addressed in these articles. They therefore take last place in the hope that their relevance will by then be apparent.

Reference ranges

No measurement in clinical medicine is of use in the assessment of a patient unless the physician knows what its level should be, and how much it might deviate from the "norm" before it should be considered "abnormal." A healthy man aged 30, height 1·85 metres, on average has an FEV_1 of about 4·6 litres,[1] but we would not be surprised if such a man had an FEV_1 of 4·0 litres, nor would we be concerned unless his FEV_1 was less than 3·6 litres. We commonly express this variation in terms of a *95% range*—that is, an interval in which the measurement lies for 95% of people whom we consider to be healthy. We derive such an interval by measuring a large number of healthy subjects, and then either determine the values that cut off 2·5% of subjects at either end of the distribution or make an assumption that the distribution has a particular shape. If it has a normal (Gaussian) distribution the 95% range is from (mean $-1{\cdot}96$ standard deviations) to (mean $+1{\cdot}96$ standard deviations) provided that a large sample has been studied.

The factor 1·96 should be replaced[2] by $t_{n-1,0{\cdot}05}\sqrt{(n+1)/n}$ if the sample size is less than 100, but large samples should be used in this context. For men aged 30, height 1·85 metres, the mean FEV_1 is 4·6 litres and the standard deviation is about 0·51 litres,[1] so the 95% range is from 3·6 to 5·6 litres. Thus the *standard deviation* is a step towards deriving more useful quantities (now called *reference* ranges to avoid the confusing term "normal" range), but it can be used also as a measure of variation per se. Of course, adjustment of lung function for age and height adds to the complexity of the problem, but does not change it fundamentally.

Confidence intervals

In research we are more often interested in the average value for a group of individuals, and whether this differs from that of another group. For example, does the mean FEV_1 of children exposed to passive smoking differ from that of non-exposed children? If we studied two such groups of children aged 8, we might wish when reporting the results to describe the variation of lung function in each group first. For this we could use the standard deviation of FEV_1 within each group; this is better than the range from minimum to maximum, which is influenced by sample size and does not use all the data. We could derive a 95% range for each group if we wished.

Having described each group, however, we would want to decide whether the mean FEV_1 values for the two groups indicated an effect of passive smoking on lung function. The two mean values are very unlikely to be exactly the same. The mean for the non-exposed group is, we would hope, a good estimate of the mean FEV_1 of all children aged 8 not exposed to passive smoking. From the sample mean and its *standard error* we can calculate a *confidence interval*, in which we expect the true, or population, mean to lie. Much of statistics is about how to calculate standard errors correctly in complicated circumstances, but if we leave aside the question of adjusting lung function for height the calculation is straightforward in this case. The standard error of a sample mean is calculated as the sample standard deviation divided by the square root of the sample size; it is a measure of the chance variation we expect to see in mean values of

Department of Public Health Medicine, United Medical and Dental Schools of Guy's and St Thomas's, St Thomas's Campus, London SE1 7EH
S Chinn

Address for reprint requests: Miss Chinn

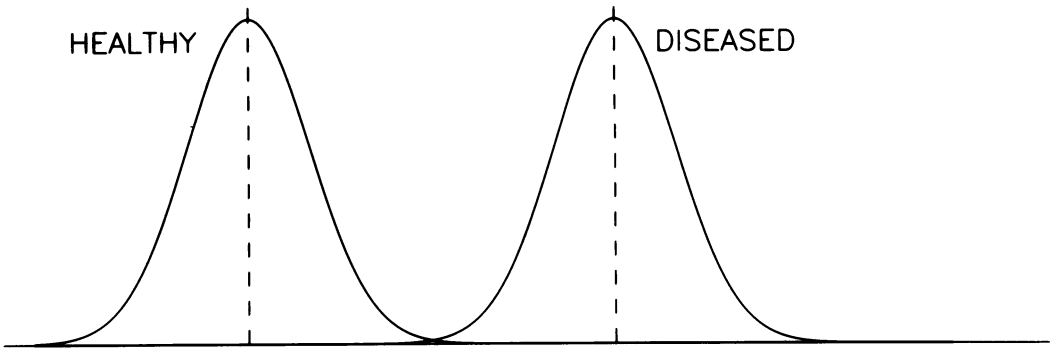

Figure 1 Hypothetical distributions of a measurement that separates healthy and diseased subjects.

samples of the chosen size from the underlying distribution of FEV_1 of children aged 8 years not exposed to passive smoking. A standard error is thus a step towards obtaining the more useful quantity, the confidence interval.

If the sample size n is greater than 60 then the 95% confidence interval will be approximately the mean ± 2 standard errors. If the sample size is less than 60 the factor 2 should be replaced by the value from the t distribution with $(n - 1)$ degrees of freedom that cuts off 2·5% of the distribution at each end (the critical value for a two tailed significance test at the 5% level). We can calculate the mean and confidence interval similarly for the exposed group. If the two confidence intervals do not overlap then we can be sure that the probability that the observed (or a greater) difference in the mean values occurred by chance is less than 0·05, the probability threshold at which we conventionally prefer to believe that something other than chance has an effect. Usually, however, we would like to know how big or small the difference between the two mean values is likely to be, and if the confidence intervals overlap we must calculate a confidence interval for the *difference* in the means. Again the first step is to calculate a standard error for the difference in the mean values, which should be that used in the unpaired, or two sample, t test (that is, based on a pooled variance) provided that the assumptions of the t test are valid. A 95% confidence interval for the mean difference can then be calculated. If the confidence interval does not overlap zero we say that the two mean values are "significantly different at the 5% level," and reject the notion (null hypothesis) that the mean values of the two populations are identical. The confidence interval for the difference in means is more useful because it shows how big the difference in mean values might be and whether the maximum difference between non-exposed and exposed children is big enough to suggest that passive smoking is an important hazard to lung function. Whether we can interpret the difference in mean values in this way depends on how well we chose our two samples of children, and whether they were comparable in all respects other than passive exposure to smoking.

The assumptions of t tests will be described in the third article. The confidence interval calculations used most frequently are described by Gardner and Altman.[3]

Notation and graphical representation

Altman and Gardner[4] drew attention to the confusion that use of a $\pm$ sign can produce, and recommended that all journals adopt the policy of using SD or SE to denote standard deviation or standard error as appropriate. If the more useful 95% range or 95% confidence interval is calculated these two must be distinguished; the latter can be abbreviated to 95% CI. The choice will depend on the purpose, as described above. Figures in papers or shown in verbal presentations must distinguish between "error bars" that may denote any of four possibilities. Again 95% range or 95% confidence interval bars are preferred to mean and SD or mean and SE bars.

Diagnostic tests

A reference range may be used to suggest who is healthy and who needs further investigation.

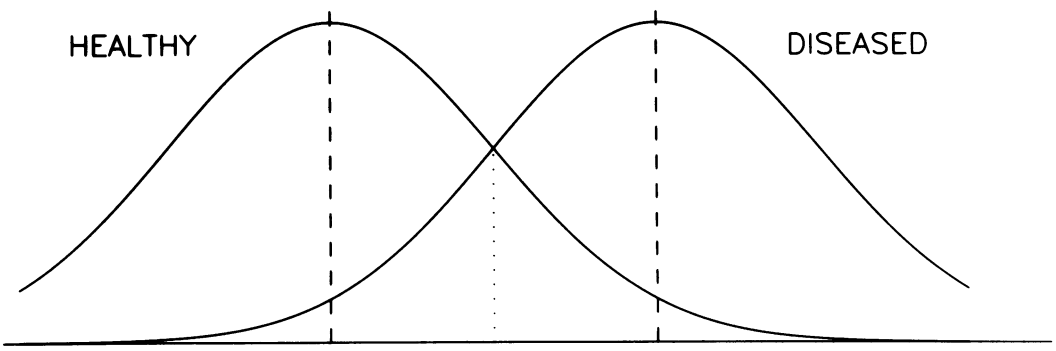

Figure 2 Hypothetical distributions of a measurement that would show a significant difference in mean values between healthy and diseased groups with very few subjects, but with a cut off point (.....) would misclassify 16% of each group.

Some measurements may be used as a diagnostic tool for a specific disease. For a measurement to be useful it must not just differ on average between subjects with the disease in question and those without, but must take distinct or nearly distinct values in the two groups (fig 1). For example, if we proposed to use percentage increase in FEV_1 on inhalation of salbutamol as an aid to diagnosis of asthma we would have to establish the 95% *range* for known asthmatic patients and for non-asthmatic individuals; only if the two ranges have little overlap is bronchodilatation useful as a diagnostic test. This is quite different from establishing a *significant* difference between mean response in the two groups, which would of course be achieved with quite small samples (fig 2). Although a difference that is not significant, unless obtained with a very small sample, would indicate that a measurement is not of use in diagnosis, a significant difference does not necessarily imply diagnostic usefulness. In determining whether a measurement is of use in diagnosis a direct measure of separation of the diagnostic groups should be used, such as the index of separation.[5,6] If a cut off point on the scale is established above or below which a subject is said to give a positive response to the test, then sensitivity and specificity can be defined.

This section deals only with tests used to make an initial diagnosis. Reference ranges for monitoring changes in patients will be described in article 2.

Conclusions

Careful thought should be given in the use of ranges and confidence intervals, or standard deviations and standard errors. Ranges and standard deviations measure variation in individual measurements, whereas confidence intervals and standard errors relate to the precision of population estimates. Unfortunately terminology has been misused often in the past; even in the pages of this journal[7] a 95% range has been called a 95% confidence interval. Emphasis that a 95% range is not a 95% confidence interval can be made by denoting it a reference range when applicable, or a 95% *tolerance* range (but this term has not been adopted universally). Research workers should beware of claiming diagnostic or discriminating power for measurements that have displayed statistically significant differences between groups.

1 Quanjer Ph, ed. Standard lung function testing. *Bull Eur Physiopathol Respir* 1983:suppl 5.
2 Healy MJR. Notes on the statistics of growth standards. *Ann Hum Biol* 1974;1:41–6.
3 Gardner MJ, Altman DG, eds. *Statistics with confidence: confidence intervals and statistical guidelines*. London: British Medical Journal, 1989.
4 Altman DG, Gardner MJ. Presentation of variability [letter]. *Lancet* 1986;ii:639.
5 Armitage P, Berry G. *Statistical methods in medical research*. 2nd ed. Oxford: Blackwell, 1987:475.
6 Chinn S. The assessment of methods of measurement. *Statistics in Medicine* 1990;**9**:351–62.
7 Dehaut P, Rachiele A, Martin RR, Malo HL. Histamine dose-response curves in asthma: reproducibility and sensitivity of different indices to assess response. *Thorax* 1983;**38**:516–22.