Chapter 5: Sampling, Variability, and Confidence

Sampling, or First Catch Your Hare!

Mrs. Beaton's famous English cookbook is alleged to have contained a recipe for Jugged Hare that started, "First. Catch your hare." It is too good a line to pass up, true or not. If you want the dish, catching the hare is the place to start. If you want to mine data, catching the "hare" in the data is the place to start. So what is the "hare" in data? The hare is the information content enfolded into the data set. Just as hare is the essence of the recipe for Jugged Hare, so information is the essence of the recipe for building training and test data sets.

Clearly, what is needed is enough data so that all of the relationships at all levels—superstructure, macrostructure, and microstructure—are captured. An easy answer would seem to be to use all the data. After all, with all of the data being used, it is a sure thing that any relationship of interest that the data contains is there to be found. Unfortunately, there are problems with the idea of using all of the data.

5.1.1 How Much Data?

One problem with trying to use all of the data, perhaps the most common problem, is simply that all of the data is not available. It is usual to call the whole of data the *population*. Strictly speaking, the data is not the population; the data is simply a set of measurements about the population of objects. Nonetheless, for convenience it is simply easier to talk about a population and understand that what is being discussed is the data, not the objects. When referring to the objects of measurement, it is easy enough to make it clear that the objects themselves are being discussed.

Suppose that a model is to be built about global forestry in which data is measured about individual trees. The population is at least all of the trees in the world. It may be, depending on the actual area of interest, all of the trees that have ever lived, or even all of the trees that could possibly live. Whatever the exact extent of the population, it is clearly unreasonable to think that it is even close to possible to have data about the whole population.

Another problem occurs when there is simply too much data. If a model of credit card transactions is proposed, most of these do actually exist on computers somewhere. But even if a computer exists that could house and process such a data set, simply accumulating all of the records would be at least ridiculously difficult if not downright impossible.

Currency of records also presents difficulties. In the case of the credit card transactions, even with the data coming in fast and furious, there would be no practical way to keep the data set being modeled reflecting the current state of the world's, or even the nation's, transactions.

For these reasons, and for any other reason that prevents having access to data about the whole population, it is necessary to deal with data that represents only some part of the population. Such data is called a *sample*.

Even if the whole of the data is available, it is still usually necessary to sample the data when building models. Many modeling processes require a set of data from which to build the model and another set of data on which to test it. Some modeling processes, such as certain decision tree algorithms, require three data sets—one to build the tree, one to prune the tree, and one to test the final result. In order to build a valid model, it is absolutely essential that each of the samples reflects the full set of relationships that are present in the whole population. If this is not the case, the model does not reflect what will be found in the population. Such a model, when used, will give inaccurate or misleading results.

So, sampling is a necessary evil. However, when preparing the data for modeling, the problem is not quite so great as when actually building the model itself. At least not in the early stages. Preparing the variables requires only that sufficient information about each individual variable be captured. Building data mining models requires that the data set used for modeling captures the full range of interactions between the variables, which is considered later, in Chapter 10. For now the focus is on capturing the variations that occur within each variable.

5.1.2 Variability

Each variable has features, many of which were discussed in Chapter 2. However, the main feature is that a variable can take on a variety of values, which is why it is called a variable! The actual values that a variable can have contain some sort of pattern and will be distributed across the variable's range in some particular way. It may be, for example, that for some parts of the range of values there are many instances bunched together, while for other parts there are very few instances, and that area of the range is particularly sparsely populated. Another variable may take on only a limited number of values, maybe only 5 or 10. Limited-value distribution is often a feature of categorical variables.

Suppose, for instance, that in a sample representative of the population, a random selection of 80 values of a numeric variable are taken as follows:

49, 63, 44, 25, 16, 34, 62, 55, 40, 31, 44, 37, 48, 65, 83, 53, 39, 15, 25, 52
68, 35, 64, 71, 43, 76, 39, 61, 51, 30, 32, 74, 28, 64, 46, 31, 79, 69, 38, 69

53, 32, 69, 39, 32, 67, 17, 52, 64, 64, 25, 28, 64, 65, 70, 44, 43, 72, 37, 31
67, 69, 64, 74, 32, 25, 65, 39, 75, 36, 26, 59, 28, 23, 40, 56, 77, 68, 46, 48

What exactly can we make of them? Is there any pattern evident? If there is, it is certainly hard to see. Perhaps if they are put into some sort of order, a pattern might be easier to see:

15, 16, 17, 23, 25, 25, 25, 25, 26, 28, 28, 28, 30, 31, 31, 31, 32, 32, 32, 32
34, 35, 36, 37, 37, 38, 39, 39, 39, 39, 40, 40, 43, 43, 44, 44, 44, 46, 46, 48
48, 49, 51, 52, 52, 53, 53, 55, 56, 59, 61, 62, 63, 64, 64, 64, 64, 64, 64, 65
65, 65, 67, 67, 68, 68, 69, 69, 69, 69, 70, 71, 72, 74, 74, 75, 76, 77, 79, 83

Maybe there is some sort of pattern here, but it is hard to tell exactly what it is or to describe it very well. Certainly it seems that some numbers turn up more often than others, but exactly what is going on is hard to tell.

Perhaps it would be easier to see any pattern if it were displayed graphically. Since the lowest number in the sample is 15, and the highest 83, that is the range of this sample. A *histogram* is a type of graph that uses columns to represent counts of features. If the sample is displayed as a histogram, some sort of pattern is easier to see, and Figure 5.1 shows a histogram of this sample. Each column in Figure 5.1 shows, by its height, the number of instances of a particular value. Each column represents one particular value. The first column on the left, for example, represents the value 15, and the column height indicates that there is one of this value in the sample.

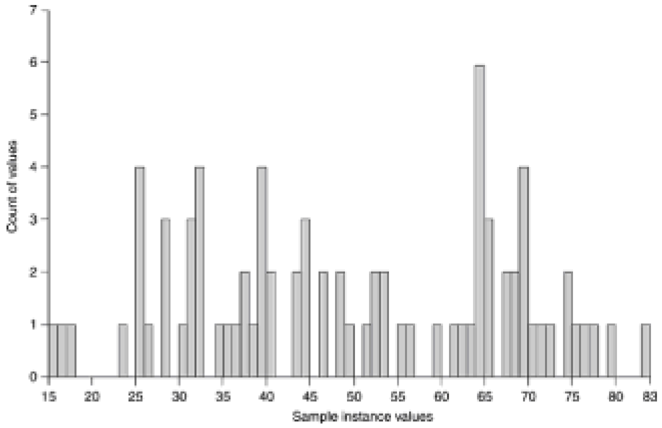

Figure 5.1 Histogram of a numeric variable sample. The column positions represent the magnitude of each of the values. The height of each column represents the count of instance values of the appropriate measured value.

The histogram in Figure 5.1 certainly makes some sort of pattern easier to see, but because of the number of columns, it is still hard to detect an overall pattern. Grouping the values together, shown in Figure 5.2, might make it easier to see a pattern. In this histogram each column represents the count of instances that are in a particular range. The leftmost column has a zero height, and a range of 0 to 9.99 (less than 10). The next column has a range from 10 to less than 20, and a height of 3. This second column aggregates the values 15, 16, and 17, which are all there are in the range of the column. In this figure the pattern is easier to see than in the previous figure.

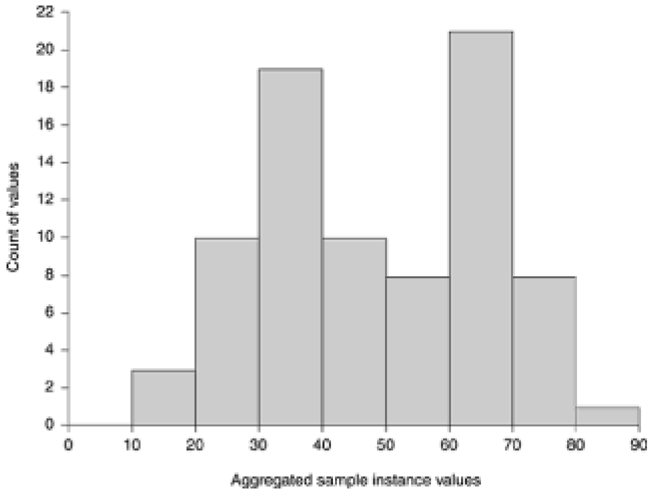

Figure 5.2 A histogram with vertical columns representing the count for a range of values.

Another way to see the distribution pattern is to use a graph that uses a continuous line, called a *curve*, instead of columns. Figure 5.3 shows a distribution curve that uses each value, just as in Figure 5.1. Again, the curve is very jagged. It would be easier to see the nature of the distribution if the curve were smoother. Curves can be easily smoothed, and Figure 5.4 shows the curve using two smoothing methods. One method (shown with the unbroken line) uses the average of three values; the other (shown with the dashed line) uses the average of five values. Smoothing does make the pattern easier to see, but it seems to be a slightly different pattern that shows up with each method. Which is the “correct” pattern shape for this distribution, if any?

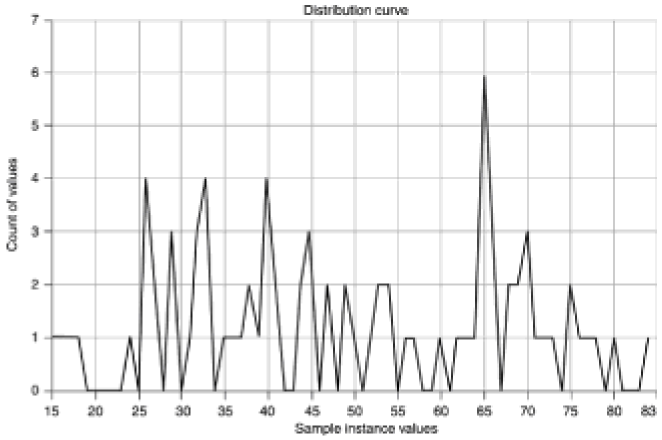

Figure 5.3 Sample value counts individually plotted and shown by a continuous line instead of using columns.

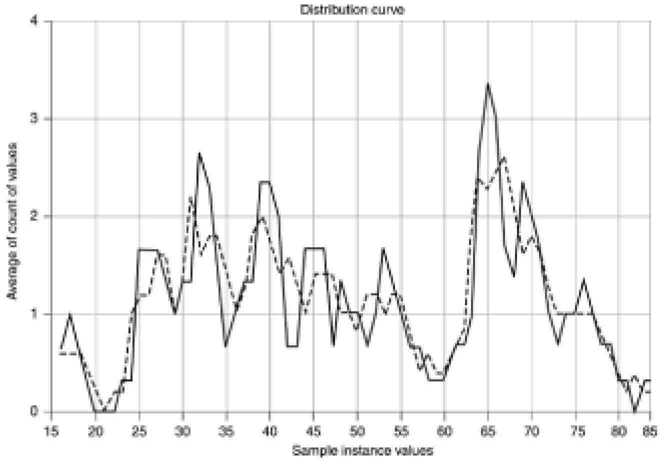

Figure 5.4 Results of two “smoothing methods, both using the average of a number of instance values. The solid line uses the average of three values, and the dashed line uses the average of five values.

There are two problems here. Recall that if this sample is indeed representative of the population, and for the purposes of this discussion we will assume that it is, then any other representative random sample drawn from the same population will show these patterns.

The first problem is that until a representative sample is obtained, and known to be representative, it is impossible to know if the pattern in some particular random sample does, in fact, represent the “true” variability of the population. In other words, if the true population distribution pattern is unknown, how can we know how similar the sample distribution curve is to the true population distribution curve?

The second problem is that, while it is obvious that there is some sort of pattern to a distribution, various ways of looking at it seem to produce slightly different patterns. Which of all these shapes, if any, is the right one to use?

5.1.3 Converging on a Representative Sample

The first problem, getting a representative sample, can be addressed by a phenomenon called *convergence*. Taking a sample starts by selecting instance values from a population, one at a time and at random. The sample starts at size 0. For any sample size a distribution curve can be created for the sample, similar to those shown in the earlier figures. In fact, although tedious for a human being, the distribution curve can be recalculated every time an instance value is added to the sample.

Suppose that the sample distribution curve is recalculated with each additional instance added. What will it look like? At first, when the number of instances in the sample is low, each addition will make a big impact on the shape of the curve. Every new instance added will make the curve “jump” up quite noticeably. Almost every instance value added to the sample will make a large change in the shape of the distribution curve. After a while, however, when the number of instances in the sample is modestly large, the overall shape of the curve will have settled down and will change little in shape as new instances are added. It will continue to increase in height because with more points in the sample, there are more points under any particular part of the curve. When there are a large number of instances in the sample, adding another instance barely makes any difference at all to the overall shape. The important point here is that the overall shape of the curve will settle down at some point.

This “settling down” of the overall curve shape is the key. As more instances are added, the actual shape of the curve becomes more like some final shape. It may never quite get there, but it gets closer and closer to settling into this final, unchanging shape. The curve can be thought of as “approaching” this ultimate shape. Things are said to converge when they come together, and in this sense the sample distribution curve converges with the final shape that the curve would have if some impossibly large number of instances were added. This impossibly large number of instances, of course, is the population. So the distribution curve in any sample converges with the distribution curve of the population as instances selected at random are added to the sample.

In fact, when capturing a sample, what is measured is not the shape of the curve, but the variability of the sample. However, the distribution curve shape is produced by the variability, so both measures represent very much the same underlying phenomenon. (And to understand what is happening, distribution curves are easier to imagine than variability.)

5.1.4 Measuring Variability

The other problem mentioned was that the distribution curve changes shape with the width of the columns, or the smoothing method. This problem is not so easy to address. What is really required instead of using column widths or smoothing is some method of measuring variability that does not need any arbitrary decision at all. Ideally, we need some method that simply allows the numbers sampled to be “plugged in,” and out comes some indication of the variability of the sample.

Statisticians have had to grapple with the problem of variability over many years and have found several measures for describing the characteristics of variables. Detailed discussion is beyond the scope of this book, but can be found in many statistical works, including those on business statistics. What they have come up with is a description of the *variability*, or *variance*, of a variable that captures the necessary variability information without being sensitive to column width or smoothing.

In many statistical texts, variability is very often described in terms of how far the individual instances of the sample are from the mean of the sample. It is, in fact, a sort of “average” distance of the instance values from the mean. It is this measure, or one derived from it, that will be used to measure variability. The measure is called the *standard deviation*. We need to look at it from a slightly different perspective than is usually found in statistics texts.

5.1.5 Variability and Deviation

Deviation is simply the name for what was described above as “a sort of average distance of instance values from the mean.” Given the same set of 80 numbers that were used before, the *mean*, often called the *arithmetic average*, or just average for short, is approximately 49.16. In order to find the distance of the instance values from the mean, it is only necessary to subtract the one from the other. To take the first five numbers as an example:

49 – 49.16 = –0.16

63 – 49.16 = 13.84

44 – 49.16 = –5.16

25 – 49.16 = –24.16

16 – 49.16 = –33.16

Unfortunately, the “–” signs make matters somewhat awkward. Since it is the mean that is being subtracted, the sum of all of the differences will add up to 0. That is what the mean is! Somehow it is necessary to make the “–” signs disappear, or at least to nullify their effect. For various reasons, in the days before computers, when calculations were all done by hand (perish the thought!), the easiest way for mathematicians to deal with the problem was not to simply ignore the “–” sign. Since “negative times negative is a positive,” as you may recall from school, squaring, or multiplying a number by itself, solves the problem. So finding the variance of just the first five numbers: The mean of only the first five numbers is

$(49 + 63 + 44 + 25 + 16)/5 = 39.4$

so squaring the instance value minus the mean:

$(49 – 39.4)^2 = 9.6^2 = 92.16$
$(63 – 39.4)^2 = 23.6^2 = 556.96$
$(44 – 39.4)^2 = 4.6^2 = 21.16$
$(25 – 39.4)^2 = –14.4^2 = 207.36$
$(16 – 39.4)^2 = –23.4^2 = 547.56$

and since the variance is the mean of these differences:

$(92.16 + 556.96 + 21.16 + 207.36 + 547.56)/5 = 285.04$

This number, 285.04, is the mean of the squares of the differences. It is therefore a variance of 285.04 square units. If these numbers represent some item of interest, say, percentage return on investments, it turns out to be hard to know exactly what a variance of 285.04 square percent actually means. Square percentage is not a very familiar or meaningful measure in general. In order to make the measure more meaningful in everyday terms, it is usual to take the square root, the opposite of squaring, which would give 16.88. For this example, this would now represent a much more meaningful variance of 16.88 percent.

The square root of the variance is called the *standard deviation*. The standard deviation is a very useful thing to know. There is a neat, mathematical notation for doing all of the things just illustrated:

Standard deviation = $\sqrt{(\Sigma(x - m)^2/(n - 1))}$

where

$\sqrt{\ }$ means to take the square root of everything under it

Σ means to sum everything in the brackets following it

x is the instance value

m is the mean

n is the number of instances

(For various technical reasons that we don’t need to get into here, when the number is divided by *n*, it is known as the standard deviation of the population, and when divided by *n* – 1, as the standard deviation of the sample. For large numbers of instances, which will usually be dealt with in data mining, the difference is miniscule.)

There is another formula for finding the value of the standard deviation that can be found in any elementary work on statistics. It is the mathematical equivalent of the formula shown above, but gives a different perspective and reveals something else that is going on inside this formula—something that is very important a little later in the data preparation process:

$$s = \sqrt{((\Sigma x^2 - nm^2)/(n - 1))}$$

What appears in this formula is “Σx^2,” which is the sum of the instance values squared. Notice also that “nm^2,” which is the number of instances multiplied by the mean, squared. Since the mean is just the sum of the *x* values divided by the number of values (or Σx/n), the formula could be rewritten as

$$s = \sqrt{((\Sigma x^2 - (n(\Sigma x/n))^2)/(n - 1))}$$

But notice that $n(\Sigma x$/n) is the same as Σx, so the formula becomes

$$s = \sqrt{((\Sigma x^2 - (\Sigma x)^2)/(n - 1))}$$

(being careful to note that Σx^2 means to add all the values of *x* squared, whereas $(\Sigma x)^2$ means to take the sum of the unsquared *x* values and square the total).

This formula means that the standard deviation can be determined from three separate pieces of information:

1. The sum of x^2, that is, adding up all squares of the instance values

2. The sum of *x*, that is, adding up all of the instance values

3. The number of instances

The standard deviation can be regarded as exploring the relationship among the sum of the squares of the instance values, the sum of the instance values, and the number of instances. The important point here is that in a sample that contains a variety of different values, the exact ratio of the sum of the numbers to the sum of the squares of the numbers is very sensitive to the exact proportion of numbers of different sizes in the sample. This sensitivity is reflected in the variance as measured by the standard deviation.

Figure 5.5 shows distribution curves for three separate samples, each from a different population. The range for each sample is 0–100. The linear (or rectangular) distribution sample is a random sample drawn from a population in which each number 0–100 has an equal chance of appearing. This sample is evidently not large enough to capture this distribution well! The bimodal sample was drawn from a population with two “humps” that do show up in this limited sample. The normal sample was drawn from a population with a normal distribution—one that would resemble the “bell curve” if a large enough sample was taken. The mean and standard deviation for each of these samples is shown in Table 5.1.

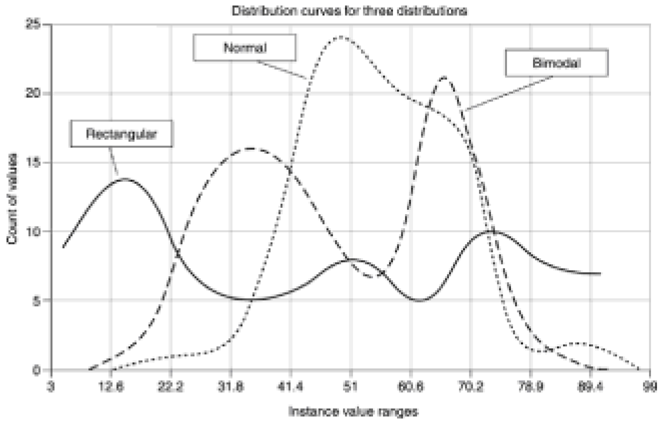

Figure 5.5 Distribution curves for samples drawn from three populations.

TABLE 5.1 Sample statistics for three distributions.

Sample distribution	Mean	Standard deviation
Linear	47.96	29.03
Bimodal	49.16	17.52
Normal	52.39	11.82

The standard deviation figures indicate that the linear distribution has the highest variance, which is not surprising as it would be expected to have the greatest average distance between the sample mean and the instance values. The normal distribution sample is the most bunched together around its sample mean and has the least standard deviation. The bimodal is more bunched than the linear, and less than the normal, and its standard deviation indicates this, as expected.

Standard deviation is a way to determine the variability of a sample that only needs to have the instance values of the sample. It results in a number that represents how the instance values are scattered about the average value of the sample.

5.2 Confidence

Now that we have an unambiguous way of measuring variability, actually capturing it requires enough instances of the variable so that the variability in the sample matches the variability in the population. Doing so captures all of the structure in the variable. However, it is only possible to be absolutely 100% certain that all of the variability in a variable has been captured if all of the population is included in the sample! But as we've already discussed, that is at best undesirable, and at worst impossible. Conundrum.

Since sampling the whole population may be impossible, and in any case cannot be achieved when it is required to split a collected data set into separate pieces, the miner needs an alternative. That alternative is to establish some acceptable degree of *confidence* that the variability of a variable is captured.

For instance, it is common for statisticians to use 95% as a satisfactory level of confidence. There is certainly nothing magical about that number. A 95% confidence means, for instance, that a judgment will be wrong 1 time in 20. That is because, since it is right 95 times in 100, it must be wrong 5 times in 100. And 5 times in 100 turns out to be 1 time in 20. The 95% confidence interval is widely used only because it is found to be generally useful in practice. “Useful in practice” is one of the most important metrics in both statistical analysis and data mining.

It is this concept of “level of confidence” that allows sampling of data sets to be made. If the miner decided to use only a 100% confidence level, it is clear that the only way that this can be done is to use the whole data set complete as a sample. A 100% sample is hardly a sample in the normal use of the word. However, there is a remarkable reduction in the amount of data needed if only a 99.99% confidence is selected, and more again for a 95% confidence.

A level of confidence in this context means that, for instance, it is 95% certain that the variability of a particular variable has been captured. Or, again, 1 time in 20 the full variability of the variable would not have been captured at the 95% confidence level, but some lesser level of variability instead. The exact level of confidence may not be important. Capturing enough of the variability is vital.

5.3 Variability of Numeric Variables

Variability of numeric variables is measured differently from the variability of nonnumeric variables. When writing computer code, or describing algorithms, it is easy to abbreviate numeric and nonnumeric to the point of confusion—“Num” and “Non.” To make the difference easier to describe, it is preferable to use distinctive abbreviations. This distinction is easy when using “Alpha” for nominals or categoricals, which are measured in nonnumeric scales, and “Numeric” for variables measured using numeric scales. Where convenient to avoid confusion, that nomenclature is used here.

Variability of numeric variables has been well described in statistical literature, and the previous sections discussing variability and the standard deviation provide a conceptual overview.

Confidence in variability capture increases with sample size. Recall that as a sample size gets larger, so the sample distribution curve converges with the population distribution curve. They may never actually be identical until the sample includes the whole population, but the sample size can, in principle, be increased until the two curves become as similar as desired. If we knew the shape of the population distribution curve, it would be easy to compare the sample distribution curve to it to tell how well the sample had captured the variability. Unfortunately, that is almost always impossible. However, it is possible to measure the rate of change of a sample distribution curve as instance values are added to the sample. When it changes very little with each addition, we can be confident that it is closer to the final shape than when it changes faster. But how confident? How can this rate of change be turned into a measure of confidence that variability has been captured?

5.3.1 Variability and Sampling

But wait! There is a critical assumption here. The assumption is that a larger sample is in fact more representative of the population as a whole than a smaller one. This is not necessarily the case. In the forestry example, if only the oldest trees were chosen, or only those in North America, for instance, taking a larger sample would not be representative. There are several ways to assure that the sample is representative, but the only one that can be assured not to introduce some bias is *random sampling.* A random sample requires that any instance of the population is just as likely to be a member of the sample as any other member of the population. With this assumption in place, larger samples will, on average, better represent the variability of the population.

It is important to note here that there are various biases that can be inadvertently introduced into a sample drawn from a population against which random sampling provides no protection whatsoever. Various aspects of sampling bias are discussed in Chapters 4 and 10. However, what a data miner starts with as a source data set is almost always a sample and not the population. When preparing variables, we cannot be sure that the original data is bias free. Fortunately, at this stage, there is no need to be. (By

Chapter 10 this is a major concern, but not here.) What is of concern is that the sample taken to evaluate variable variability is representative of the original data sample. Random sampling does that. If the original data set represents a biased sample, that is evaluated partly in the data assay (Chapter 4), again when the data set itself is prepared (Chapter 10), and again during the data survey (Chapter 11). All that is of concern here is that, on a variable-by-variable basis, the variability present in the source data set is, to some selected level of confidence, present in the sample extracted for preparation.

5.3.2 Variability and Convergence

Differently sized, randomly selected samples from the same population will have different variability measures. As a larger and larger random sample is taken, the variability of the sample tends to fluctuate less and less between the smaller and larger samples. This reduction in the amount of fluctuation between successive samples as sample size increases makes the number measuring variability converge toward a particular value.

It is this property of convergence that allows the miner to determine a degree of confidence about the level of variability of a particular variable. As the sample size increases, the average amount of variability difference for each additional instance becomes less and less. Eventually the miner can know, with any arbitrary degree of certainty, that more instances of data will not change the variability by more than a particular amount.

Figure 5.6 shows what happens to the standard deviation, measured up the side of the graph, as the number of instances in the sample increases, which is measured along the bottom of the graph. The numbers used to create this graph are from a data set provided on the CD-ROM called CREDIT. This data set contains a variable DAS that is used through the rest of the chapter to explore variability capture.

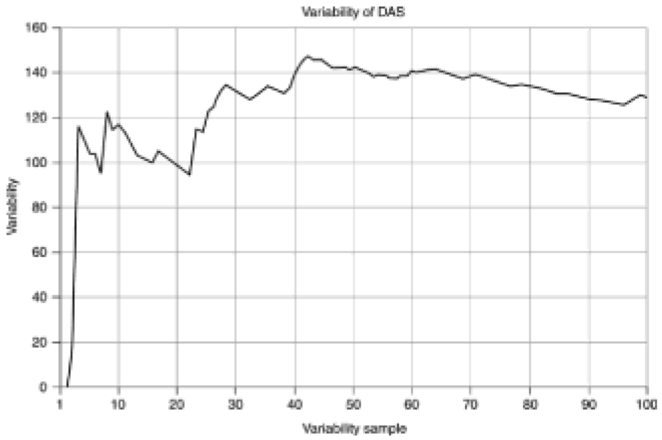

Figure 5.6 Measuring variability DAS in the CREDIT data set. Each sample contains one more instance than the previous sample. As the sample size increases, the variability seems to approach, or converge, toward about 130.

Figure 5.6 shows incremental samples, starting with a sample size of 0, and increasing the sample size by one each time. The graph shows the variability in the first 100 samples. Simply by looking at the graph, intuition suggests that the variability will end up somewhere about 130, no matter how many more instances are considered. Another way of saying this is that it has converged at about 130. It may be that intuition suggests this to be the case. The problem now is to quantify and justify exactly how confident it is possible to be. There are two things about which to express a level of confidence—first, to specify exactly the expected limits of variability, and second, to specify how confident is it possible to be that the variability actually will stay within the limits.

The essence of capturing variability is to continue to add samples until both of those confidence measures can be made at the required level—whatever that level may be. However, before considering the problem of justifying and quantifying confidence, the next step is to examine capturing variability in alpha-type variables.

5.4 Variability and Confidence in Alpha Variables

So far, much of this discussion has described variability as measured in numeric variables. Data mining often involves dealing with variables measured in nonnumeric ways. Sometimes the symbolic representation of the variable may be numeric, but the variable still is being measured nominally—such as SIC and ZIP codes.

Measuring variability in these alpha-type variables is every bit as important as in numerical variables. (Recall this is not a new variable type, just a clearer name for qualitative variables—nominals and categoricals—to save confusion.)

A measure of variability in alpha variables needs to work similarly to that for numeric variables. That is to say, increases in sample size must lead to convergence of variability. This convergence is similar in nature to that of numerical variables. So using such a method, together with standard deviation for numeric variables, gives measures of variability that can be used to sample both alpha and numeric variables. How does such a method work?

Clearly there are some alpha variables that have an almost infinite number of categories—people’s names, for instance. Each name is an alpha variable (a nominal in the terminology used in Chapter 2), and there are a great many people each with different names!

For the sake of simplicity of explanation, assume that only a limited number of alpha labels exist in a variable scale. Then the explanation will be expanded to cover alpha variables with very high numbers of distinct values.

In a particular population of alpha variables there will be a specific number of instances of each of the values. It is possible in principle to count the number of instances of each value of the variable and determine what percentage of the time each value occurs. This is exactly similar to counting how often each numeric instance value occurred when creating the histogram in Figure 5.1. Thus if, in some particular sample, "A" occurred 124 times, "B" 62 times, and "C" 99 times, then the ratio of occurrence, one to the others, is as shown in Table 5.2.

TABLE 5.2 Sample value frequency counts.

Sample distribution	Mean	Standard deviation
A	124	43.51
B	62	21.75
C	99	34.74
Total	285	100.00

If the population is sampled randomly, this proportion will not be immediately apparent. However, as the sample size increases, the relative proportion will become more and more nearly what is present in the population; that is, it converges to match that of the population. This is altogether similar to the way that the numeric variable variability converges. The main difference here is that since the values are alpha, not numeric, standard deviation can't be calculated.

Instead of determining variability using standard deviation, which measures the way numeric values are distributed about the mean, alpha variability measures the rate of change of the relative proportion of the values discovered. This rate of change is analogous to the rate of change in variability for numerics. Establishing a selected degree of confidence that the relative proportion of alpha values will not change, within certain limits, is analogous to capturing variability for a numeric variable.

5.4.1 Ordering and Rate of Discovery

One solution to capturing the variability of alpha variables might be to assign numbers to each alpha and use those arbitrarily assigned numbers in the usual standard deviation formula. There are several problems with this approach. For one thing, it assumes that each alpha value is equidistant from one another. For another, it arbitrarily assigns an ordering to the alphas, which may or may not be significant in the variability calculation, but certainly doesn't exist in the real world for alphas other than ordinals. There are other problems so far as variability capture goes also, but the main one for sampling is that it gives no clue whether all of the unique alpha values have been seen, nor what chance there is of finding a new one if sampling continues. What is needed is some method that avoids these particular problems.

Numeric variables all have a fixed ordering. They also have fixed distances between values. (The number "1" is a fixed distance from "10"—9 units.) These fixed relationships allow a determination of the range of values in any numeric distribution (described further in Chapter 7). So for numeric variables, it is a fairly easy matter to determine the chance that new values will turn up in further sampling that are outside of the range so far sampled.

Alphas have no such fixed relationship to one another, nor is there any order for the alpha values (at this stage). So what is the assurance that the variability of an alpha variable has been captured, unless we know how likely it is that some so far unencountered value will turn up in further sampling? And therein lies the answer—measuring the rate of discovery of new alpha values.

As the sample size increases, so the rate of discovery (ROD) of new values falls. At first, when the sample size is low, new values are often discovered. As the sampling goes on, the rate of discovery falls, converging toward 0. In any fixed population of alphas, no matter how large, the more values seen, the less new ones there are to see. The chance of seeing a new value is exactly proportional to the number of unencountered values in the population.

For some alphas, such as binary variables, ROD falls quickly toward 0, and it is soon easy to be confident (to any needed level of confidence) that new values are very unlikely. With other alphas—such as, say, a comprehensive list of cities in the U.S.—the probability would fall more slowly. However, in sampling alphas, because ROD changes, the miner can estimate to any required degree of confidence the chance that new alpha values will turn up. This in turn allows an estimate not only of the variability of an alpha, but of the comprehensiveness of the sample in terms of discovering all the alpha labels.

5.5 Measuring Confidence

Measuring confidence is a critical part of sampling data. The actual level of confidence selected is quite arbitrary. It is selected by the miner or domain expert to represent some level of confidence in the results that is appropriate. But whatever level is chosen, it is so important in sampling that it demands closer inspection as to what it means in practice, and why it has to be selected arbitrarily.

5.5.1 Modeling and Confidence with the Whole Population

If the whole population of instances were available, predictive modeling would be quite unnecessary. So would sampling. If the population really is available, all that needs to be done to “predict” the value of some variable, given the values of others, is to look up the appropriate case in the population. If the population is truly present, it is possible to find an instance of measurements that represents the exact instance being predicted—not just one similar or close to it.

Inferential modeling would still be of use to discover what was in the data. It might provide a useful model of a very large data set and give useful insights into related structures. No training and test sets would be needed, however, because, since the population is completely represented, it would not be possible to *overtrain*. Overtraining occurs when the model learns idiosyncrasies present in the training set but not in the whole population. Given that the whole population is present for training, anything that is learned is, by definition, present in the population. (An example of this is shown in Chapter 11.)

With the whole population present, sampling becomes a much easier task. If the population were too large to model, a sample would be useful for training. A sample of some particular proportion of the population, taken at random, has statistically well known properties. If it is known that some event happens in, say, a 10% random sample with a particular frequency, it is quite easy to determine what level of confidence this implies about the frequency of the event in the population. When the population is not available, and even the size of the population is quite unknown, no such estimates can be made. This is almost always the case in modeling.

Because the population is not available, it is impossible to give any level of confidence in any result, based on the data itself. All levels of confidence are based on assumptions about the data and about the population. All kinds of assumptions are made about the randomness of the sample and the nature of the data. It is then possible to say that *if* these assumptions hold true, then certain results follow. The only way to test the assumptions, however, is to look at the population, which is the very thing that can’t be done!

5.5.2 Testing for Confidence

There is another way to justify particular levels of confidence in results. It relies on the quantitative discriminatory power of tests. If, for instance, book reviewers can consistently and accurately predict a top 10 best-selling book 10% of the time, clearly they are wrong 90% of the time. If a particular reviewer stated that a particular book just reviewed was certain to be a best-seller, you would be justified in being skeptical of the claim. In fact, you would be quite justified in being 10% sure (or confident) that it would be a success, and 90% confident in its failure. However, if at a convention of book reviewers, every one of hundreds or thousands of reviewers each separately stated that the book was sure to be a best-seller, even though each reviewer had only a 10% chance of success, you would become more and more convinced of the book's chance of success.

Each reviewer performs an independent reading, or test, of the book. It is this independence of tests that allows an accumulation of confidence. The question is, how much additional confidence is justified if two independent tests are made, each with a 10% accuracy of being correct in their result, and both agree? In other words, suppose that after the first reviewer assured you of the book's success, a second one did the same. How much more confident, if at all, are you justified in being as a result of the second opinion? What happens if there are third and fourth confirming opinions? How much additional confidence are you justified in feeling?

At the beginning you are 100% skeptical. The first reviewer's judgment persuades you to an opinion of 10% in favor, 90% against the proposition for top 10 success. If the first reviewer justified a 10/90% split, surely the second does too, but how does this change the level of confidence you are justified in feeling?

Table 5.3 shows that after the first reviewer's assessment, you assigned 10% confidence to success and 90% to skepticism. The second opinion (test) should also justify the assignment of an additional 10%. However, you are now only 90% skeptical, so it is 10% of that 90% that needs to be transferred, which amounts to an additional 9% confidence. Two independent opinions justify a 19% confidence that the book will be a best-seller. Similar reasoning applies to opinions 3, 4, 5, and 6. More and more positive opinions further reinforce your justified confidence of success. With an indefinite amount of opinions (tests) available, you can continue to get opinions until any particular level of confidence in success is justified.

TABLE 5.3 Reviewer assurance charges confidence level.

Reviewer number	Start level	Transfer amount (start level x10%)	Confidence of success	Your remaining skeptical balance
1	100%	10	10	90
2	90%	9	9	81

3	81%	8.1	27.1	72.9
4	72.9%	7.29	34.39	65.61
5	65.61%	6.561	40.951	59.049
6	59.049%	5.9049%	46.8559	53.1441

Of course, a negative opinion would increase your skepticism and decrease your confidence in success. Unfortunately, without more information it is impossible to say by how much you are justified in revising your opinion. Why?

Suppose each reviewer reads all available books and predicts the fate of all of them. One month 100 books are available, 10 are (by definition) on the top 10 list. The reviewer predicts 10 as best-sellers and 90 as non-best-sellers. Being consistently 10% accurate, one of those predicted to be on the best-seller list was on it, 9 were not. Table 5.4 shows the reviewer's hit rate this month.

TABLE 5.4 Results of the book reviewer's predictions for month 1.

Month 1	**Best-seller**	**Non-best seller**
Predicted bestseller	1	9
Predicted non-best-seller	9	81

Since one of the 10 best-sellers was predicted correctly, we see a 10% rate of accuracy. There were also 90 predicted to be non-best-sellers, of which 81 were predicted correctly as non-best-sellers. (81 out of 90 = 81/90 = 90% incorrectly predicted.)

In month 2 there were 200 books published. The reviewer read them all and made 10 best-seller predictions. Once again, a 10% correct prediction was achieved, as Table 5.5 shows.

TABLE 5.5 Results of the book reviewer's predictions for month 2.

Month 1	Best-seller	Non-best seller
Predicted bestseller	1	9
Predicted non-best-seller	9	181

Once again, there are 10 best-sellers and one was correctly predicted for a correct pick rate of 10%. However, there were 190 books predicted to be non-best-sellers this month, of which 181 were correctly predicted because they weren't best-sellers. However, 181 out of 190 is a correct prediction rate for non-best-sellers of 95.26%!

What is going on here? The problem is that predicting best-sellers and predicting non-best-sellers are *not* two sides of the same problem, although they look like they might be. The chances of being right about best-sellers are not the opposite of the chances of being right about non-best-sellers. This is because of the old bugaboo of knowledge of the size of the population. What changed here is the size of the population from 100 to 200. The number of best-sellers is always 10 because they are defined as being the 10 best-selling books. The number of non-best-sellers depends entirely on how many books are published that month.

However (and this is a very important point), deciding how much confidence can be justified after a given number of tests depends *only on the success ratio of the tests*. This means that if the success/fail ratio of the test is known, or assumed, knowledge of the size of the population is not needed in order to establish a level of confidence. With this knowledge it is possible to construct a test that doesn't depend on the size of the population, but only on the consecutive number of confirmatory tests.

The confidence generated in the example is based on predicting best-sellers. The number of best-sellers is a purely arbitrary number. It was just chosen to suit the needs of the selector. After all, it could have been the top 12, or 17, or any other number. The term "best-seller" was defined to suit someone's convenience. It is very likely that the success of reviewers in picking best-sellers would change if the definition of what constituted a best-seller changed. The point here is that if the chosen assumptions meet the needs of whoever selected them, then a rational assessment of confidence can be made based on those assumptions.

5.5.3 Confidence Tests and Variability

The consequence for determining variability of a variable is that the modeler must make assumptions that meet the modeler's needs. Choosing a 95% level of confidence implies saying, among other things, "If this test is wrong 95% of the time, how many times must independent tests confirm its correctness before the cumulative judgment can be accepted as correct at least 95% of the time?"

In practical terms (using the 95% level of confidence for this discussion), this implies several consequences. Key, based on the level of confidence, is that a single test for convergence of variability is incorrect 95% of the time and correct 5% of the time. From that it is possible to rationally accumulate confidence in a continuing series of positive results. (Positive results indicate variability convergence.) After some unbroken series of positive results, a level of confidence is accumulated that exceeds 95%. When that happens you can be sure that accepting the convergence as complete will only be a mistake 1 time in 20, or less.

At the end, the result is a very simple formula that is transformed a little and used in the demonstration software to know when enough is enough. That is,

$$s = e^{t}$$

where

s = Justified level of skepticism

e = Error rate

t = Number of positive tests

Results of this formula, using the 90% confidence level from the best-seller example, are given in Table 5.6.

TABLE 5.6 Results of the book reviewer's predictions for month 2.

Skepticism	**Error rate**	**Number of tests**
0.9	0.9	1
0.81	0.9	2

0.729	0.9	3
0.6561	0.9	4
0.59049	0.9	5
0.531441	0.9	6

This is the same series of numbers shown in the “Your remaining skeptical balance” column in Table 5.3 except that these are expressed as decimals rather than percentages.

Of course, this diminishing level of skepticism indicates the confidence that you are *wrong*. The confidence that you are right is what is left after subtracting the confidence that you are wrong! The confidence level in being right is, therefore,

$c = 1 – e^t$

where

c = Confidence

e = Error rate

t = Number of positive tests

The Supplemental Material section at the end of this chapter briefly shows the transformation from this statement into one that allows the number of consecutive true tests to be directly found from the error rate. It is this version of the formula that is used in the demonstration software. However, for those who need only to understand the concepts and issues, that section may be safely skipped.

5.6 Confidence in Capturing Variability

Capturing the variability of a variable means, in practice, determining to a selected level of confidence that the measured variability of a sample is similar to that of the population, within specified limits. The measure of sample variability closeness to population variability is measured by convergence in increasingly larger samples. In other words, converged variability means that the amount of variability remains within particular limits for enough independent tests to be convincing that the convergence is real. When the variability is converged, we are justified in accepting, to a certain level of confidence, that the variability is captured. This leads to two questions. First, exactly what is measured to determine convergence, and second, what are the “particular limits” and how they are discovered?

On the accompanying CD-ROM there is a data set CREDIT. This includes a sample of real-world credit information. One of the fields in that data set is “DAS,” which is a particular credit score rating. All of the data used in this example is available on the CD-ROM, but since the sample is built by random selection, it is very unlikely that the results shown here will be duplicated exactly in any subsequent sample. The chance of a random sampling procedure encountering exactly the same sequence of instance values is very, very low. (If it weren’t, it would be of little value as a “random” sequence!) However, the results do remain consistent in that they converge to the same variability level, for a given degree of confidence, but the precise path to get there may vary a little.

5.6.1 A Brief Introduction to the Normal Distribution

Capturing variability relies on assuming normality in the distribution of the test results, and using the known statistical properties of the normal distribution. The assumption of normality of the distribution of the test results is particularly important in estimating the probability that variability has actually converged. A brief and nontechnical examination of some facets of the normal distribution is needed before looking at variability capture.

The normal distribution is well studied, and its properties are well known. Detailed discussion of this distribution, and justification for some of the assertions made here, can be found in almost any statistical text, including those on business statistics. Specifically, the distribution of values within the range of a normally distributed variable form a very specific pattern. When variables’ values are distributed in this way, the standard deviation can be used to discover exactly how likely it is that any particular instance value will be found. To put it another way, given a normally distributed sample of a particular number of instances, it is possible to say how many instances are likely to fall between any two values.

As an example, about 68% of the instance values of a normally distributed variable fall inside the boundary values set at the mean-plus-1 standard deviation and the mean-minus-1 standard deviation. This is normally expressed as $m \pm s$, where m is the mean and s the standard deviation. It is also known, for instance, that about 95.5% of the sample’s instance values will lie within $m \pm 2s$, and 99.7% within $m \pm 3s$.

What this means is that if the distance and direction from the mean is known in standard deviation units for any two values, it is possible to determine precisely the probability of discovering instance values in that range. For instance, using tables found in any elementary statistics text, it is easy to discover that for a value of the mean-minus-1 standard deviation, approximately 0.1587 (i.e., about 16%) of the instances lie in the direction away from the mean, and therefore 0.8413 (about 84%) lie in the other direction.

The normal curve shown in Figure 5.7 plots values in the sample along the horizontal axis (labeled Standard deviation) and the probability of finding a value on the vertical axis. In a normally distributed sample of numbers, any value has some probability of occurring, but with a vanishingly small probability as the distance of the values moves far from the mean. The curve is centered on the sample mean and is usually measured in standard deviation units from the mean. The total area under the curve corresponds to a probability of 100% for finding a value in the distribution. The chance of finding a specific value corresponds to the area under the curve for that value. It is easier to imagine finding a value in some interval between two values.

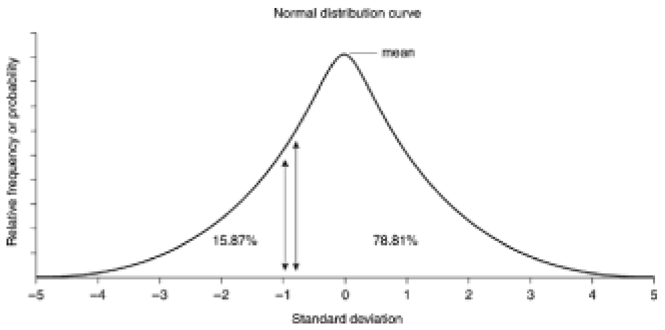

Figure 5.7 Normal distribution curve with values plotted for standard deviation (x-axis) and probability of finding a value (y-axis).

This figure shows the normal curve with the interval between –1 and –0.8 standard deviations. It can be found, by looking in standard deviation tables, that approximately 15.87% of the instance values lie to the left of the mean-minus-1 standard deviation line, 78.81% lie to the right of the –0.8 standard deviation line, which leaves 5.32% of the distribution between the two (100% – 78.81% – 15.78% = 5.32%). So, for instance, if it were known that some feature fell between these two limits consistently, then the feature is “pinned down” with a 94.68% confidence (100% – 5.32% = 94.68%).

5.6.2 Normally Distributed Probabilities

Measuring such probabilities using normally distributed phenomena is only important, of course, if the phenomena are indeed normally distributed.

Looking back at Figure 5.6 will very clearly show that the variability is not normally distributed, nor even is the convergence. Fortunately, the changes in convergence, that is, the size of the change in variance from one increment to the next increment, can easily be adjusted to resemble a normal distribution. This statement can be theoretically supported, but it is easy to intuitively see that this is reasonable.

Early in the convergence cycle, the changes tend to be large compared to later changes.

This is true no matter how long the convergence cycle continues. This means that the proportion of relatively small changes always predominates. In turn, this leads to the conclusion that the more instances that are considered, the more the later changes in variance cluster about the mean. Relatively large changes in variance, both positive and negative, are much less likely than small changes. And that is exactly what the normal distribution describes.

To be sure, this is only a descriptive insight, not proof. Unfortunately, proof of this takes us beyond the scope of this book. The Further Reading section at the end of this book has pointers to where to look for further exploration of this and many other areas.

It must be noted that the variance distribution is not actually normal since convergence can continue arbitrarily long, which can make the number of small changes in variability far outweigh the large changes. Figure 5.8 shows part of the distribution for the first 100 samples of DAS variance. This distribution is hardly normal! However, although outside the scope of this conceptual introduction, adjustment for the reduction in change of size of variance with the sample is fairly straightforward. Figure 5.9 shows the effect of making the adjustment, which approximately normalizes the distribution of changes in variance—sufficient to make the assumptions for testing for convergence workably valid. Figure 5.9 shows the distribution for 1000 variability samples.

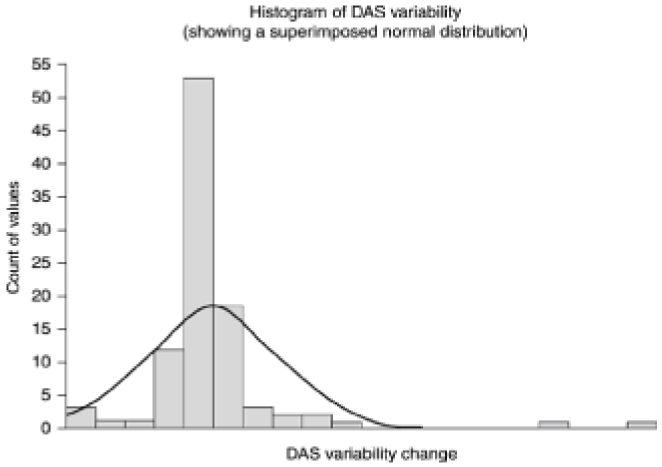

Figure 5.8 The actual distribution of the changes in variability of DAS for the first 100 samples shown in Figure 5.6.

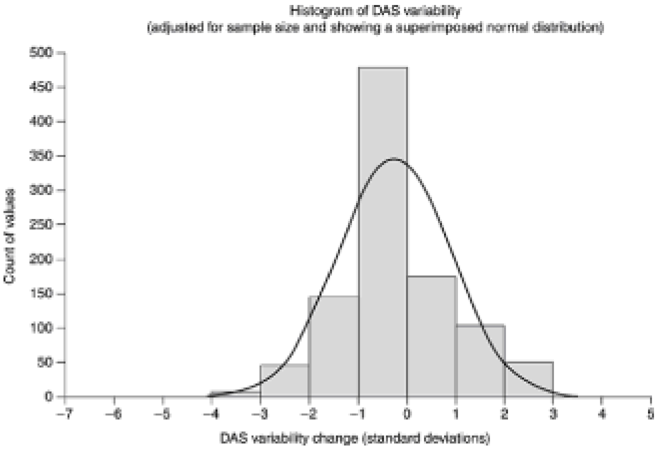

Figure 5.9 The variability of DAS for a 1000-instance sample when adjusted for sample size.

5.6.3 Capturing Normally Distributed Probabilities: An Example

After much preamble, we are at last ready to actually capture DAS variability. Recall that Figure 5.6 showed how the variance of DAS changes as additional instances are sampled. The variance rises very swiftly in the early instances, but settles down (converges) to remain, after 100 samples, somewhere about 130. Figure 5.10 shows the process of capturing variability for DAS. At first glance there is a lot on this graph! In order to get everything onto the same graph together, the figure shows the values “normalized” to fit into a range between 0 and 1. This is only for the purposes of displaying everything on the graph.

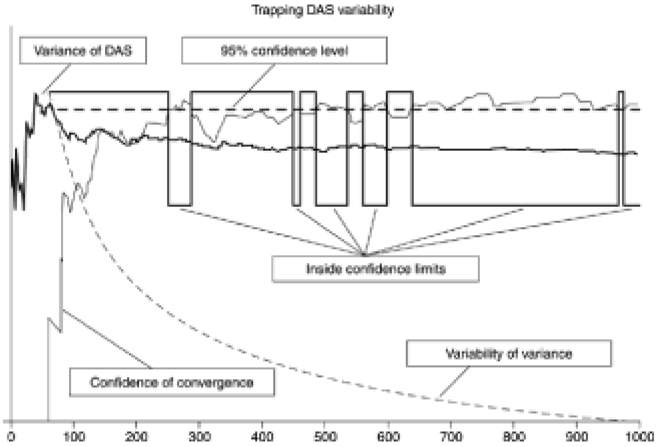

Figure 5.10 Various features shown on a common scale so that their relationships are more easily seen.

This example uses a 95% confidence level, which requires that the variability be inside 0.05 (or 5%) of its range for 59 consecutive tests. In this example, the sampling is continued long after variability was captured to see if the confidence was justified. A total of 1000 instance-value samples are used.

As the variance settles down, the confidence level that the variability has been captured rises closer to “1,” which would indicate 100% confidence. When the confidence of capture reaches 0.95, in actual data preparation, the needed confidence level is reached and sampling of this variable can stop. It means that at that point there is a 95% confidence that at least 95% of the variability has been captured.

Because the example does not stop there, the variability pops out of limits from time to time. Does this mean that the measurement of variability failed? When the variability first reached the 0.95 confidence level, the variability was 127.04. After all 1000 samples were completed, the variability was at 121.18. The second variance figure is 4.6% distant from the first time the required confidence level was reached. The variance did indeed stay within 5% for the measured period. Perhaps it might have moved further distant if more instances had been sampled, or perhaps it might have moved closer. The measurement was made at the 95% confidence interval for a 95% variability capture. So far as was measured, this confidence level remains justified.

5.6.4 Capturing Confidence, Capturing Variance

This is a complex subject, and it is easy to confuse what actually has been captured here. In the example the 95% level was used. The difficulty is in distinguishing between capturing 95% of the DAS variability, and having a 95% confidence that the variability is captured. Shouldn’t the 95% confidence interval of capturing 95% of the variability really indicate a 0.95 x 0.95 = 0.9025 confidence?

The problem here is the difficulty of comparing apples and oranges. Capturing 95% of the variability is not the same as having 95% confidence that it is captured. An example might help to clarify.

If you have an interest-bearing bank account, you have some degree of assurance, based on past history, that the interest rate will not vary more than a certain amount over a given time. Let’s assume that you think it will vary less than 5% from where it is now over the next six months. You could be said to be quite certain that you have locked in at least 95% of the current interest rate. But how certain are you? Perhaps, for the sake of this discussion, you can justify a 95% confidence level in your opinion.

So, you are 95% confident of capturing 95% of the current interest rate. By some strange coincidence, those are the numbers that we had in capturing the variability! Ask this: because 0.95 x 0.95 = 0.9025, does this mean that you are really 90.25% confident? Does it mean the interest rate is only 90.25% of what you thought it was? No. It means