Sampling and Samples

Contributed by Joanne Birchall

Sampling and Samples written by Joanne Birchall from Rainbow ResearchUnless you are in the luxurious position of having access to everyone who forms your population, you will need to take some form of sample from which to glean information for Market Research purposes. In addition to accessibility, the method chosen will depend upon a variety of statistical and practical factors. You will want to ensure your sample size is sufficient for the purpose of the analysis you intend to perform, ensure your sample is representative of the population you are attempting to say something about, and of course you will need to take into account your affordability.

This section covers the following:
- Sampling methods
- Calculating a sample size
- Calculating a sampling error

Sampling Methods

In most surveys, access to the entire population is near on impossible, however, the results from a survey with a carefully selected sample will reflect extremely closely those that would have been obtained had the population provided the data.

Sampling therefore is a very important part of the Market Research process. If you have surveyed using an appropriate sampling technique, you can be confident that your results will be generalised to the population in question. If the sample were biased in any way, for example, if the selection technique gave older people more of a chance of selection than younger people, it would be inadvisable to make generalisations from the findings.

There are essentiality two types of sampling: probability and non-probability sampling.

Probability Sampling Methods

Probability or random sampling gives all members of the population a known chance of being selected for inclusion in the sample and this does not depend upon previous events in the selection process. In other words, the selection of individuals does not affect the chance of anyone else in the population being selected.

Many statistical techniques assume that a sample was selected on a random basis. There are four basic types of random sampling techniques:

1) Simple Random Sampling

This is the ideal choice as it is a ‘perfect’ random method. Using this method, individuals are randomly selected from a list of the population and every single individual has an equal chance of selection.

This method is ideal, but if it cannot be adopted, one of the following alternatives may be chosen if any shortfall in accuracy.

2) Systematic Sampling

Systematic sampling is a frequently used variant of simple random sampling. When performing systematic sampling, every kth element from the list is selected (this is referred to as the sample interval) from a randomly selected starting point. For example, if we have a listed population of 6000 members and wish to draw a sample of 2000, we would select every 30th (6000 divided by 200) person from the list. In practice, we would randomly select a number between 1 and 30 to act as our starting point.

The one potential problem with this method of sampling concerns the arrangement of elements in the list.? If the list is arranged in any kind of order e.g. if every 30th house is smaller than the others from which the sample is being recruited, there is a possibility that the sample produced could be seriously biased.

3) Stratified Sampling

Stratified sampling is a variant on simple random and systematic methods and is used when there are a number of distinct subgroups, within each of which it is required that there is full representation. A stratified sample is constructed by classifying the population in sub-populations (or strata), base on some well-known characteristics of the population, such as age, gender or socio-economic status. The selection of elements is then made separately from within each strata, usually by random or systematic sampling methods.

Stratified sampling methods also come in two types – proportionate and disproportionate.

In proportionate sampling, the strata sample sizes are made proportional to the strata population sizes.For example if the first strata is made up of males, then as there are around 50% of males in the UK population, the male strata will need to represent around 50% of the total sample.

In disproportionate methods, the strata are not sampled according to the population sizes, but higher proportions are selected from some groups and not others. This technique is typically used in a number of distinct situations:

The costs of collecting data may differ from subgroup to subgroup.

We might require more cases in some groups if estimations of populations values are likely to be harder to make i.e. the

larger the sample size (up to certain limits), the more accurate any estimations are likely to be.
We expect different response rates from different groups of people. Therefore, the less co-operative groups might be ‘over-sampled’ to compensate.

{mospagebreak title=Sampling Methods}
4) Cluster or Multi-stage Sampling
Cluster sampling is a frequently-used, and usually more practical, random sampling method. It is particularly useful in situations for which no list of the elements within a population is available and therefore cannot be selected directly. As this form of sampling is conducted by randomly selecting subgroups of the population, possibly in several stages, it should produce results equivalent to a simple random sample.
The sample is generally done by first sampling at the higher level(s) e.g. randomly sampled countries, then sampling from subsequent levels in turn e.g. within the selected countries sample counties, then within these postcodes, the within these households, until the final stage is reached, at which point the sampling is done in a simple random manner e.g. sampling people within the selected households. The ‘levels’ in question are defined by subgroups into which it is appropriate to subdivide your population.

Cluster samples are generally used if:

- No list of the population exists.
- Well-defined clusters, which will often be geographic areas exist.
- A reasonable estimate of the number of elements in each level of clustering can be made.
- Often the total sample size must be fairly large to enable cluster sampling to be used effectively.
Non-probability Sampling Methods
Non-probability sampling procedures are much less desirable, as they will almost certainly contain sampling biases. Unfortunately, in some circumstances such methods are unavoidable.
In a Market Research context, the most frequently-adopted form of non-probability sampling is known as quota sampling.? In some ways this is similar to cluster sampling in that it requires the definition of key subgroups. The main difference lies in the fact that quotas (i.e. the amount of people to be surveyed) within subgroups are set beforehand (e.g. 25% 16-24 yr olds, 30% 25-34 yr olds, 20% 35-55 yr olds, and 25% 56+ yr olds) usually proportions are set to match known population distributions. Interviewers then select respondents according to these criteria rather than at random. The subjective nature of this selection means that only about a proportion of the population has a chance of being selected in a typical quota sampling strategy.

If you are forced into using a non-random method, you must be extremely careful when drawing conclusions. You should always be honest about the sampling technique used and that a non-random approach will probably mean that biases are present within the data. In order to convert the sample to be representative of the true population, you may want to use weighting techniques.

The importance of sampling should not be underestimated, as it determines to whom the results of your research will be applicable. It is important, therefore to give full consideration to the sampling strategy to be used and to select the most appropriate. Your most important consideration should be whether you could adopt a simple random sample.? If not, could one of the other random methods be used? Only when you have no choice should a non-random method be used.

All to often, researchers succumb to the temptation of generalising their results to a much broader range of people than those from whom the data was originally gathered. This is poor practice and you should always aim to adopt an appropriate sampling technique. The key is not to guess, but take some advice.?

{mospagebreak title=Calculating Sample Size}
Calculating a Sample Size
A frequently asked question is “How many people should I sample?” It is an extremely good question, although unfortunately there is no single answer! In general, the larger the sample size, the more closely your sample data will match that from the population. However in practice, you need to work out how many responses will give you sufficient precision at an affordable cost.

Calculation of an appropriate sample size depends upon a number of factors unique to each survey and it is down to you to make the decision regarding these factors. The three most important are:
- How accurate you wish to be
- How confident you are in the results
- What budget you have available
The temptation is to say all should be as high as possible. The problem is that an increase in either accuracy or confidence (or both) will always require a larger sample and higher budget. Therefore a compromise must be reached and you must work out the degree of inaccuracy and confidence you are prepared to accept.
There are two types of figures that you may wish to estimate in your Market Research project: values such as mean income, mean height etc. and proportions (the percentage of people who intend to vote for party X). There are slightly different sample size calculations for each:

For a meanThe required formula is: s = (z / e)2

Where:
s = the sample size
z = a number relating to the degree of confidence you wish to have in the result. 95% confidence* is most frequently used and accepted. The value of ‘z’ should be 2.58 for 99% confidence, 1.96 for 95% confidence, 1.64 for 90% confidence and 1.28 for 80% confidence.
e = the error you are prepared to accept, measured as a proportion of the standard deviation (accuracy)

For example, imagine we are estimating mean income, and wish to know what sample size to aim for in order that we can be 95% confident in the result. Assuming that we are prepared to accept an error of 10% of the population standard deviation (previous research might have shown the standard deviation of income to be 8000 and we might be prepared to accept an error of 800 (10%)), we would do the following calculation:

s = (1.96 / 0.1)2
Therefore s = 384.16
In other words, 385 people would need to be sampled to meet our criterion.
*Because we interviewed a sample and not the whole population (if we had done this we could be 100% confident in our results), we have to be prepared to be less confident and because we based our sample size calculation on the 95% confidence level, we can be confident that amongst the whole population there is a 95% chance that the mean is inside our acceptable error limit. There is of course a 5% chance that the measure is outside this limit. If we wanted to be more confident, we would base our sample size calculation on a 99% confidence level and if we were prepared to accept a lower level of confidence, we would base our calculation on the 90% confidence level.`
For a proportion
Although we are doing the same thing here, the formula is different:

s = z2(p(1-p))
???????? e2
Where:
s = the sample size
z = the number relating to the degree of confidence you wish to have in the result
p = an estimate of the proportion of people falling into the group in which you are interested in the population
e = the proportion of error we are prepared to accept
As an example, imagine we are attempting to assess the percentage of voters who will vote for candidate X. If we assume that we wish to be 99% confident of the result i.e. z = 2.85 and that we will allow for errors in the region of +/-3% i.e. e = 0.03. But in terms of an estimate of the proportion of the population who would vote for the candidate (p), if a previous survey had been carried out, we could use the percentage from that survey as an estimate. However, if this were the first survey, we would assume that 50% (i.e. p = 0.05) of people would vote for candidate X and 50% would not. Choosing 50% will provide the most conservative estimate of sample size. If the true percentage were 10%, we will still have an accurate estimate; we will simply have sampled more people than was absolutely necessary. The reverse situation, not having enough data to make reliable estimates, is much less desirable.

In the example:

s = 2.582(0.5*0.5)
???????? 0.032
Therefore s = 1,849

This rather large sample was necessary because we wanted to be 99% sure of the result and desired and desired a very narrow (+/-3%) margin of error. It does, however reveal why many political polls tend to interview between 1,000 and 2,000 people.
{mospagebreak title=Calculating a Sampling Error}
Calculating a Sampling Error
In estimating the accuracy of a sample (sampling error), or selecting a sample to meet a required level of accuracy, there are two critical variables; the size of the sample and the measure being taken which for simplicity we shall take as a single percentage e.g. the percentage aware of a brand. A common mistake about sample size is to assume that accuracy is determined by the proportion of a population included in a sample (e.g. 10% of a population). Assuming a large population, this is not the case and what matters is the absolute size of the sample regardless of the size of the population – a sample of 500 drawn from a population of one million will be as accurate as a sample of 500 from a population of five million (assuming both are truly random samples of the respective populations).

The relationship between sampling error, a percentage measure and a sample size can be expressed as a formula.

e = z√(p%(100-p%))

?????????????? √ s

Where:
e = sampling error (the proportion of error we are prepared to accept)
s = the sample size
z = the number relating to the degree of confidence you wish to have in the result
p = an estimate of the proportion of people falling into the group in which you are interested in the population
By applying the formula it can be calculated, for example, that from a sample of 500 respondents (s), a measure of 20% aware of a brand (p), will have a sample error of +/-3.5% at the 95% confidence level.

e = 1.96√(20(80))
??????????? √ 500
This means, therefore, that based on a sample of 500 we can be 95% sure that the true measure (e.g. of brand awareness) among the whole population from which the sample was drawn will be within +/-3.5% of 20% i.e. between 16.5% and 23.5%.
If you are put off by these calculations, help is at hand.
Please do not hesitate to give Rainbow Research a call on +44 (0) 1772 743235. Follow Market Research World on Twitter or join in the conversation with our LinkedIn Group