Statistics in Question

SHEILA M GORE

ASSESSING METHODS— TRANSFORMING THE DATA

Transformation is often needed to analyse medical data competently. There are four reasons for transforming data, but most often in practice a transformation selected for one purpose conveniently achieves the others as well. The reasons for transformation are as follows. (a) To arrange that the dispersion of coded observations is similar across all treatment or diagnostic groups¹—so that changes in mean or median are not obscured by greater variability in some groups than in others. (b) To induce normality—because many statistical tests such as Student's *t*-test require that the data are normally distributed. It is better to work with derived observations that have a Gaussian (normal) distribution than to use less powerful statistical methods on the untransformed data—for example, the logarithm of plasma cholesterol, being normally distributed,² could be compared between two groups of 20 patients each, say, using Student's *t*-test but not the plasma cholesterol concentration itself. (c) To produce simple relationships—the eye easily assimilates straight lines but is less good at detecting other patterns, such as quadratic, cubic, or exponential. By transforming one or both of a pair of variables an association can sometimes be presented as linear on the transformed data when it was more complicated on the original observations. (d) So that there is good agreement with biological or physiological principles—the plasma creatinine concentration indicates the level of renal function.³ The kidney may be thought of as a sink or reservoir maintained at a constant level by drainage, and since the product of its input and output is constant the one is the inverse of the other, so that it makes sense to study the inverse of plasma creatinine as a monitor of renal function. The impressive work of Dr M S Knapp and his colleagues at the City Hospital, Nottingham in collaboration with Professor A F M Smith (Department of Mathematics, University of

Nottingham), who have used inverse plasma creatinine to monitor the function of kidney transplants, shows that transformation is far from being a statistical nicety, it also has clinical relevance.^{3 4}

(5) Which of the three plots shown in figures 1 and 2 gives the clearest indication of when rejection occurred in a patient who received a renal transplant?

—inverse plasma creatinine concentrations corrected for body weight

COMMENT

Plotting plasma creatinine concentrations against time tells us little because as plasma creatinine concentrations decrease so, dramatically, does their variance. Plotted as in figure 1 the points therefore do not convey equal information, making the graph difficult to interpret. The need to transform the data was

Rejection: day 24 approximately

FIG 1—Before transformation.

Conversion: SI to traditional units—Plasma creatinine: 1 μmol/l ≈ 88.4 mg/100 ml.

recognised because the distribution of plasma creatinine concentrations is highly skewed and by thinking about renal physiology. By plotting the inverse of plasma creatinine against time an abrupt change at approximately 24 days after transplantation for this patient becomes apparent. Because changes in plasma creatinine concentrations are caused by variations in the volume of distribution, which is commonly a result of fluid retention or diuresis, the Nottingham team first corrected plasma creatinine for body weight.³

The adjustment for weight is important (see figure 2) because the time of rejection now appears as the intersection of two

FIG 2—After transformation and correction for weight. The numbers in parentheses are plasma creatinine concentrations on the original scale of $\mu\text{mol/l}$ —for example, 1 over 1000 $\mu\text{mol/l}$ = reciprocal plasma creatinine, namely 0.001.

straight lines having different slope. This simple representation, by no means obvious from figure 1, has been arrived at by carefully considering the statistical and physiological reasons for transforming the original observations. An interesting result of the Nottingham studies is that rejection is more common at night, having apparently a circadian rhythm, and is likely first to alter creatinine clearance at around 6 am.⁴ The idea of making a correction for weight when concentrations are measured has been investigated also by Wald and his colleagues at Oxford.⁵ They found that the maternal serum alpha-fetoprotein concentration is influenced by maternal weight and propose an appropriate adjustment. This type of transformation is likely to be generally important.

(6) *If asked to investigate an association between leucocyte count at initial assessment and subsequent survival time for patients with unresectable lung cancer, which transformations would you consider when plotting the data?*

—logarithmic transformation of survival time

—square-root transformation of leucocyte count

—when graphs are used to explore data the choice of

the correct transformation is less important in practical terms than realising when some transformation is called for

COMMENT

When graphs are used to explore data the choice of the "correct" transformation is less important in practical terms than realising that "some" transformation ought to be tried and knowing broadly how to proceed. Mention of survival time should conjure up a picture of non-negative positively-skewed data and the possibility that variance increases as mean survival time squared—which it does indeed if survival is exponential. Consideration should therefore be given to plotting the logarithm of survival time—the "correct" transformation when variance increases as mean squared, and a sensible approach when dealing with survival data generally, provided that survival times are not too close to zero. Likewise, mention of counts triggers the thought that as mean count increases so, as a rule, does the variance—making transformation appropriate. The "correct" transformation in this case would be the square root of the original counts, but taking logarithms instead—a more extreme transformation—will not usually do more harm than failing to transform the data at all.

As an aside I mention the logit transformation for proportions, which will be discussed in a later article. When plotting proportions—for example, of schizophrenic patients who develop tardive dyskinesia—against some risk factor, such as age or neuroleptic dose measured in equivalents of chlorpromazine, there may be need for transformation: because the numbers at risk in different subgroups vary considerably or because the proportion of failures—that is, patients with tardive dyskinesia—ranges outside 0.3–0.7 wherein variance is relatively stable provided that the groups at risk are of similar size, or both. The reason for transformation is again to find a scale on which variance will depend neither on the estimated proportion nor on the number of patients in different groups at risk.

Reverting to the question of whether leucocyte count at initial assessment and survival time are associated for patients with unresectable lung cancer, all that remains is to plot for each patient the logarithm of survival time against the square root of leucocyte count and inspect the scattergram for evidence of association. Take care to distinguish between deaths and the survival times of patients who are still alive, using different symbols for exact (●) and censored (x) observations. The scattergram would be difficult to interpret if follow-up was incomplete on many patients so that censored observations were frequent, but this would be unlikely in a retrospective survey of patients with unresectable lung cancer. Life-tables have the advantage that they take proper account of censored survival times. An alternative plan, if sample size was moderate, would be to separate patients into three or more groups according to initial leucocyte count and to plot the life-table for each group of patients. The one difficulty here is deciding on the appropriate grouping. Any classification that is suggested by the data needs to be viewed cautiously and requires independent verification. A defensible solution is to group together the lower, middle, and upper-third of the distribution for leucocyte counts.

The suggestion⁶ that the survival time of patients with unresectable lung cancer is related to initial leucocyte count was not supported by correspondents.⁷ * None of the authors published a scattergram for the information of readers, though Snell⁷ had clearly constructed such a plot when exploring his data, and the observation by Huhti *et al.*⁸ that high leucocyte counts were predominant only for patients who survived less than one month, could have derived from similar inspection. Publication of these figures would have needed little extra space and would have been more informative and less arbitrary than the original tabulation.⁶ Authors should note that when transformed data are plotted or reported it is highly desirable to link the new observations to the original scale, providing, as

it were, a translation for the reader into the language he is familiar with.

(7) *Tumour mass was measured clinically and recorded for 294 patients with breast cancer; the mean tumour size was 4.2 cm and standard deviation 2.7 cm. What do you infer from the above summary about the distribution of tumour size?*

- tumour size is not normally distributed
- there is a small proportion of very large tumours

COMMENT

If tumour size were normally distributed we should expect 95% of the observations to lie in the interval from approximately two standard deviations below the mean to roughly two standard deviations above the mean (the exact multiple is 1.96). This is the interval from $4.2 - 2 \times 2.7 = -1.2$ cm to $4.2 + 2 \times 2.7 = 9.6$ cm, but the implication that 2½% of patients have tumour size less than “-1.2 cm” is clearly nonsense. The observations on tumour size are therefore not normally distributed (*reductio ad absurdum*); it is likely that the distribution is positively skewed, as indeed is the case (see table I). Some

60% of patients have tumour size less than 4.2 cm instead of the 50% we should expect from a Gaussian distribution. How could we have derived a more sensible interval? The long upper tail to the distribution of tumour size points to a logarithmic transformation, although taking square roots gives a satisfactory enough answer (see table II)—better at least than no transformation. The mean of $\log_e$ (tumour size) was 1.24 and standard deviation 0.66. The antilog of 1.24 is 3.5 cm, and the corresponding interval from two standard deviations below to two standard deviations above the mean—reported in the original units—is from antilog $(1.24 - 2 \times 0.66) = 0.9$ cm to antilog $(1.24 + 2 \times 0.66) = 12.9$ cm. Notice that this interval excludes the lower 4% and upper 2% of patients, agreeing reasonably well with the expectation that about 2½% would be excluded in each tail if $\log_e$ (tumour size) were normally distributed. Intervals calculated using logarithmic, square root, and no transformation are compared in table II, as are the observed numbers of patients whose tumour size exceeds the estimated 99th percentile—we should expect there to be three such patients. The interval derived using the logarithmic transformation is the best representation, in other words, $\log_e$ (tumour size) most closely of the three follows a normal distribution.

I am grateful to Dr M S Knapp and his colleagues at the Renal Unit, Nottingham and to Professor A F M Smith, Department of Mathematics, Nottingham University for permission to reproduce figures 1 and 2.

Sheila M Gore, MA, is a statistician in the MRC Biostatistics Unit, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH.

TABLE I—Measured tumour size (rounded up to the nearest centimetre)

Tumour size (cm)	Frequency (n=294)
1	12
2	29
3	73
4	56
5	50
6	25
7	11
8	12
9	6
10	10
11	0
12	4
13	0
14	1
15	3
16	1
17	0
18	0
19	0
20	1

References

¹ Rogers K, Roberts GM, Williams GT. Gastric-juice enzymes—an aid in the diagnosis of gastric cancer? *Lancet* 1981;ii:1124-5.
² Flynn FV, Piper KAJ, Garcia-Webb P, McPherson K, Healy MJR. The frequency distributions of commonly determined blood constituents in healthy blood donors. *Clin Chim Acta* 1974;52:163-71.
³ Knapp MS, Blamey R, Cove-Smith R, Health M. Monitoring the function of renal transplants. *Lancet* 1977;ii:1183.
⁴ Knapp MS, Cove-Smith JR, Dugdale R, MacKenzie N, Pownall R. Possible effect of time on renal allograft rejection. *Br Med J* 1979;ii:75-7.
⁵ Wald N, Cuckle H, Boreham J, Perzian E, Redman C. Effect of maternal weight on maternal serum alpha-fetoprotein levels. *Br J Obstet Gynaecol* 1981 (in press).
⁶ Check IJ, DeMeester T, Vardiman J, Hunter RL. Differential counts and survival in lung cancer. *Lancet* 1978;ii:1317-8.
⁷ Snell NJC. Leucocyte-counts and survival in unresectable lung cancer. *Lancet* 1979;ii:383-4.
⁸ Huhti E, Poukkula A, Saloheimo M. Leucocyte-counts and survival in lung cancer. *Lancet* 1979;ii:1348.

No reprints will be available from the author.

TABLE II—Tumour size: comparison of transformations

Transformation	Interval representing the central 95% of tumour size distribution (cm)	Excluded		Estimated 99th percentile	
		Lower tail	Upper tail		Exceeded by (expectation: 3)
Logarithmic	0.9-12.9	4%	2%	16.0 cm	1 patient
Square root	0.5-10.2	2%	3%	11.5 cm	8 patients
None	-1.2- 9.6	0%	5%	10.5 cm	10 patients