

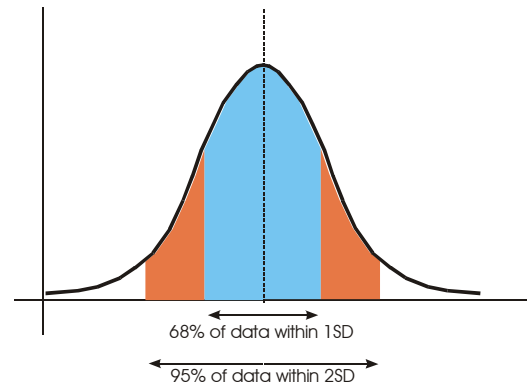
USING SAMPLES TO GENERATE COST PROJECTIONS

1. Introduction

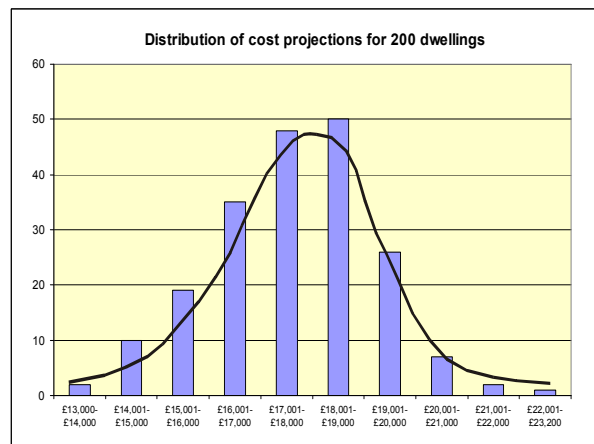
An effective property strategy cannot be developed without an assessment of future elemental repairs and renewals. This assessment is usually calculated from data collected during a stock condition survey. Large samples, indeed 100% ones, may be necessary to develop programmed plans of maintenance or to collect detailed information on individual dwellings. However, where the primary aim is to establish a 20 or 30 year cost projection of repairs for business planning purpose etc. samples can be quite small. This paper explains how sample size can be calculated and includes a number of calculations using real data. A later paper will examine how sampling can be used to assess specific attributes, eg, how many dwellings have hard wired smoke detectors?

2. Normal Distribution

As far as statistics is concerned we have been blessed with one amazing stroke of luck: the normal distribution. Life is not as random as it might seem because many sets of data, whether it be the height of adult men and women, their shoe sizes, or the blood pressure in healthy people, all conform to what is known as a normal distribution. Furthermore, we can be fairly confident that 68% of the population will be within 1 standard deviation of the mean, and 95% within 2 standard deviations (it's actually 1.96). Calculating the standard deviation is explained in more detail in Section 3.



If condition survey work follows the norm most sets of data should give a normal pattern of distribution. If we plot the data used in Section 4 (30 year cost projections for repairing 200 new-build houses) the pattern is familiar. In this data set the lowest value is £13,400, the highest £22,650 and the mean is £17,633. The standard deviation is 1,597 (see Section 3).



We know that in a normal distribution 68% of the sample will be within 1 standard deviation of the mean. The mean is £17,633; therefore 68% of the data will lie within the mean -1597 (the sd) and the mean +1597. 95% of the data will lie in the range £14,439 and £20,827(see the table below).

Data items	200
Lowest	13,400
Highest	22,650
Mean	17,633
SD	1,597

		From	To		
All data		13,400	22,650	All data lies in this range	
+ or - 1 SD	mean - (1 x sd)	16,036	mean + (1 x sd)	19,230	68% of data should be in this range
+ or - 2SD	mean - (2 x sd)	14,439	mean + (2 x sd)	20,827	95 % of data should be in this range

3. Sampling

Statistical theory can help to assess the likely errors involved in using a sample of data rather than a complete data set. This involves thinking about a particular sample as one of a large number of possible samples that might be selected and assessing how different the results could be across this range of samples. If a large number of samples of the same size were selected at *random* from the stock, and the estimated mean repair cost of the stock obtained from each sample were plotted on a graph, the results would give the distinctive bell shaped curve: the *normal distribution*.

Most samples would give a mean near the centre of the distribution. A few samples would give results that were some way away from the centre.

There is another reassuring conclusion from statistical theory: the larger the sample, the narrower is the spread or variation in the sampling distribution. In housing the required sample size will depend on the nature, age and condition of the stock.

Standard deviation tells how tightly a set of values is clustered around the (mean) average of those same values. It's a measure of dispersal, or variation, in a group of numbers. The size of the required sample is partly determined by the standard deviation.

Consider the following examples.

30 year Cost Projection for Elemental Renewals			
H1	10,000	10,000	10,000
H2	11,000	10,000	15,000
H3	12,000	10,000	15,000
H4	13,000	10,000	15,000
H5	14,000	10,000	15,000
H6	16,000	20,000	15,000
H7	17,000	20,000	15,000
H8	18,000	20,000	15,000
H9	19,000	20,000	15,000
H10	20,000	20,000	20,000
Mean	15,000	15,000	15,000
SD	3317	5000	2236

The table shows the 30 year cost projection for three sets of data. Each set has a mean of £15,00 and a range of £10,00 to 20,000. The first one has an even spread across the range and has a standard deviation of 3,317. The second one has 5 dwellings at £10,000 and 5 dwellings at £20,000; hence the SD is larger at 5,000. The third set of data has a SD of 2,236. This is the smallest SD because most of the costs are the same as the mean.

4. Worked examples

Consider the cost profile below. This shows a 30 year cost projection (at current prices) for the main building elements of a 3 bed house built in 1996.

House - 1996											
	: 0 :	: 1 :	: 2 :	: 3 :	: 4 :	: 5 :	: 6-10 :	: 11-15 :	: 16-20 :	: 21-25 :	: 26-30 :
Gutters/RW pipes	0	0	0	0	0	0	0	450	0	0	0
Windows	0	0	0	0	0	0	0	0	2750	0	0
Entrance doors	0	0	0	0	0	0	0	0	390	0	0
Fencing/gates	0	0	0	0	0	0	0	1500	0	0	1500
Internal doors	0	0	0	0	0	0	200	0	0	200	0
Kitchen units	0	0	0	0	0	0	1750	0	0	1750	0
Bathroom fittings	0	0	0	0	0	0	0	1000	0	0	0
2nd WC	0	0	0	0	0	0	0	0	400	0	0
Plumbing/hot water	0	0	0	0	0	0	0	0	400	0	0
Heating	0	0	0	0	1200	0	0	0	1200	1500	1200
Wiring	0	0	0	0	0	0	0	0	0	2250	0
Totals	£0	£0	£0	£0	£1,200	£0	£1,950	£2,950	£5,140	£5,700	£2,700
Report total	£19,640										

It is likely that most houses of similar size and construction, and built, say from 1990 to 2000, will have similar cost profiles. Some 'real' data helps illustrate this point. As part of a detailed asset management strategy a housing association, let's call it Harley Housing, has surveyed all of its 290 dwellings. An assessment of repair costs over the next 30 years formed part of the survey.

200 of the association's units were built in the 1990s. These are mostly 2, 3 and 4 bed semi-detached or terraced houses. The lowest cost projection was £13,400 and the highest £22,650 (all the surveys were carried out by the same surveyor - using strict rules to ensure consistency). The mean of the 200 projections was £17,633 and the standard deviation (easily calculated through Excel or any other spreadsheet) was 1597.

If the purpose of the survey was solely to provide a 30 year cost projection, how many units would have to be included in the sample? This depends, of course, on the level of accuracy required. If we accept that in 95 out of 100 samples (95% confidence level) the mean of the sample should be within 5 % (confidence interval) of the actual mean (17,633) the formula is:

$$\text{No. surveys} = ((\text{sd})/(\text{confidence interval}/1.96))^2$$

$$\text{No. surveys} = ((1597)/(881/1.96))^2$$

$$\text{No. surveys} = 13 \text{ (actually 12.6)}$$

This is less than 7% of the stock

In other words if we select 13 out of the 200 properties 100 times we can be confident that in 95 cases the mean of each sample will be £17,633+ or - £881.

The calculation can be repeated with other confidence intervals. For example, if we wanted a confidence interval of 2.5%, in other words the mean of each sample should be £17,633 + or - £440, we would have to survey 4 times as many buildings - just over 50.

Harley Housing also owns and manages 90 acquired (in other words older) dwellings. The cost profile for one of these is shown below.

House - 1930											
	: 0 :	: 1 :	: 2 :	: 3 :	: 4 :	: 5 :	: 6-10 :	: 11-15 :	: 16-20 :	: 21-25 :	: 26-30 :
Coverings	0	0	0	0	5000	0	0	0	0	0	0
Verges/parapets	0	0	0	0	500	0	0	0	0	0	0
Fascias, bargeboards	0	650	0	0	0	0	0	0	0	0	0
Gutters/RW pipes	0	0	0	0	500	0	0	0	0	500	0
Walls structure	0	3000	0	0	0	0	0	0	0	0	0
Walls pointing	0	175	0	0	0	0	0	0	0	0	0
Windows	0	0	0	0	0	0	2750	0	0	0	0
Entrance doors	0	0	0	0	0	1300	0	0	0	0	1300
Soil vent pipes	0	300	0	0	0	0	0	0	0	0	300
Fencing/gates	0	0	0	0	0	0	0	500	0	0	0
Paths/pavings	0	0	0	0	0	0	500	0	0	0	0
Ext lighting	0	0	0	0	0	200	0	0	200	0	0
Drainage	500	0	0	0	0	0	0	0	0	0	0
Floor finish	0	400	0	0	0	0	0	400	0	400	0
Internal walls/finish	0	500	0	0	0	0	0	0	0	0	0
Internal doors	0	300	0	0	0	0	0	0	0	0	0
Kitchen units	0	1750	0	0	0	0	0	1750	0	0	0
Bathroom fittings	0	1000	0	0	0	0	0	0	1000	0	0
Plumbing/hotwater	0	0	0	0	0	0	0	0	400	0	0
Heating	0	0	0	0	0	2850	0	250	1250	0	250
Wiring	0	1500	0	0	0	0	0	0	0	0	0
Totals	£500	£9,575	£0	£0	£6,000	£4,350	£3,250	£1,150	£3,600	£1,900	£1,850
Report total	£32,175										

These properties date from 1880 to 1970; they are individually purchased dwellings of various sizes, and in various states of repair. Data analysis shows that the mean cost projection of these 90 properties is £29,951, the lowest cost projection is £17,650 and the highest £52,575. The standard deviation is 6699.

If we require a confidence level of 95% and a confidence interval of 5% then the number of surveys will be:

$$\text{No. surveys} = ((6699)/(1498/1.96))^2$$

$$\text{No. surveys} = 77.$$

This is over 70 % of the stock.

So we need to undertake 13 surveys of the new- build dwellings and 77 of the acquired ones.

If we had not divided the stock into its two discrete types even more surveys would have been required. Analysis of all the data shows that the mean of the 290 dwellings is £21,456 and the standard deviation 6940. In this case the number of surveys would be:

$$\text{No. surveys} = ((6940)/(1072/1.96))^2$$

$$\text{No. surveys} = 161.$$

The advantages of stratification are obvious.

5. Practical Implications

All the above examples are based on the assumption that correct standard deviations are available for the stock. This is not usually the case - that after all is the purpose of the survey! There are two ways round this; one is to carry out a series of pilot surveys to calculate standard deviations. Another is to examine the relationship between the standard deviation and the mean, and to relate this to specific property types. For example, Harley Housing's new

build stock has a mean of £17,633 and a standard deviation of 1597. The standard deviation divided by the mean is 0.09. The comparable figure for the acquired stock is 0.22 and for all the stock 0.32. By re-jigging the formula above we can produce required sample sizes using the standard deviation divided by the mean and replacing the confidence interval with a % rather than an exact value.

The original formula for the new build stock was:

$$\text{No. surveys} = ((\text{sd})/(\text{confidence interval}/1.96))^2$$

$$\text{No. surveys} = ((1597)/(881/1.96))^2$$

$$\text{No. surveys} = 13 \text{ (actually 12.6)}$$

If it's replaced with this formula the answer is the same even though we have not specified a V monetary value for the confidence interval or a specific figure for the standard deviation.

$$\text{No. surveys} = ((\text{sd}/\text{mean})/(\% \text{ confidence interval}/1.96))^2$$

$$\text{No. surveys} = ((0.09)/(0.05/1.96))^2$$

$$\text{No. surveys} = ((0.09)/(0.0255))^2$$

$$\text{No. surveys} = 13$$

Based on the extensive data we have collected over the years we have produce some typical values for four types of dwelling. These are shown below.



Purpose built dwellings, post 1985 - 0.1 to 0.2



Purpose built older dwellings - 0.15 to 0.25



**Acquired two storey dwellings
Late Victorian to- 1930s - 0.25 to 0.35**



Georgian, Victorian, 3 or more storeys - 0.35 to 0.50

Using this approach we can construct a simple table which shows the number of surveys required for properties of various types and for a given confidence interval (% either side of the mean cost projection).

Sd/mean	Confidence Interval		
	2.5	5	10
0.1	62	16	4
0.2	246	62	16
0.3	554	139	35
0.4	985	246	62
0.5	1537	385	96

Notice that as the sd/mean doubles, or if the confidence interval is halved the number of surveys increase four fold.

6. Allowing for the size of the population

In many examples of survey work, the population from which the sample is drawn is very large and the size of the sample is small in relation to this population. In these cases, the standard error is most significantly influenced by the size of the sample. However, in some cases where Housing Associations are assessing the characteristics of their stock, the size of the stock is small from a few hundred up to, perhaps, a few thousand dwellings and a sample might easily represent 10% or more of the stock.

In general, as the sample size becomes a larger proportion of the population, the sampling error decreases. The 'large sample correction' formula (sometimes also known as the finite population correction factor) is:

$$\sqrt{(1-(n/N))}$$

If the stock of dwellings owned by an association – the population (N) - is 2000 dwellings, and the sample surveyed (n) is 100 dwellings, then the large sample correction factor becomes:

$$\sqrt{(1-(100/2000))} = \sqrt{(1-0.05)} = 0.97$$

The size of the sample required (97 as opposed to 100) is thus not reduced by very much in this case. However, if the sample surveyed was 1000 dwellings, the large sample correction factor begins to reduce the sample size by a significant amount:

$$\sqrt{(1-(1000/2000))} = \sqrt{(1-0.5)} = 0.71$$

Obviously as the sample size approaches the size of the stock, the correction factor becomes almost insignificant.

In the case of Harley Housing the total stock (N) is only 290 units. Thus the required sample size is reduced by the large sample correction factor:

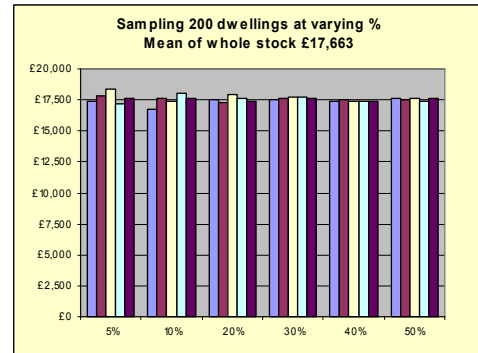
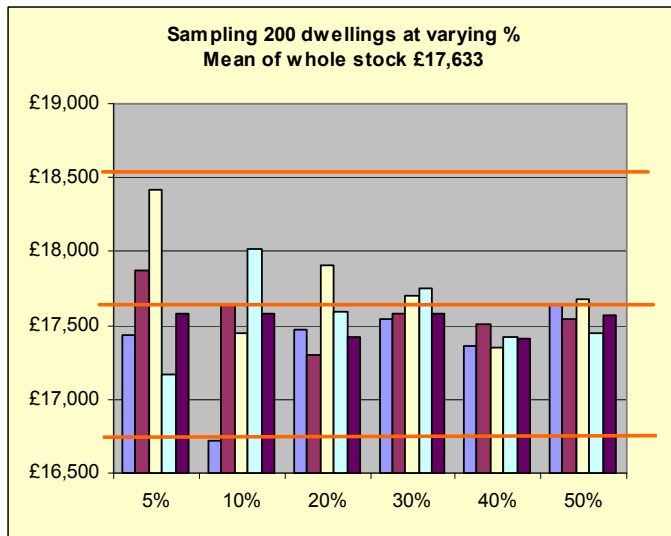
$$\sqrt{(1-(161/290))} = \sqrt{(1-0.55)} = 0.67$$

$$n = 161 * .67 = 107$$

When we did the calculation for the required sample size without taking account of the population size of the stock as a whole, we calculated that we would need a sample of 161 dwellings. When we take account of the fact that the size of the stock is only 290 units, the sample can be reduced to 107 units.

7. Looking at the data another way

To test this sampling theory, and to examine a more common-sense approach we randomly sampled the new build data from Harley Housing. Using Excel we sampled the 200 dwellings 5 times using various percentages. The chart below shows the results. The orange lines represent the mean (£17,633) and the mean + or - 5%. The statistical theory shows that we needed a sample of 13. Our first sample was 10 (5% of the 200 units). You can see that each of the samples falls within the mean, + or - 5%. As the samples get bigger (10%, 20% etc.) so the variance reduces.



The chart on the right shows the complete 'Y' axis – this shows how close all the samples are.

It is clear that the larger samples are more consistent although all except one of the samples are within the mean, + or - 5%. The table below shows the actual figures on which the chart is based. Notice that, for each % sample, the mean of the 5 samples is pretty close to the actual sample, £17,639.

% sample	S1	S2	S3	S4	S5	Mean
5%	£17,433	£17,871	£18,415	£17,170	£17,580	£17,694
10%	£16,718	£17,645	£17,445	£18,020	£17,584	£17,482
20%	£17,471	£17,301	£17,909	£17,596	£17,424	£17,540
30%	£17,546	£17,581	£17,701	£17,745	£17,576	£17,630
40%	£17,359	£17,512	£17,353	£17,426	£17,416	£17,413
50%	£17,640	£17,546	£17,674	£17,448	£17,572	£17,576

8. Survey Errors.

Our experience in validating surveys is that, in many cases, an unnecessary number of surveys are carried out, and that cost extrapolations based on these surveys are not usually incorrect because of sampling errors; they are wrong because of mistakes made by the surveyors on-site.

A number of typical examples will be added to this article when time permits.

9. Sampling for Attributes

If we want to measure the proportion of dwellings with a given attribute, say 2nd WCs, because a measure of stock upgrading is thought to be necessary, then we need to use a different formula from the one we have been using. We ignore the large sample correction factor initially:

$$n = (p*(1-p)/(confidence\ interval/confidence\ level))^2$$

p = the proportion of the stock with the attribute

1-p = the proportion without that attribute

n = the sample size

confidence interval = ie say 3% or 5%

confidence level = usually 95%

What size of sample should be selected?

ZED Housing Association has a large stock of units and wants to commission a survey from consultants to assess the proportion which have 2nd WCs. If it wants the 95% CI to be + or – 3% and confidence level of 95% then all we need to do now is to make an educated guess about the proportion of the stock with 2nd WCs. Managers reckon this may be around 10%. This is a low proportion of the stock, and it means that in accepting the CI of + or – 3%, the managers will have to accept quite a large margin of error in the results. Having conducted the survey, if the sample estimate of the percentage of dwellings with a 2nd WC turns out to be 10% then the 95% CI will be from 7% to 13%.

We can now plug these values into the re-arranged formula to calculate the sample size required:

$$n = (p*(1-p))/(confidence\ interval/confidence\ level)^2$$

$$n = 0.1 * 0.9/(3\%/1.96)^2$$

$$n = 0.09/(0.015306)^2$$

$$n = 0.09/0.000234$$

$$n = 384$$

The consequence of asking for a more demanding level of accuracy is to increase the size of sample required quite substantially. For example, if we want a confidence interval of 1.5 % (not 3%) we would have to survey 4 times as many dwellings.

$$n = 1536$$

If we are not sure about the % of dwellings that are likely to have an attribute we can use the worst case scenario: $p = 0.5$

$$\text{In this case } n = 1067$$

Note: as in the previous examples (using mean costs) we can apply the large sample correction factor by multiplying the answer (n) by $\sqrt{1-(n/N)}$ where N is the population and (n) the sample surveyed.