LETTERS TO THE EDITOR

Estimating sample sizes for studies using the SF-36 health survey

SIR-You published a paper by Julious *et al*[1] on the calculation of sample sizes for intervention studies using the short form 36 (SF-36) health survey. The paper highlighted discrepancies between parametric and non-parametric techniques when calculating these sample sizes. Taking into account the discrete (ordinal) nature and the non-normal distributions of the SF-36 scale scores, the authors proposed the use of non-parametric methods for this purpose. Based on a non-parametric equation, these authors proposed a sample size of 1401 individuals in *each* group when comparing results for the pain dimension of the SF-36. This number of subjects is clearly much larger than the sample size estimation of 70 individuals calculated with a parametric equation for the same purpose. It also disregards empirical published evidence which reports significant changes in the SF-36 scales using dramatically smaller sample sizes.[2] When the economic aspects of any intervention study are considered, and taking into account a number of theoretical considerations we discuss below, it seems that by using a non-parametric method pounds are being thrown away to save pennies.

The use of non-parametric statistics when considering discrete scales was first introduced by Stevens.[3] He distinguished four scales of measurement - nominal, ordinal, interval, and ratio. Stevens specified the statistics which would be permissible for each scale: non-parametric procedures were appropriate with nominal and ordinal scales, whereas parametric procedures were required for interval and ratio scales. The first obstacle in this reasoning concerns the distinction between scale types. Between ordinal and interval scales there are a number of instruments (such as the SF-36) that can be labelled as 'summated scales' and in which the total score is the sum of a set of ordinal rankings. There is no claim that equal increments in the observed score along the summated scale represent equal increments in the underlying latent variable being measured, but the mode of construction suggests that the deviations from interval properties will not be extreme. Secondly, several authors have pointed out[3] that there is no relationship between type of scale and statistical techniques used. Even though the use of parametric methods requires more assumptions than non-parametric methods, failure to meet these assumptions does not appear to have serious consequences in most instances.[4] For these reasons, and in opposition to Julious *et al*,[1] we believe that parametric techniques might be used for SF-36 sample size calculations.

These authors[1] additionally presented their results based on an arbitrarily chosen difference (one discrete value away from the population mean or median). This may lead to confusion and to the incorrect interpretation of the proposed sample sizes (obtained with parametric or non-parametric equations) as the standard to consider in future research. Parametric estimates of sample sizes necessary to detect small to large group differences in average SF-36 scale scores have already been published in the *SF-36 Manual and Interpretation Guide*[5] and are reproduced in table 1, with permission. The differences in scores modelled were selected to represent very small to large differences (2-20 points of difference). An alternative approach - the minimal clinically important difference (MCID) - has been suggested when calculating these estimations. The MCID can be defined as the smallest difference in score in the domain of interest perceived as beneficial.[6] Unfortunately, the MCID for the SF-36 scales are still unknown and further research is needed to elucidate the clinical significance of its score changes.

Table 1 Sample size needed to detect 2-20 point differences in the short form 36 (SF-36) between a group mean and a fixed norm

SF-36 scale	No of points difference			
	2	5	10	20
Physical functioning	1067	171	44	12
Role - physical	2282	366	92	24
Bodily pain	1103	177	45	12
General health	818	132	34	9
Vitality	866	139	36	10
Social functioning	1012	163	41	11
Role - emotional	2152	345	87	22
Mental health	644	104	27	8

Estimate assumes alpha = 0.05, two tailed *t* test, power = 80%. Source: reference 5.

Table 2 Short form 36 (SF-36) minimum score differences to consider change when assessing differences between a group mean and a fixed norm

SF-36 scale	SD	No of points difference		
		Small (ES=0.2)	Moderate (ES=0.5)	Large (ES=0.8)
Physical functioning	21.27	4.3	10.6	17.0
Role - physical	32.40	6.5	16.2	25.9
Bodily pain	23.24	4.6	11.6	18.6
General health	21.08	4.2	10.5	16.9
Vitality	21.28	4.3	10.6	17.0
Social functioning	21.17	4.2	10.6	16.9
Role - emotional	33.32	6.7	16.7	26.7
Mental health	19.07	3.8	9.5	15.3
Parametric sample size		**393**	**64**	**25**

ES = effect size. Estimates assume alpha = 0.05, two tailed *t* test, power = 80%.

An alternative score difference to be considered is based in the traditional statistical 'effect size' (ES) calculation.[7] The usual calculation of ES is given by: –

$$ES=(m_1-m_2)/s_1 \quad (equation\ 1)$$

where m_1 and m_2 are the fixed norm and the study group means, and s_1 is the standard deviation of the fixed norm. Some authors have suggested[7] that an ES of 0.2 represents a small change, one of 0.5 a moderate change, and one of 0.8 or greater a large change in health status. Using these benchmarks as a reference and the standard deviation of the fixed norm, the difference between the means needed to produce small, moderate or great changes can be easily calculated by replacing values in equation 1. This difference may be used then to obtain more appropriate sample sizes.

Considering the results provided by Julious *et al*,[1] table 2 presents minimum score differences to obtain a small, a moderate, or a large change between a group mean and a fixed norm for each SF-36 scale. Parametric sample size estimates considering these differences were also calculated resulting in sample sizes of 25 to 393 individuals per group. While using the approach of Julious *et al*,[1] researchers should identify the SF-36 dimension of primary interest upon which to base the sample size and treat the others as a secondary, under the ES approach, investigators can use the same sample size estimate for each SF-36 scale.

We believe that these alternative sample size calculations, which are statistically justified and conceptually appropriate, will not discourage including the SF-36 in clinical and epidemiological research.

LUIS PRIETO
JORDI ALONSO
JOSEP M ANTÓ
Institut Municipal d'Investigació Mèdica Dr Aiguader 80,E-08003 Barcelona, Spain

1 Julious SA, George S, Campbell MJ. Sample sizes for studies using the short form 36 (SF-36). *J Epidemiol Community Health* 1995;**49**:642-4.
2 Ware JE, Kosinski M, Keller SD. *SF-36 Physical and mental health summary scales: a user's manual.* Boston, MA: New England Medical Center, 1994.
3 Gaito J. Measurement scales and statistics: resurgence of an old misconception. *Psychol Bull* 1980;**87**:564-7.
4 Heerman EF, Braskamp LA. *Readings in statistics for the behavioural sciences.* Englewood Cliffs, NJ: Prentice-Hall, 1970.
5 Ware JE, Snow KK, Kosinski M, Gandek B. *SF-36 health survey manual and interpretation guide.* Boston, MA: New England Medical Center, 1993.
6 Jaeschke R, Singer J, Guyatt GH. Ascertaining the minimal clinically important difference. *Controlled Clin Trials* 1989;**10**:407-15.
7 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;**27**:S178-89.

Reply

Prieto *et al* suggest that because the failure of assumptions governing parametric tests is often not serious (presumably in terms of the achieved significance level), one can use parametric methods for determining sample sizes, even in circumstances where parametric assumptions are not warranted. Unfortunately, this is based on a misunderstanding of the basis of sample size calculation.

The reason that parametric tests are generally successful is that, because of the central limit theorem, estimates of parameters tend to be normally distributed, even if the original distribution is not. This is highlighted by the fact that many confidence intervals are based on ± 2 standard errors.

However, for determining a sample size we need to specify an effect size based upon a standard deviation, not upon a standard error. It is the distribution of the population, not the estimate, that is important, and the standard deviation for data that are not Normally distributed is uninterpretable. Prieto *et al* also state that our calculations,[1] are 'against empirical published evidence which reports significant changes in the SF-36 scales using dramatically smaller sample sizes'. This, however, is a circular argument: such 'significant' results will have been obtained from tests using parametric methods, which are inappropriate to the data, and they certainly do not prove the applicability of parametric methods to sample size calculation in this case.

In our paper we highlighted the importance of considering the distribution of the SF-36 scores when considering a sample size. For example, the dimension role limitation can only take four values and, in the data set reported by Brazier,[2] most of the population score 100%. In practice, an apparent interval scale may be composed of several highly correlated responses, so that the final response is bimodal. In this case, our methods demonstrate that the sample size approaches the size required for a binary variable, as one might sensibly expect.[3] Incidentally, scoring the results as a percentage hides the fact that the data are discrete, not continuous.

For data with a marked skewness it is also important to consider the direction of the effect. Directionality is not taken into account using a normal approach to skewed data. For example, detecting a difference of 20% success versus 10% success requires 199 patients per group, with 80% power and 5% (two sided) significance level, whereas 20% success versus 30% success requires 294 patients per group.

Lastly, Prieto *et al* state that MCIDs for the SF-36 are still unknown. We are currently engaged in research to establish what are realistic and clinically meaningful effect sizes for a number of quality of life measures. In the meantime, authors should present their results as median, rather than mean, scores, together with ranges and interquartile ranges rather than standard deviations. Where appropriate, hypothesis testing should be performed using non-parametric methods.

M J CAMPBELL
S A JULIOUS
Medical Statistics and Computing

S L GEORGE
Public Health Medicine, University of Southampton.

1 Julious SA, George S, Campbell MJ. Sample sizes for studies using the short form 36 (SF-36). *J Epidemiol Community Health* 1995;**49**:642-4.
2 Brazier JE, Harper R. Jones NNB, O'Cathain A, Thomas MJ, Usherwood T, Westlake L. Validating the SF-36 health questionnaire; new outcome measure for primary care. *BMJ* 1993;305:160-4.
3 Julious SA, George S. Sample size estimates for quality of life data. ISCB 16. *Proceedings of the International Society of Clinical Biostatistics Conference*, Barcelona, 1995.

Suicide from the Clifton Suspension Bridge in England

SIR - Nowers and Gunnell in their study on suicides from the Clifton Bridge[1] quite rightly emphasise the notoriety of this bridge as a suicide site. It is, however, interesting that some equally famous bridges do not attract this reputation.

The Humber Suspension Bridge opened in Hull in 1982, becoming both the world's longest suspension bridge and a highly visible local landmark. In the 10 years after its opening, however, only four deaths were reported from falls from the bridge.[2] All received suicide verdicts at the Coroner's Court, and all had travelled between 8 and 40 miles to their chosen site of death. During this period, two Hull residents had also travelled 40 miles to the Valley Bridge at Scarborough where they jumped to their deaths.

The preference for one bridge over another is most clearly seen in the study of Seiden.[3] Fifty per cent of his series of 115 suicides travelled over the Oakland Bay Bridge to reach the Golden Gate Bridge where they jumped. No cases of people travelling over the Golden Gate Bridge to jump off the Oakland Bay Bridge were found. Similarly, Cantor and Hill in their later study from Australia[4] reported a recent preference for the newly opened Gateway Bridge over the neighbouring Story Bridge in Brisbane. While the Story Bridge had been a traditional suicide site since its opening in 1935, the newer bridge had 17 suicides in its first 17 months of use compared with two from the older bridge. The authors suggested that the first suicide from the new bridge, which occurred during the televised opening ceremony in 1986, may have influenced patterns of choice and also demonstrated significant differences between those who used each bridge both demographically and in regard to psychiatric history.

Further comparison between bridges may provide us with insight as to why certain bridges are more attractive for suicide for certain people. Are there factors in the design or location of bridges that should be considered to prevent deadly reputations developing? If such existed, a further means of suicide prevention would be available. Modifications to existing structures would be able to utilise these findings and new bridges could be designed accordingly. This is clearly an area which warrants further research.

ANDREW M ELLIS
Mossley Hill Hospital, Park Avenue, Liverpool L18 8BU.

1 Nowers M, Gunnell D. Suicide from the Clifton Suspension Bridge in England. *J Epidemiol Community Health* 1996; **50**:30-32
2 Record books of HM Coroner for Hull and North Humberside. Hull: 1972-92
3 Seiden RH. A tale of two bridges: comparative suicide incidence on the Golden Gate and San Francisco – Oakland Bay bridges. *Omega* 1992;**14**:200-9
4 Cantor CH, Hill MA. Suicide from river bridges. *Aust N Z J Psychiatry* 1990; **24**: 377-80

NOTICES

17th International Colloquium, ISSA Chemistry Section, Plant Safety in the Chemical Industry, 9-11 June 1997 in Frankfurt/Main, Germany. For further information: Secretariat of the ISSA Chemistry Section, c/o BG Chemie, Kurfürsten-Anlage 62, D-69115 Heidelberg, Germany. Tel: + 49 6221 523 498. Fax: + 49 6221 523 420.

European Journal of Public Health - volume 6, number 2, June 1996