27th Australasian Transport Research Forum, Adelaide, 29 September – 1 October 2004

Paper title: **Sample size requirements for measuring a change in behaviour**

Author(s) name(s): **Peter R Stopher and Stephen P Greaves**

Organisation(s): Institute of Transport Studies, University of Sydney

Contact details:

Postal address: Faculty of Business & Economics, Burren Street Campus (C37), University of Sydney, Sydney, NSW 2006

Telephone: 61-2-9351-0078
Facsimile: 61-2-9351-0088
email: stepheng@its.usyd.edu.au

Abstract (200 words):
Before and after surveys are designed to detect a change in travel-behaviour following an intervention policy, such as a travel-modification program. Longitudinal panel surveys are the preferred method for detecting such changes, because the variance of the difference between the before and after surveys is substantially reduced, enabling changes to be detected with smaller sample sizes than if a repeated cross-sectional survey is used. A key issue concerns the size of sample required to be able to generalise the Panel results to the population; that is to state, with 95% confidence that if there is a ∂% change in behaviour for the sample, there is a ∂%± e% change in the behaviour of the population, where e is the sampling error. In this paper we present the rationale for an alternative formulation and demonstrate its applicability both hypothetically and then empirically using data from the Puget Sound Transportation Panel. The results have important ramifications both for those implementing future behaviour change programs and those interpreting the results reported in previous studies.

Introduction

A critical issue in transportation planning and policy development is *reliably* evaluating the extent to which a particular intervention results in a change in behaviour for the population under investigation, and to be able to estimate the size of the change to some level of acceptable statistical significance. This issue has received particular attention recently in the context of evaluating voluntary travel behaviour change (VTBC) programs, where the primary question is whether it is possible to achieve significant reductions in vehicle kilometres travelled (VKT) by private automobile, through provision of improved information on travel alternatives (Stopher, 2004). Intuitively, such a question requires a study of travel behaviour before and after an intervention for the population under investigation, from which an assessment of change (if any) can be made. The size of the change is then used to determine whether the VTBC program is worthwhile extending to a larger population.

Before and after surveys involve a host of issues, which are well summarised in Richardson *et al.* (2003). The issue under investigation here is what sample size is necessary to measure a change in a population parameter (e.g., average VKT) by a specified amount. In this paper we argue that an inappropriate formula is currently being used to answer such a question for reasons that are elucidated on in a later section. In response, we propose an alternative formulation for calculating required sample sizes, which enables results to be generalised to the population, based on a simple re-arrangement of the formula for calculating sampling error for the difference in two means (Kish, 1965).

Following the theoretical development, we use this formula to demonstrate the implications of changing statistical parameters (sampling error, covariance) on the sample sizes required to measure behavioural change using data from two recent waves of the Puget Sound Panel Survey. The results have important ramifications both for those conducting future evaluations of before and after programs, and those interpreting the results provided in previous studies.

Measuring a Change in Behaviour – Key Issues

Sampling Error

The issue of concern is what sample size is necessary to measure whether or not a certain size change has occurred in the behaviour of a population over a period of time, either in response to some external stimulus, or simply as a result of normal evolutionary changes. The main concern in using a sample survey to estimate whether or not change has taken place is that a sample always contains error. This error arises from the fact that a sample is only a representation of the population. Indeed, the only way to eliminate such error is to conduct a full enumeration of the population itself. Also, even though we draw a representative sample from the population, there will always be a chance that we miss certain extreme behaviours in the sample, or that we measure an extreme behaviour but that it is disproportionately present in the sample compared to the total population. When measuring a change in behaviour of the population, there is a further contributor to error. This arises from the fact that behaviour is itself variable. Therefore, there will be some level of change that might be detected in two successive samples that is not measuring change over time, but is simply reflective of the inherent variability in behaviour.

As with all survey sampling, as one increases the size of the sample, one expects these errors to be reduced. This comes about because it becomes more likely that extreme and rare behaviours will be represented in the sample in correct proportions. It also occurs because the increasing sample size will reduce the impact of the inherent variability in the phenomenon being measured. Therefore, as sample size increases, one becomes increasingly sure as to whether or not a change is actually being measured – clearly we still need to interpret such change appropriately.

The other principal issue here is that the item of most interest is the measurement of a difference between two occasions in a measure that is itself variable. It is a statistical fact that the variance (*V)* of a difference in the means of two variables, x_1 and x_2 is equal to the sum of the variances of the two variables less twice the covariance (if any) between the two variables (Kish, 1965; Stopher, 1976, Weatherburn, 1962; Benjamin and Cornell, 1970). That is:

$$V(\bar{x}_2 - \bar{x}_1) = V(x_2) + V(x_1) - 2\,\mathrm{cov}(x_1, x_2) \tag{1}$$

$$s.e.(\bar{x}) = \sqrt{\frac{V(x)}{n}} \tag{2}$$

It should be noted that this formula is true, no matter what the underlying distributions are of the variables x_1 and x_2. The sampling error of the mean of any variable is equal to the standard deviation divided by the square root of the sample size *(n)*, so that the sampling error is given by equation (2):

From this, it is clear that the sampling error in measuring a difference between two means must be given by equation (3), where n_1 is the sample size for the first mean and n_2 is the sample size for the second mean (Kish, 1965):

$$s.e.(\bar{x}_2 - \bar{x}_1) = \sqrt{\frac{V(x_1)}{n_1} + \frac{V(x_2)}{n_2} - \frac{2\,\mathrm{cov}(x_1,x_2)}{n_1}} \tag{3}$$

Methods for Measuring Change

There are fundamentally two different methods for measuring a change over time with a survey. The first is to undertake two or more independent, cross-sectional surveys at different points in time and use these surveys to estimate whether or not a change has occurred. In this case, there is no covariance between the two samples, so that the third term in the sampling error equation becomes zero, and the sampling error of the difference in means on the two occasions is given by equation (4):

$$s.e.(\bar{x}_2 - \bar{x}_1) = \sqrt{\frac{V(x_1)}{n_1} + \frac{V(x_2)}{n_2}} \tag{4}$$

The second method is to use a panel to measure change. By definition, a panel entails measuring the same sample on each occasion. In this case, n_1 and n_2 are the same (unlike in

equation (3)), and we can set them equal to a value n. Also, the covariance term will now exist. Therefore, the sampling error is given by equation (5):

$$s.e.(\bar{x}_2 - \bar{x}_1) = \sqrt{\frac{V(x_1)}{n} + \frac{V(x_2)}{n} - \frac{2\,\mathrm{cov}(\bar{x}_1, \bar{x}_2)}{n}} \tag{5}$$

Two major variations exist for a panel. In one the second occasion is a subsample of the first occasion, while in the other, the second occasion partially overlaps the first occasion. We will not concern ourselves with these two situations in this paper.

Determining the Sample Size

Generally, when we wish to determine if a change has taken place over time, using a sample survey, the issue becomes one of specifying with what level of precision we would like to measure that change. This is usually done by deciding what is the approximate expected size of the change, and then determining from that how much error we can tolerate in that measure, and at what level of confidence we would want to accept that level of error. Suppose, for example, that the change we were expecting to measure would be approximately 10 percent. We may then wish to specify that the level of error we could accept would be ±1 percent at 95 percent confidence, meaning that we would know, with 95 percent confidence that the change was between 9 and 11 percent. Similarly, if the expected change would be on the order of 50 percent, we might be willing to accept a level of error of ±5 percent at 95% confidence, meaning that we would know that the change was between 45 and 55 percent with 95 percent confidence. Such a level of error would probably not be acceptable for a change of 10 percent, and even less so if the expected change was only 5 percent.

Assuming that the sample distribution of the means is normal (i.e., if we took many samples, the estimates of the means from the many samples would be normally distributed), the 95 percent confidence bounds are at 1.96 times the sampling error from the sample mean difference, specifying that the difference must be known to within ±1 percent with 95 percent confidence means that we need to have a sampling error that is no larger than 1 percent of the mean divided by 1.96. By specifying the maximum allowable sampling error, we can take the formula for the sampling error, and rearrange it to allow us to estimate the sample size in terms of the allowable sampling error and the variance of the measure on each occasion.

Independent Cross-Sectional Surveys

We will simplify this problem slightly by assuming that it is our intention to draw the same size sample on each occasion. Thus, in the sampling error formula (equation 4), we set n_1 and n_2 both to n. After squaring and rearranging the terms, this results in equation (6):

$$n = \frac{V(x_1) + V(x_2)}{[s.e.(\bar{x}_2 - \bar{x}_1)]^2} \tag{6}$$

By substituting into this equation the acceptable value of the sampling error and the estimated variances for each of the before and the after situation, we can estimate the required sample size. This is demonstrated under the ‘Practical Application’ later in the paper.

Panel Survey

In the case of a panel survey, we proceed in much the same way. The difference now is that we have the covariance term in the numerator of the expression. We also assume, for a true panel, that the sample size is the same on both occasions. Thus, the sample size required for a panel survey is given by equation (7):

$$n = \frac{V(x_1) + V(x_2) - 2\,\mathrm{cov}(x_1, x_2)}{[s.e.(\bar{x}_2 - \bar{x}_1)]^2} \tag{7}$$

Given that the covariance will always be greater than zero for a panel, the numerator of this equation will always be smaller than the numerator of equation (6). The implication is that the sample size required for a given level of precision in the measurement of the mean difference will always be smaller from a panel than from two independent cross-sectional samples. How much smaller depends on the size of the covariance term. We can understand a little more about this, by noting that the correlation between the two panel measurements is given by equation (8):

$$R_{12} = \frac{\mathrm{cov}(x_1, x_2)}{\sqrt{V(x_1)V(x_2)}} \tag{8}$$

This means that we can replace the covariance term by R multiplied by the product of the standard deviations of the two occasions. In other words, substituting into equation (7), we get equation (9):

$$n = \frac{V(x_1) + V(x_2) - 2R_{12}s_1s_2}{[s.e.(\bar{x}_2 - \bar{x}_1)]^2} \tag{9}$$

Interpreting equation (9), we can see a major determining factor is going to be how much correlation there is between the two waves of the panel. If most respondents' behaviours change very little, we can expect the correlation to be high and sample sizes to be reduced.

A Practical Application

It is easiest to see what all of this means if we apply these formulae to a practical situation. Let us suppose that the issue at hand is to measure the change in VKT from before a Voluntary Travel Behaviour Change initiative to after the initiative. The expectation, we shall assume, is that the change in VKT is about 10 percent. We will also assume that the desire is to know this with an accuracy of about ±1 percent at 95 percent confidence. We will also assume that the average VKT per household is 40 kms per day, and that the standard deviation of this measure is about 1.2 times the mean value. Therefore, the standard deviation before the initiative would be 48 kms per day, the expected mean value after the initiative is 36 kms per day, with a standard deviation of 43 kms per day. Since the change in behaviour is relatively small, we will assume that the correlation between the before and after panel is 0.75. Given the specification of the error level as ±1 percent with 95 percent confidence, it follows that the allowable sampling error is 0.4/1.96 kms or 0.2041 kms.

If it was decided to undertake two independent cross-sectional samples to measure this change, then the required household sample size from equation (6) is given by:

$$n = \frac{V(x_1) + V(x_2)}{[s.e.(\bar{x}_2 - \bar{x}_1)]^2} = \frac{48^2 + 43^2}{(0.2041)^2} = 99{,}714$$

For the same degree of precision, the panel sample of households would be:

$$n = \frac{V(x_1) + V(x_2) - 2R_{12}s_1s_2}{[s.e.(\bar{x}_2 - \bar{x}_1)]^2} = \frac{48^2 + 43^2 - 2*0.75*48*43}{0.2041^2} = 25{,}379$$

In both cases, the sample sizes are extremely large, and would probably require the finite population correction factor to be applied. Suppose, in fact, that the survey was to be done in a community of 30,000 households. The finite population correction factor is *1/(1+n/N)*, which would reduce the cross-sectional sample to 23,062 and the panel to 13,748. Almost certainly, we would conclude that we cannot afford this level of precision.

Table 1 demonstrates the impacts of changing the level of precision required. For a level of precision of ±2 percent with 95 percent confidence, the maximum allowable sampling error doubles from 0.2041 to 0.4082, while the corrected sample size drops to 13,615 for the cross-sectional sample and 5,237 for the panel. These are still very large samples. If we increase the allowable error still further to ±5 percent, the corrected sample size for the cross-section is 3,989 and for the panel, 982, a number which although still quite large, looks more achievable. The problem is we are now saying that our allowable sampling error is 1.0204 kms, which it is critical to think about in the context of the actual magnitude of the change in VKT. If the change is 10 percent, we would only be able to say that the actual change in VKT lay somewhere between 5 and 15 percent with 95 percent confidence. If however, the measured change were only 6 percent, the 95 percent confidence bounds would still be ±5 percent, so we would only know the change in VKT was between 1 and 11 percent. Clearly, this implies we could reach vastly different conclusions about the effectiveness of the intervention.

Table 1 Impacts of Changing Sample Error on Sample Size Requirements

Survey Type	Sample Error	Allowable Sample Error	Correlation (R)	Sample Size	Population	After FPCF
Cross-Section	0.01	0.20408	-----	99,714	30,000	23,062
Panel	0.01	0.20408	0.75	25,379	30,000	13,748
Cross-Section	0.02	0.40816	-----	24,928	30,000	13,615
Panel	0.02	0.40816	0.75	6,345	30,000	5,237
Cross-Section	0.03	0.61224	-----	11,079	30,000	8,091
Panel	0.03	0.61224	0.75	2,820	30,000	2,578
Cross-Section	0.04	0.81633	-----	6,232	30,000	5,160
Panel	0.04	0.81633	0.75	1,586	30,000	1,507
Cross-Section	0.05	1.02041	-----	3,989	30,000	3,520
Panel	0.05	1.02041	0.75	1,015	30,000	982

Assumptions: $\bar{x}_1 = 40$ *km;* $\bar{x}_2 = 36$ *km;* $s_1 = 48$ *km;* $s_2 = 43$ *km;* $\alpha = 0.05$

FPCF = finite population correction factor

This previous point is highly pertinent when it comes to assessing the results of previous studies, many of which have reported substantial reductions in VKT, based on small, cross-sectional samples. In this case, we can calculate the sample error, using equation (4). For instance, say we are only able to sample 500 households using a cross-sectional method,

which indicate a 15% decrease in VKT, we would only be able to state with 95% confidence the actual change in VKT for the population was between 1 and 29 percent!

These sample size calculations have also incorporated assumptions about the correlation between the before and after survey VKT. A correlation of 1 would imply an exact match in VKT for all households in the before and the after surveys, while a correlation of zero implies no match and is synonymous with two independent cross-sectional surveys. In reality, we expect inherent variability in travel between the before and after period for several reasons, some of which we cannot control (e.g., changes in household structure, a new job, obtaining a drivers licence, rising petrol prices), and some of which we can (same day-of-the-week, multi-day sampling periods, length of time between the before and after surveys).

Table 2 illustrates that changes in the correlation have a marked impact on the sample size required to achieve the same level of precision. Clearly it is an assumption that must be investigated and verified further through empirical evidence. It is however, worth stressing that even if the correlation is relatively weak, say 0.6, the corrected sample size is still substantially less (3,892) than for the zero correlation/cross-section (11,079) case at 95 percent confidence. This adds further credence to the belief of the authors that a panel is the preferred method for measuring behavioural change.

Table 2 Impacts of Changing Correlation on Sample Size Requirements

Sample Error	Allowable Sample Error	Correlation (R)	Sample Size	Population	After FPCF
0.03	0.61	0.95	617	30000	605
0.03	0.61	0.90	1,168	30,000	1,124
0.03	0.61	0.75	2,820	30,000	2,578
0.03	0.61	0.50	5,573	30,000	4,700
0.03	0.61	0.25	8,326	30,000	6,517
0.03	0.61	0.00*	11,079	30,000	8,091

Assumptions: $\bar{x}_1 = 40$ *km;* $\bar{x}_2 = 36$ *km;* $s_1 = 48$ *km;* $s_2 = 43$ *km;* $\alpha = 0.05$

FPCF = finite population correction factor

**A correlation of zero is that obtained using two independent cross-sectional samples*

Empirical Testing

The hypothetical example shows the critical importance of the assumptions made about the parameters – means, standard deviations, and particularly the correlation between the before and after scenario. Establishing appropriate values is a non-trivial matter and clearly requires measurements for a panel of participants before and after an intervention has been introduced. At time of writing, because no such data set was available to the authors for assessment of the assumptions locally, use was made of a U.S. panel survey, the Puget Sound Transportation Panel (PSTP) for this phase of the analysis.

The PSTP is an approximately yearly panel survey, begun in 1989, with subsequent waves in 1990, 1992, 1993, 1994, 1996, 1997, as well as more recent waves, which are not yet public issue. Data were recorded from October to January with no collection on Thanksgiving, Christmas and Boxing Days for the first five waves and May to August for the next 2 waves. The format of the survey is a 2-day weekday only travel diary with the same 2 days used in as

near as possible the same time of the year (apart from between the 1994 and 1996 waves). In addition, the sample is stratified according to three 'modal' classifications: 1) those that travel by bus for four or more trips/week, 2) those that travel to work by carpool/vanpool four or more trips/week, and 3) those that do neither of the above (single occupant vehicle – SOV – users).

For the purposes of this analysis, the 1996 wave was selected for the 'before' scenario and 1997 as the 'after'. It was critical to include only those households who had joined in 1996 or earlier – as Table 3 shows, this resulted in the omission of 659 (33 percent) households. It is also interesting to note the year of entry for those households in the 1997 wave. Almost one-quarter have been in from the inception of the panel (what might be termed the 'hard-core' panel members), with a similar percentage entering in the immediately preceding wave. After removing households who had been replaced and those which had only recorded travel for one of the two days, a final sample of 1,318 households was used in the analysis.

Table 3 Year of Entry for the 1997 Wave of the Puget Sound Transportation Panel

Year	Frequency	Percent
1989	476	23.72
1990	69	3.44
1992	94	4.68
1993	136	6.78
1994	127	6.33
1996	446	22.22
1997	659	32.84
Total	2007	100.00

While we can use the PSTP to determine the 'Before' survey VKT mean and standard deviation, we cannot use it to determine directly the 'After' survey mean and standard deviation or the covariance, following an intervention. What we can do, however, is attempt to simulate on a household-by-household case what might happen in the event of an intervention and use this as the basis for determining the three unknown parameters. This clearly needs to account for whether the household takes up the intervention and (for those that do) must separate out changes due to the intervention and those due to external factors. Clearly, we could control this process to determine what the impacts of various assumed changes in VKT might have on the parameters and ultimately the sample size.

The issues outlined above demonstrate this is a non-trivial task and is one that the research team are currently undertaking. For the purposes of the current paper, the approach taken was to simulate an overall reduction of 10% in VKT and determine the covariance based on the actual 1996 and 1997 data. [Note that intuition suggests we might expect the covariance to be reduced following an intervention, which could increase the sample size requirements if this outweighs the corresponding reduction in the variance of VKT in the 'After' case].

Table 4 shows the sample size requirements to detect a 10 percent decrease in VKT for various scenarios using the PSTP data. Notable points include the much higher average and variability of VKT, a reflection of the greater automobile availability and usage in the U.S. than Australia. The correlation might seem on first reflection to be lower than expected.

However, it must be realised, this is based on a survey conducted on two days, one year apart, during which time several factors could have changed that might impact VKT – these were speculated on in the previous section and could be investigated with the PSTP data, something that is planned in future work.

Table 4 Sample Size Requirements to detect a 10 Percent Decrease in VKT (1-day)

Daily (matched to day of week)	Desired Sample Error	$\bar{x}_B$	$\bar{x}_A$	S_B	S_A	Allowable Sample Error	Correlation (R)	Sample Size
Cross-Section	0.02	77.22	69.50	69.94	62.94	0.79	-----	14,258
Cross-Section	0.03	77.22	69.50	69.94	62.94	1.18	-----	6,337
Panel	0.02	77.22	69.50	69.94	62.94	0.79	0.542	6,573
Panel	0.03	77.22	69.50	69.94	62.94	1.18	0.542	2,921

α=0.05

From the perspective of sample sizes, they are clearly large (even for the panel) to detect a change with ±3 percent (7 to 13 percent) with 95 percent confidence. However, it is notable how the sample sizes are substantially reduced when two days as opposed to one day are used as shown in Table 5. This occurs because extending the duration of the survey reduces the relative variability of the mean within the waves and the correlation between the waves, something which is well-known in practice and proven elsewhere (Richardson *et al.*, 2003).

Table 5 Sample Size Requirements to detect a 10 Percent Decrease in VKT (2-day)

2-Day (matched to same 2 days of week)	Required Sample Error	$\bar{x}_B$	$\bar{x}_A$	S_B	S_A	Allowable Sample Error	Correlation (R)	Sample Size
Cross-Section	0.02	153.22	137.90	126.32	113.69	1.56	-----	11,814
Cross-Section	0.03	153.22	137.90	126.32	113.69	2.35	-----	5,251
Panel	0.02	153.22	137.90	126.32	113.69	1.56	0.6747	3,887
Panel	0.03	153.22	137.90	126.32	113.69	2.35	0.6747	1,727

α=0.05

Pursuing this last point further, evidence suggests that sample size requirements for similar levels of precision could be reduced further by extending the survey duration to one week. For instance, referring back to the example from earlier and using Richardson *et al.*'s (2003) figure for the coefficient of variation (CV) for one week of trip reporting (0.37), and a correlation of 0.9 (we would logically expect greater consistency between weeks for the before and after waves), we would only need to sample 250 households to achieve a sample error of ±2 percent at 95 percent confidence. Turning this around, even if we were only able to sample 100 households for a week, we could still achieve a level of precision of ±3.2

percent, which reviewing the sample sizes in Table 1, clearly represents a substantial savings. Of course, multi-day surveys bring additional challenges, but the fact remains that they offer tremendous potential savings in sample size and ultimately costs of these types of surveys.

Measuring a Change in Behaviour – An Alternative Method

An alternative method has sometimes been applied to determine the sample size required for measuring a change in travel behaviour. This method is based on structuring a hypothesis test about an effect in a population, and applying this to the type of situation described in the previous section. This method is based on a consideration of Type I and Type II errors. Type I error is the error of rejecting an hypothesis when it is actually true, and Type II error is the error of accepting an hypothesis when it is actually false.

The method is based on assuming that we wish to know how big a sample is required in order to detect a certain pre-specified size of effect within a population. It is a method that is most commonly found in medical statistics, but which we would argue is not the most appropriate method for the situation at hand. The main problem with the formula used in this method for determining the sample size for a panel is that it has nothing to do with panels, and nothing to do with sampling, *per se*. It also is based on a test for the difference in means. It is also extremely important to keep in mind that the probabilities of Type I and Type II errors are *conditional probabilities*, because, only if the null hypothesis is in fact true can Type I error exist, and only if the alternative hypothesis is true can the Type II error exist. This becomes clearer if we look at a plot of the Type I and Type II errors. An example of such a plot is shown in Figure 1.

It is important to note that the Type I error does not exist for values of π greater than 0.5, while Type II error does not exist for values of π smaller than 0.5, in this example. The value of 0.5 represents the maximum value for the null hypothesis in this case.

Now, Walpole and Myers (1978) suggest that one can determine a sample size for testing a hypothesis that a mean is equal to some hypothesised value, μ_0. Thus if the null hypothesis is that the mean is equal to μ_0 and the alternate hypothesis is that the mean is greater than μ_0, then they show that the sample size required for this hypothesis test is given by:

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{\delta^2}$$

where:

n	=	desired sample size
z_α	=	the normal deviate for a Type I error probability of α
z_β	=	the normal deviate for a Type II error probability of β
σ^2	=	the variance of the measure
δ	=	the minimum difference between μ and μ_0 that would result in rejection of the null hypothesis

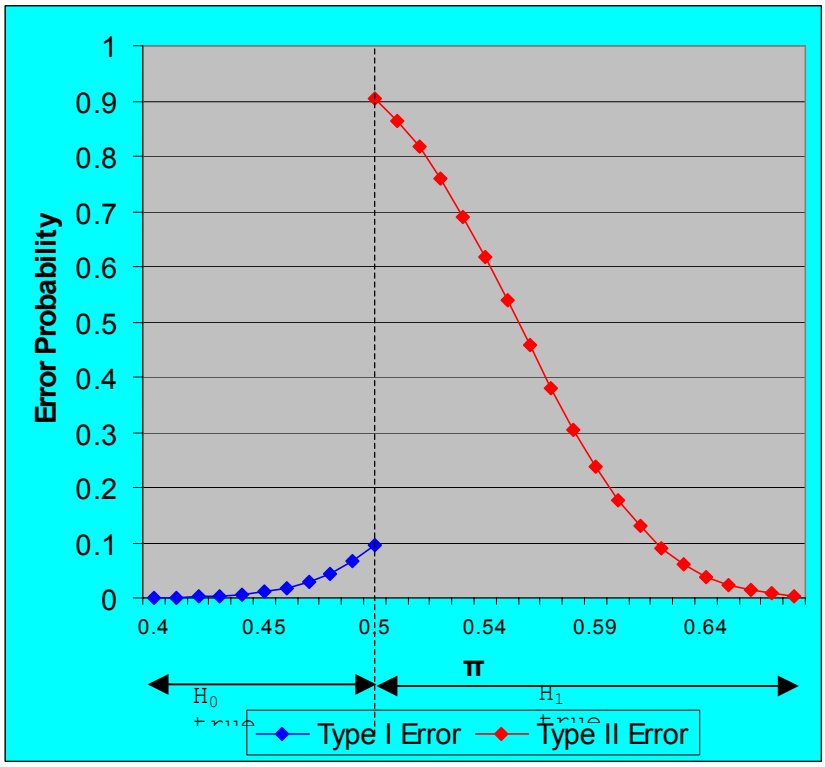

Figure 1: Type I and Type II errors for a hypothetical situation

In other words, the hypothesis that is being tested here is:

H_0: $\mu = \mu_0$
H_1: $\mu = \mu_0 + \delta$

and n is then the sample size needed to determine if the null hypothesis should be rejected with a probability of Type I error of α and a probability of Type II error of β.

Walpole and Myers (1978) then go on to look at another hypothesis test, which is to compare two population means. It is important to note here that the hypothesis test and the resulting calculation of n is correct only if the two populations are independent of one another. In this case, Walpole and Myers show that, if the same sample size is to be used for both populations, then the following is the result. The hypothesis may be stated as:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 = \mu_2 + \delta$

and

$$n = \frac{(z_\alpha + z_\beta)^2 (\sigma_1^2 + \sigma_2^2)}{\delta^2}$$

where the symbols mean the same as before, but σ_1^2 and σ_2^2 are the variances of the two populations. If one were to assume that the two populations had the same variance, then this could be written as:

$$n = \frac{2(z_\alpha + z_\beta)^2 \sigma^2}{\delta^2}$$

This is the equation that Richardson *et al.* (1995) use for determining the sample size of a panel to measure a change in behaviour of the population of an amount of δ. It is also the basis of the computations in Richardson *et al.* (2003).

Immediately, we see some problems here. First, if there has been a change in behaviour, then the variance of the population in the before sample will not be equal to the variance in the after sample, so that the modification of the formula is not appropriate. Second, and much more importantly, in a panel, the two populations are not independent, but are actually strongly dependent on one another. This suggests that this is not the appropriate method to calculate the sample size of a panel. Third, as stated by Walpole and Myers (1978), when the population variance (or variances) is unknown, the choice of sample size is more complex, and these formulas do not apply directly. In application to transport issues, the population variance is unknown, and we have only sample estimates, which require a more complex estimation.

If, for the moment, however, we ignore these problems, and assume that this formula is appropriate to use, then we next have to consider how to define z_α and z_β. First, it is useful to note that we have shown here, and it is a statistical fact, that α, β, and n are all interrelated. As stated by Ewart *et al.* (1982),

> "Because of the relationships among α, β, and n, the specification of any two of these three quantities automatically determines the third. Thus, it is theoretically and mechanically possible to derive the decision rule by specifying either (1) the significance level [(α)] and sample size, or (2) the significance level and power of the test [$(1-\beta)$], or (3) the sample size and the power of the test. However, because of practical difficulties in specifying a desired power for the test, only the first of these three possible approaches is used to any great extent in actual practice." (pages 264-265.)

Clearly, in developing the method of Walpole and Myers, we see a case where the attempt is made to specify the second case of Ewart *et al.* (1982), in which the significance level and the power of the test are specified, and the sample size is then determined from it. The real problem with this is in knowing how to choose the power of the test. Walpole and Myers suggest this is a simple matter, and choose to suggest a significance level of 0.05 and a power of the test of 0.95, under which conditions, for a one-tailed test, $z_\alpha = z_\beta = 1.645$, and the formula can be solved for n. This appears to be rather too simple. In the first place, as one decreases the probability of a Type I error, the probability of a Type II error must increase. Second, there are an extremely large number of possible values of Type II error for a given test situation (hypothesis), given that we have specified the acceptable level of Type I error. The only way to fix the Type II error is to restrict the alternative hypothesis to a point value. Usually, we fix the Type I error as the maximum Type I error that can occur under the null hypothesis. However, we cannot do the same for Type II error.

In the method outlined by Walpole and Myers (1978), they fix the Type II error by recasting the alternative hypothesis to a point value. This means that, if we have evidence that leads to rejecting the null hypothesis, we have to accept the alternative hypothesis. In most cases, this is a very restrictive test, and one that would not be considered appropriate. However, there are circumstances (albeit restricted ones) where the use of point hypotheses are potentially of value. For example, suppose a pig farmer is considering introducing a new feed, that has been promoted as resulting in more rapid weight gain. The farmer decides to divide his pigs into two groups, one of which will receive the new feed, while the other continues to have the old feed. He decides further that he will adopt the new feed only if the pigs gain 500 grams more on the new feed in the space of two weeks than on the old feed. Therefore, his hypotheses are:

H_0: There is no change in the weight of the pigs
H_1: There is a change of 500 grams in the weight of the pigs

Therefore, if the farmer has evidence to reject the null hypothesis of no weight change, he accepts the alternative hypothesis and adopts the new feed. If he does not have evidence to reject the null hypothesis, then he stays with the old feed.

In the case of applying this method to travel behaviour change, if all other assumptions were valid, then it would be useful if we knew what was the minimum level of change that would make it worthwhile to pursue implementation of travel behaviour change on a wider basis. If that was then specified as the value δ, the sample size to establish whether or not that amount of change had occurred would be useful. However, we would argue that the number, in that case, is probably much smaller than 10 percent, because it has been shown that behaviour changes as large as 10 percent lead to benefit-cost ratios that are very large. Possibly, the appropriate value of δ might be closer to 1 or 2 percent. In that case, the sample sizes necessary to detect such a behaviour change would be about 25 times as big as have previously been estimated, for an infinite or very large population, and would probably approach the population size, when populations are much smaller.

By the nature of sampling statistics, it is true that, if we have drawn a sample large enough to reject an hypothesis of 0 percent change, and our alternative hypothesis is that the change is 10 percent, then the sample size is also necessarily large enough to work for any change larger than 10 percent. However, it will not be a sufficient size to permit testing the alternative hypothesis of less than a 10 percent change.

In summary, there are a number of reasons for rejecting the use of the formula, based on setting the levels of both Type I and Type II error. These are:

a) The formula is designed to determine the size of sample needed to determine whether or not a specific hypothesis of change can be rejected;
b) The formula is based on an assumption of independence between the before and the after sample, and is not applicable to a panel;
c) The formula assumes that the variances before the behaviour change and after it are the same;
d) The formula assumes that the population variances are known (not sample estimates only);
e) The formula allows testing two point hypotheses; and
f) The formula ignores sampling error and its effect on sample size determination.

Hence, we recommend that this formula is not used to estimate sample size requirements for a panel of for travel behaviour change measurement, especially when it is not certain what the minimum level of change is that would be required, or where the level of change is largely unknown.

Conclusions/recommendations

This paper marks an important milestone in the issue of sample size requirements for measuring changes in (travel) behaviour. An approach that has been used rather widely to date is questioned and an alternative formulation proposed, which is built on solid foundations and classical statistical theory. The formulation enables designers of behavioural change surveys to investigate the trade-offs between the desired precision of their predictions, the levels of confidence required in results, and the sample size (and costs). In addition, it provides a statistically-robust method for assessing the results of previous behavioural change programs.

Through the hypothetical and empirical examples, which focused on the issue of predicting an *x%* decrease in VKT, several issues emerged. First (perhaps fundamentally), large sample sizes are required (several thousand households) to detect a pre-specified change in behaviour if high (or even reasonable) levels of precision are required in results. Second, sample sizes are reduced using a panel as opposed to two independent cross-sectional samples. The extent of this reduction depends on the strength of the correlation between the before and after measurement, which is weakened by any major changes in the household that potentially impact VKT (e.g., new car, change in occupation, new baby, etc.) as well as the inherent variability in VKT. The longer the time between waves, the more these factors will negatively impact the correlation, something that must be considered carefully in sample design. Third, sample sizes can also be reduced substantially by increasing the duration of the survey period to two days or (preferably) for a longer period such as a week.

As Richardson *et al.*, (2003) note, while a weekly panel is the 'optimum' in terms of sample size requirements, one has to be aware of the increased difficulties of implementing such a survey to ensure unbiased results. Seemingly new technologies such as global positioning systems (GPS) may offer the potential to overcome some of these issues by automating the collection of those data, which are prone to unreliable reporting (e.g., locations, distances, travel times). If one balances the costs of the technology against the savings in sample size requirements, data processing, and data accuracy/quality, this could seemingly offer an increasingly viable method for inclusion within such surveys in the future.

References

Benjamin, J.R. and C.A. Cornell (1970) *Probability, Statistics, and Decision for Civil Engineers*, McGraw-Hill Book Co., New York.

Ewart, P.J., J.S. Ford, and C-Y Lin (1982) *Applied Managerial Statistics*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Kish, L (1965) *Survey Sampling,* John Wiley and Sons.

Richardson, A.J., R.K. Seethaler and P.L. Harbutt (2003) Design Issues for Before and After Surveys of Travel Behaviour Change, *Proceedings of* the *26th Australasian Transport Research Forum*, Wellington, New Zealand.

Richardson, A.J., E.S. Ampt, and A.H. Meyburg (1995) *Survey Methods for Transport Planning*, Eucalyptus Press.

Stopher, P.R. and A.H. Meyburg (1976) *Urban Transportation Planning and Modelling,* Lexington Books, D.C. Heath and Co., Lexington, MA.

Stopher, P.R. (2004) **"**Voluntary Travel Behaviour Change" in D. Hensher and K. Button (editors), *Handbook on Transport Policy*, Elsevier (Handbook No. 6), (in press).

Walpole, R.E. and R.H. Myers (1978) *Probability and Statistics for Engineers and Scientists*, 2nd Edition, MacMillan Publishing Co., New York.

Weatherburn, C.E. (1962) *A First Course in Mathematical Statistics*, Cambridge University Press.