Journal of Epidemiology and Community Health 1995;49:642-644

Sample sizes for studies using the short form 36 (SF-36)

Steven A Julious, Steve George, Michael J Campbell

When designing a study to compare the outcomes of an intervention, an essential step is the calculation of sample sizes that will allow a reasonable chance (power) of detecting a predetermined difference (effect size) in the outcome variable, at a given level of significance. Sample size is critically dependent on the proposed effect size: half the effect size and the sample size is quadrupled. Unlike power and level of significance, for which norms and precedents dictate values, the effect size must be determined from experience, published data, or pilot studies. It is variously defined as the "minimum value worth detecting" or a "clinically important effect", or "quantitatively significant effect".[1]

The short form 36 (SF-36) health survey questionnaire is a multi-dimensional measure of perceived health status originally developed in the USA.[2] It has been adapted for use in UK populations,[3] and UK population norms have recently been made available for adults of working age.[4] The SF-36 has been compared in "normal" populations with the Nottingham health profile,[3] and has been reported to be preferable for measuring improvements in health in a population with relatively minor conditions such as in general practice or in the community. This is because more subjects use a wider range of scores, which leads to a greater power to discriminate between groups. This variation has lead Ziebland to suggest that it is an inadequate tool to assess health interventions aimed at heterogeneous communities.[5] Scores have been quoted as means and standard deviations, implying that parametric methods should be used to estimate sample sizes,[4] and indeed this is the methodology recommended in the SF-36 manual.[6] However, the distributions of SF-36 dimension scores are not "normal".[3] This paper highlights discrepancies between sample sizes for intervention studies using the SF-36 calculated using conventional parametric techniques[7] and a non-parametric approach.[8]

Departments of Medical Statistics and Computing
S A Julious
M J Campbell

and Public Health Medicine
S George

University of Southampton, Level B, South Academic Block, Southampton General Hospital, Tremona Road, Southampton SO16 6YD

Correspondence to:
Mr S A Julious.

Accepted for publication June 1995

J Epidemiol Community Health 1995;49:642-644

Methods

The SF-36 is a short questionnaire with 36 items which give scores in eight dimensions. Scores for each dimension are expressed as a percentage. However, each dimension score is composed of responses to a small number of questions (see table 2) with the scores for each question ranging from 1 to 2 for "yes/no" questions, through to 1 to 6 where subjects have to grade some aspect of their health. Hence, in a single individual there is a finite number of discrete values that can be obtained, in some dimensions as low as four. For the purposes of calculating sample sizes we have arbitrarily assumed that the effect size of interest is given by one discrete value (plus or minus) away from the population mean or median. On the percentage scale this could be as high as 33%. Data were obtained from a previously published study of subjects on GP lists in Sheffield, a population with relatively minor medical conditions.[3]

Sample sizes were calculated using a parametric approach where the standardised difference is used[7] (effect size over the SD) and by the technique described by Whitehead[8] using the assumption of proportional odds between groups. For increases in pain and physical functioning dimensions the scores were collapsed into just two distinct values so methods for binary sample sizes were employed.[7]

The following two formulas give the number of subjects required, m, in each groups for a two sided significance level α and power $1-\beta$. $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the appropriate values from the standard normal distribution for the $100(1-\alpha/2)$ and $100(1-\beta)$ percentiles respectively.

PARAMETRIC METHOD The sample size required when assuming that the data have a normal distribution for a given effect size δ is given by:

$$m=\frac{2*(z_{1-\alpha/2}+z_{1-\beta})^2}{d^2}+\frac{z_{1-\alpha/2}^2}{4} \qquad (1)$$

where d is the standardised difference, defined as $d=\delta/\sigma$, and σ is the population SD of the measurements.

NON-PARAMETRIC METHOD When no assumptions are made about the data (apart from proportional odds), the estimated sample size can be obtained from:

$$m=\frac{6(z_{1-\alpha/2}+z_{1-\beta})^2/(\log OR)^2}{[1-\sum_{i=1}^{k}\bar{p}_i^3]}$$

OR is the odds ratio of a subject being in category i or worse in one group compared to the other, k is the number of categories, and $\bar{p}_i$ is the mean proportion expected in category i - that is, $\bar{p}_i=(p_{Ai}+p_{Bi})/2$ where p_{Ai} and p_{Bi} are the proportions expected in category i for the two groups A and B respectively.

Table 1 Frequency of responses for a GP population group and study group

Social functioning		*Population group*		*Study group*	
Category	*Percentage scale*	*Percentage* (p_{A_i})	*Cumulation percentage* (c_{A_i})	*Percentage* (p_{B_i})	*Cumulative percentage* (c_{B_i})
1	0·00	0·5	0·5	0·7	0·7
2	11·11	1·3	1·8	1·8	2·5
3	22·22	1·3	3·1	1·8	4·3
4	33·33	1·9	5·0	2·6	6·9
5	44·44	2·7	7·7	3·7	10·6
6	55·66	3·9	11·6	5·2	15·8
7	66·67	7·3	18·9	9·3	25·1
8	77·78	9·0	27·9	10·6	35·7
9	88·89	13·0	40·9	14·3	50·0
10	100·00	59·1	100·0	50·0	100·0

Results

The first two columns of table 1 give the frequency and cumulative frequency, respectively, of responses for social functioning from subjects on GP lists in Sheffield.[3] The median score for this dimension is 100%. The population group (A) is taken from the survey. Now suppose we wished to undertake another study (B) to investigate differences between the study and population groups. Assuming that the effect size of interest is one discrete value away, then the median value of interest in the second group would be 88·89 or category 9 – that is we expect 50% of subjects to be in category 9 or less.

The OR is a measure of the chance of a subject being in a given category or less in one group compared to the other. For category 9 it is given by $OR = \{c_{A9}/(1-c_{A9})\}/\{c_{B9}/(1-c_{B9})\}$ – that is, $(40{\cdot}9/59{\cdot}1)/(50{\cdot}0/50{\cdot}0) = 0{\cdot}692$. Under the assumption of proportional odds, the expected proportions in the other categories can now be calculated, such that for category 8 $(27{\cdot}9/72{\cdot}1)/\{c_{B8}/(1-c_{B8})\} = 0{\cdot}69_2$, hence, $c_{B8} = 35{\cdot}7$. Similarly, the cumulative proportions can be calculated for the other 7 categories and from these the expected proportions derived and the final 2 columns in table 1 completed. The mean proportion for each category can now be estimated such that: $\bar{p}_1 = (0{\cdot}005 + 0{\cdot}007)/2 = 0{\cdot}006$, $\bar{p}_2 = (0{\cdot}013 + 0{\cdot}018)/2 = 0{\cdot}016$, $\bar{p}_3 = 0{\cdot}016$, $\bar{p}_4 = 0{\cdot}023$ ... etc. Thus, the sample size can now be calculated using the above equation.

The results for social functioning and the other dimensions are displayed in table 2. Results for sample sizes making no parametric assumptions reflect the asymmetric nature of each dimension and are therefore given separately for one discrete value above (+) and below (−) the population median. Each sample size is calculated assuming a significance level of 5% and power of 80%. For three dimensions (role limitations, physical and emotional, and social functioning) no value can be given for one discrete value above the median as the population median lies in the uppermost category.

Comment

The results given by these two methods are similar in some dimensions (particularly general health perception), where there are a large number of categories, but are markedly different in others (especially pain), where the SF-36 dimension scores are highly skewed. For asymmetric distributions the parametric methods give the same sample sizes for effects that are one above and one below the expected population mean as they assume the data have a symmetric (normal) distribution, whereas the non-parametric methods may give markedly different sample sizes according to the direction of the expected difference.

In general, statistics such as means and standard deviations are not suitable summary measures for non-normal distributions, and standardised differences are not a suitable basis for calculation of sample sizes. Non-parametric methods should therefore be used for sample size calculation. Workers calculating sample sizes for studies involving the SF-36 should identify the dimension of primary interest upon which to base the sample size estimate and treat the others as secondary. These results should enable investigators to plan their studies and justify the sample size requirements in the

Table 2 Descriptive statistics and sample size for each dimension of the SF-36

Dimension	*Descriptive statistics and sample size*							
	No of questions	*Incremental difference (%)*	*Mean*	*(SD)*	*Median*	*Sample size in each group*		
						Parametric	*Non-parametric*	
							−1	*+1*
General health perception	5	5·00	71·22	(21·08)	77·00	280	251	247
Mental health	5	4·00	72·55	(19·07)	76·00	358	738	217
Pain	2	11·11	79·04	(23·24)	88·89	70	1401	277
Physical functioning	10	5·00	86·23	(21·27)	95·00	285	544	247
Role limitations (physical	4	25·00	82·38	(32·40)	100·00	28	68	—
Role limitations (emotional)	3	33·33	81·64	(33·32)	100·00	17	64	—
Social functioning	2	11·11	86·96	(21·17)	100·00	58	417	—
Vitality	4	5·00	60·80	(21·28)	65·00	286	366	844

protocol. Further work is needed to discover what are realistic changes in scores for health technology interventions.

We would like to thank John Brazier and colleagues at Sheffield for the use of their data.

1 Burnard B, Kernan WN, Feinstein AR. Indexes and boundaries for 'quantitative significance' in statistical decisions. *J Clin Epidemiol* 1990;**43**:1273–84.
2 Ware JE, Brook RH, Williams KN, Stewart AL, Davies-Avery A. *Conceptualisation and measurement of health for adults in the health insurance study. Vol 1. Model of health and methodology.* Santa Monica, CA, Rand Corporation, 1980. Publication no R-1987/1-HEW.
3 Brazier JE, Harper R, Jones NMB, O'Cathain A, Thomas KJ, Usherwood T, Westlake L. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ* 1992;**305**:160–4.
4 Jenkinson C, Coulter A, Wright L. Short form 36 (SF-36) health survey questionnaire: normative data for adults of working age. *BMJ* 1993;**306**:1437–40.
5 Ziebland S. The short form 36 health status questionnaire: clues from the Oxford region's normative data about its usefulness in measuring health gain in population surveys. *J Epidemiol Community Health* 1995;**49**:102–5.
6 Ware JE, Snow KK, Kosinski M, Gandek B. *SF-36 health survey manual and interpretation guide.* Boston, MA: New England Medical Centre, The Health Institute, 1993.
7 Machin D, Campbell MJ. *Statistical tables for the design of clinical trials.* Oxford: Blackwell Scientific, 1987.
8 Whitehead J. Sample size calculations for ordered categorical data. *Statistics in Medicine* 1993;**12**:2257–72.