8 Exploratory Data Analyses

Broadly defined, the analysis of data can be considered in terms of two, largely non-overlapping aspects: *exploratory* and *confirmatory*. The former is usually an open-ended undertaking, focused on uncovering previously unknown insights, relationships or dependencies. The latter is focused primarily on hypothesis testing, or validating theoretically or otherwise derived beliefs. Typically, data exploration precedes confirmatory analysis—in fact, the initial exploratory analysis might give rise to hypothesis, to be tested within the realm of confirmatory analysis.

The *exploratory* → *confirmatory* progression is bound in the context of a particular informational need; furthermore, there are a number of methodologically distinct analytic options. More specifically, the combination of the nature of the business question and the characteristics of the available data will jointly determine the scope and the character of the initial data exploration, as discussed in this chapter.

Initiating Data Analyses

In a conceptual sense, the analysis of data has a very straightforward meaning in the context of the database analytical process: *The conversion of informationally meaningless raw inputs into useful and specific knowledge*. Implied in this conceptualization is the reliance on appropriate statistical techniques, the choice of which is determined by the end objectives of the data analytical process. In an operational sense, this means choosing from among an array of methods that, though similar in some regards, tend to be quite dissimilar in terms of their applicability limits.

As pointed out earlier, data analysis is often equated with summarization, tabulation and reporting. And though there certainly is value in keeping current on business-related outcomes and emerging trends, this type of information rarely, if ever, gives rise to informational advantage, for reasons detailed in Chapter 1 (i.e., the proliferation of generic data capture and reporting systems). Organizations that effectively use data to out-smart their competitors are those that found the way to systematically translate it into decision-aiding and competitively unique insights. And although the analytically proficient universe of organizations is growing, many firms nonetheless struggle to reap the benefits of their often considerable database infrastructure investments. One of the more frequently encountered reasons is the lack of methodological sophistication, not necessarily in the sense of academic knowledge.

There are a number of significant differences between the theory-building-focused academic research and the competitive-edge-oriented practical marketing analytics. First and foremost, the former seeks universally true generalizations, while the latter pursues entity (i.e., an organization) and situation (i.e., a specific business context at a point in time) unique insights. Hence the most important methodological considerations surrounding theory-building research pertain to sample-to-universe generalizability, which is in sharp contrast to future replicability demands of applied business analyses. Though seemingly of more philosophical than practical importance, these differences have a profound impact on the applicability limits of some of the more commonly used statistical techniques, as detailed later. The degree to which otherwise (i.e., academically)

proficient analysts do not recognize these fundamental incongruities will diminish the quality of their results.

Subsumed under the global considerations of applicability limits of broadly defined methodological approaches is the selection of specific techniques, or computational algorithms. It is intuitively obvious that a given dataset can be analyzed in a variety of ways, particularly in the sense of specific statistical formulations. For instance, the goal of identifying segments of the customer base can be reached with the help of cluster analysis, perceptual mapping, classification trees or latent class models, to name a few distinctly different grouping formulations (see Chapter 6 for details). The proliferation of choices is, as expected, a direct consequence of progress. The availability of comprehensive and computationally powerful statistical software packages, such as SAS, SPSS or R, offering a relatively easy access to a wide array of purpose-similar techniques, eases the task of physically "crunching the data," but adds a layer of complexity to choosing the "right" data crunching technique. This underscores the importance of thorough and well-thought-out analytical planning (and a plan) described earlier.

In a more prescriptive sense, the success of virtually all data analytical endeavors hinges on structuring the knowledge creation efforts around a clear *informational needs–available data–analytic approaches and tools* rationale. As pointed out earlier, the nucleus of this process is an in-depth comparison of the stated informational needs with the informational content of the available data. Assuming the presence of clearly delineated informational objectives, this chapter offers a comprehensive overview of a process of systematically exploring the data as the first step in a much broader undertaking of extracting edge-producing knowledge.

Exploration vs. Hypothesis Testing

To a statistician, all data analytical endeavors can be categorized as either *exploratory* or *confirmatory* (also called *hypothesis testing*). Before delving into specifics of data exploration, which is the focus of this section, we should establish definitional clarity of this notion, as it relates to the goals of database analytics.

According to Wikipedia, *exploratory data analysis (EDA) is that part of data analysis concerned with reviewing, communicating and using data where there is a low level of knowledge about its cause system*. It is also sometimes referred to as "data mining," in recognition of the fact that a common objective of initial data explorations is the identification of not-yet-known patterns and/or relationships in the data. Exploratory investigations may exhibit varying degrees of complexity and sophistication and generally make use of various numeric/statistical tests as bases for establishing the validity and reliability of its conclusions. Traditionally, analysts "manually" sifted through datasets in search of (numeric test-supported and thus statistically) significant relationships and patterns. For example, a correlation matrix of appropriately selected variables in conjunction with statistical significance tests can be used as the starting point in evaluating relationships among a number of metrics, based upon which, an analyst can uncover not-yet-known relationships.

Over the last couple of decades, certain aspects of "manual" data explorations began to benefit from the explosive wave of innovations sweeping across virtually all corners of the IT sector. Of most interest to marketing analytics have been two particular sets of developments: *Automated data mining* and *data visualization*. The former represents an attempt at leveraging the advancements in data processing and software technologies to develop stand-alone, self-operating automated data mining systems. These complex applications governed by sophisticated algorithms

with smart-sounding names, such as genetic algorithms of neural networks, and are intended to perform the job of an analyst by exploring data for hints of noteworthy patterns and relationships.

A related, albeit substantively quite different is a family of data processing tools focused on data visualization. Like the aforementioned automated data mining, these too are stand-alone software applications, but in contrast to data mining tools, data visualization systems require an active participation of the part of the analyst. Their goal is to replace some of the obtuse and frequently misinterpreted (and misunderstood) numeric tests with visual representations of the relationships found in data, all with the goal of making the results easier to consume by non-technical users. It follows that, although not all relationships can be depicted visually, a number of simpler, basic reporting functions can be handled quite effectively and elegantly in that format. After all, a picture is worth a thousand words (or numbers)…

In contrast to the "let's see what's there" goal of exploratory data analyses, *confirmatory analyses* are primarily concerned with the assessment of the viability of specific knowledge claims, or stated more formally, with the testing of specific hypotheses. Among the more common database analytical applications of confirmatory data analyses are analyses aimed at *confirming* average $$ spend-based customer value categorization (testing the hypothesis[1] that all customers are equally valuable), or cross-customer segment promotional-response elasticity estimation (testing the null hypothesis that all targets are equally likely to repurchase). In general, these types of analyses start from the premise that certain patterns or relationships exist though ongoing validation is necessary, in part to rationalize investment allocations.

Figure 8.1 captures the different faces of exploratory data analysis.

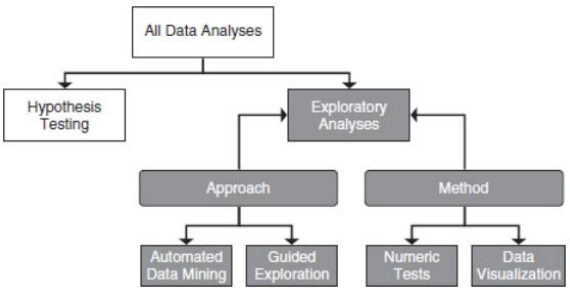


Figure 8.1 Exploratory Data Analyses

The difference between exploratory and confirmatory categories of analyses is best illustrated by customer segmentation methodologies discussed in Chapter 9. While some types of segmentation methodologies are focused on the description of an underlying structure of the customer base (i.e., are exploratory in nature), others are directed at predicting expected future behaviors (i.e., are built around difference hypotheses positing that some customer segments will outperform others). As detailed in subsequent chapters, the two entail sharp methodological and procedural differences. The remaining part of this chapter will be focused on an in-depth discussion of the exploratory dimension of the database analytical process.

Database Analytics and Data Exploration

Data exploration can take many forms, but in the broadest sense, it can entail either self-directed, open-ended data mining, or an analyst-driven, specific informational objective-focused analysis. Unfortunately, these two philosophically and methodologically distinct approaches are at times used—at least name-wise—synonymously, which is problematic, considering that, as mentioned earlier, the two can yield substantially different results: The outcomes of automated data mining are in practice highly unpredictable, simply because this type of database exploration tends to entail the widest possible scope of searching for *any* (statistically significant) patterns or systematic associations. On the other hand, analyst-directed exploratory database analyses are, from the very beginning, focused on testing the viability and reliability of a-priori-identified potential relationships, which in the context of the knowledge creation process outlined in this book would typically stem from the stated informational objectives. In other words, the former is usually driven by the question of "what's there?," while the latter is directed by the question of "what's there, that is related to those particular goals?" Not surprisingly, when it comes to data exploration, approach/methodological clarity is the key.

First and foremost: Automated pattern and/or relationship discovery certainly has its place in applied analytics, but it is not appropriate for the database analytical process outlined here, particularly when large, transactional databases come into play. As noted earlier, as a mode of insight discovery, this particular approach to database exploration is generally informationally unconstrained, in the sense of being limited to only a subset of all potential relationships. In practice, it leads to a "lumping together" of all statistically significant relationships, without regard for their business value.[2] Considering the typically large volumes of data (record counts are almost always in millions and quite frequently billions of individual records), coupled with a large number of available metrics, it is easy to see that any automated data mining tools,[3] designed to sift through the contents of these voluminous data repositories in search of any relationships that might be "hidden" there, are likely to generate an overwhelming number of "significant" relationships. Given the practical limitations of statistical significance tests (see the *Beware of Statistical Significance Tests* section later in this chapter), which offer the only objective means of identifying patterns and relationships in the data, there is oftentimes very little separating true informational nuggets from practically trivial informational tidbits. Quite commonly, an analyst will more-or-less arbitrarily pare down the list of all statistically significant relationships to a smaller, more manageable set of practically significant ones, which though understandable on a practical level, is hard to defend on the methodological level.

Automated data mining is also philosophically incongruent with the goal of organizational objectives-driven knowledge creation process, as it represents an *inductive* (or *bottom-up*, to use a business vernacular) approach to the creation of knowledge. Directed exploratory analyses, on the other hand, represent a *deductive* (also called *top-down*) insight-discovery mechanism. Although seemingly trifling, this distinction is actually quite critical in view of the database analytical knowledge creation processes described in this book.

As detailed in Chapter 2, the difference between the competitive advantage-creating knowledge generation and the accumulation of assorted informational tidbits is most evident in the extent of problem solving specificity. While the former answers specific organizational strategy-related questions, the latter tends to yield an assortment of happenstance details in the hope that those will somehow spark a Eureka-like moment. Though a Eureka-like moment is certainly always a possibility, the probability of it is usually rather remote and practically nonexistent on a repeated

, systematic basis. A winning business strategy is usually "90% perspiration and 10% inspiration," to use a common expression. Counting on consistently picking the "right" practically

significant insight out of a sea of statistically significant ones bears the resemblance of basing one's retirement plan on hitting a jackpot in Las Vegas.

As depicted in Figure 8.1, exploratory data analysis (EDA) can also take on two distinctly different methodological directions: numerical/statistical testing-based decision rule vs. visual/spatial data interpretation. In keeping with the belief that a "picture is worth a thousand words," spatially based data exploration has steadily grown in popularity; however it is best suited for diagnostic rather than relationship-testing purposes. The main reason behind this conclusion is that the knowledge creation demands a relationship-testing facility that can support an ongoing and unbiased assessment of the relationships of interest. Spatial data presentation delegates the task of identifying and assessing patterns and relationships to individual analysts, which confounds conclusions with subjective opinions, all of which may ultimately bias the results. In the end, objective, properly used and interpreted numerical tests are preferred over the more subjective visual data interpretation (see the *Beware of Statistical Significance Tests* discussion presented below). Spatial data representation, however, is a potent knowledge communication tool and can be a very effective method of conveying otherwise abstract or complex relationships.

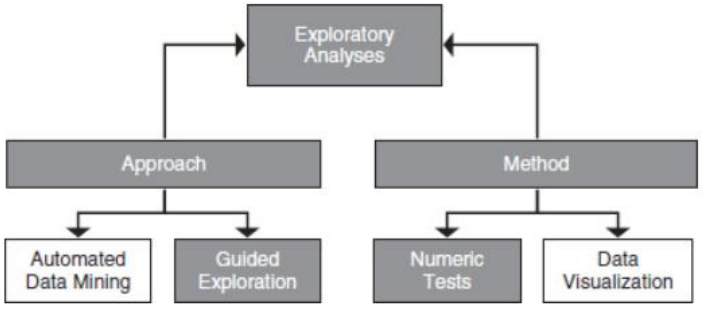


Figure 8.2 Database Analytics' EDA

Figure 8.3 EDA Guided by Strategic Goals

Summarizing the above discussion, it follows that in the context of database analytics, *exploratory data analysis* is defined as *a multi-step process guided by the firm's strategic goals and tasked with identifying specific informational foundations to support the development of competitively advantageous knowledge*, as graphically depicted in Figure 8.2.

Data Exploration Process

To yield the desired outcomes, data exploration should follow a well-defined process. Here, the seminal work of Tukey[4] (who is also credited with creating the term *exploratory data analysis*) offers a good and widely accepted overall framework. However, the database-analytics-driven knowledge creation process outlined in this book imposes a new set of limitations and requirements, most of which stem from the previously discussed characteristics of many of the

transactional databases. Specifically, the said analytical process limits the conventional data exploration process to the identification and quantification of specific data elements, as dictated by the specific informational needs stemming from the stated organizational objectives, shown in Figure 8.3 below. At the same time, it extends the said process to also include the creation of summary and index variables. Hence it is appropriate to differentiate between the general, open-ended *exploratory data analysis* and the *exploratory database analysis* guided by specific goals. In the case of the latter, the term "exploration" is taken to mean assessing the informational value of the currently available data assets, rather than identifying any significant patterns or relationships.

The so-defined EDA process can be broken down into a number of distinct steps, with each geared toward a specific end objective (see Table 8.1).

The Exploratory Data Analysis Process

Table 8.1 The EDA Process and End Objectives

EDA Process Steps	*End Objectives*
Step 1: List the states informational needs. Step 2: Revisit the Meta Analysis.	Review of previous findings
Step 3: Assess the sustainability of the available date. • Analytic usefulness (variability)	Examine data's informational quality
Step 4: Graphically examine key variables	
Step 5: Univariate exploration: Describe the dataset.	Describe–Assess–Explain
Step 6: Bivariate exploration: Assess the strength of relationships.	
Step 7: Multivariate exploration.	
Step 8: Identify indexing and summary measures.	

The general EDA process detailed in Figure 8.1 is comprised of eight distinct stages grouped into three aggregate categories: 1. Review of previous findings; 2. Examination of the informational quality of the available data; and 3. Analyses of data (describe–assess–explain). The first of the three phases—the review of previous findings—offers an opportunity to "take inventory" of previous findings and assess their potential contributions to the current informational needs. The second phase—assessment of the informational quality of the available data—calls for a critical and thorough examination of the value of the data assets vis-à-vis the stated analytical objectives. The third and the final phase of the EDA process—analyses of data—encompasses a broad array of exploration-minded, informational needs-driven and analyst-directed investigations. As stated earlier, the overriding objective of the EDA process detailed below is the identification of data insights that can be expected to give rise to informational and ultimately, competitive advantage.

Placing exploratory data analyses detailed below in the context of the knowledge creation analytical framework presented in this book, it can be shown that both the scope and the direction of the initial database exploration are shaped by a couple of process steps discussed earlier (see Figure 8.3 above). First, specific informational objectives, the delineation of which was discussed in Chapter 3, gives the ensuing analyses a clear focus by directing attention to specific relationships. Secondly, Meta Analysis detailed in Chapter 7 offers a convenient method of assessing the potential applicability of some of the already-on-hand to the stated informational goals. Hence the remainder of this chapter will be devoted to the discussion of key considerations surrounding the reconciliation of earlier identified informational needs of the organization and the informational content of the available data. Doing so entails the bridging of the "level of aggregation" divide separating the highly operationally specific evaluations of the available data (i.e., the results of Meta Analysis) and the (relatively) more loosely defined informational need-dictated data requirements. In other words, *what are the stated informational needs-implied data requirements vs. the informational content of the available data?*

I should emphasize that in this case, both the robustness of the aforementioned comparison and its timing are quite important. That is because being able to complete a thorough "available data" vs. "stated informational goals" comparison at the onset, as opposed to somewhere in the mid-stream of data analyses will speed up the availability of important insights, while keeping the data-analyses-related costs down.

EDA Part I: Review Previous Findings

Knowledge creation is a cumulative process. Even the most eye-opening and revolutionary insights ultimately exist within the realm of other, previously created knowledge. This is particularly important when an organization intends to build a comprehensive reservoir of decision-aiding knowledge. That is because unless proper steps are taken, these insight repositories can become nothing more than collections of unconnected, random tidbits of information lacking collective power. In other words, even though some of the individual pieces of information can shed insight onto specific issues, there is little-to-no informational synergy being created by their aggregate total.

The knowledge creation process outlined in this book emphasizes a purposeful and systematic accumulation of data-derived insights, which stipulates that, on the one hand, each additional insight contributes something new to the already-in-place knowledge base, while at the same time, all individual elements are connected by an underlying theme. Recalling the earlier overview of the differences between automated data mining and analyst-guided data exploration, exploratory analyses should be shaped by both the stated informational goals (see Chapter 3) and the already-on-hand information (see the Meta Analysis section in Chapter 7). However, as suggested earlier, these two quite dissimilar elements do not present a "natural" fit, thus bringing them together can benefit from an objective evaluative framework.

The MECE Framework

The MECE framework (*m*utually *e*xclusive and *c*ollectively *e*xhaustive) is the most appropriate conceptual tool that can be used to jointly evaluate the already-on-hand information in the context of the stated information gathering goals. Its basic premise is that the best way to approach an analytical (but not necessarily quantitative) problem is by identifying—within the confines of the scope of the analysis—all independent components in such a way as to provide a maximally

complete coverage or an explanation, while avoiding double-counting. Hence the framework stipulates that each element of knowledge should be informationally non-overlapping with other ones, but collectively, all of them should exhaustively cover the informational demands of the organization's strategic objectives.

The use of the MECE evaluative framework can instill a certain amount of informational discipline by drawing attention to pieces of information that contribute the most, individually as well as collectively, to reaching the stated informational objectives. Since the importance or value of an individual piece of information can be highly situational—i.e., it can depend on what are the stated informational goals—it follows that individual data elements should not be treated as being universally important or universally unimportant. In addition, the MECE framework also promotes thoughtful accumulation of knowledge, by helping to differentiate between causally or otherwise related explanatory factors and spurious informational tidbits.

However, in order to yield robust results, the evaluative framework requires a high degree of definitional clarity. A definitionally clear informational element is one accompanied by an unambiguous explanation of its interpretative meaning along with the detailing of its measurement, or operational qualities. Specifically, it is important to express the already available insights as well as the planned ones in maximally operational terms, which includes measurement properties (i.e., continuous vs. discrete), the unit of analysis and the acceptable value ranges, all with the goal of diminishing potential misinterpretations of individual informational elements. The importance of operational specificity carries far beyond the exploratory analyses presented in this section and will become even more evident in the context of segmentation and behavioral predictions overviews presented in the ensuing chapters. In the sense of technical analyses of data, the definitional precision matters for reasons ranging from effect specification to methodological appropriateness (i.e., making sure that metrics meet the distributional and other requirements of specific statistical methodologies).

Given the requirements governing a robust application of the MECE framework, the ensuing process of reconciling the organization's stated informational objectives with the already-on-hand informational assets entails some specific considerations discussed next.

Reconciling Stated Needs and Meta Analytic Insights

The key to any comparison is the establishment of robust and objective evaluative thresholds. Keeping in mind the MECE evaluative framework, meaningful commonalities between the stated informational needs and the currently-on-hand insights (i.e., the results of Meta Analysis) require an a priori assessment of the *appropriateness of individual variables*, followed by the determination of *sufficiency of coverage of the combined variable set*. The purpose of these two preliminary steps is to make sure the scope and contents of the raw inputs to be used in the analysis are both appropriate.

Appropriateness of Individual Variables

An *appropriate* variable is one that exhibits the desired value availability (i.e., % of missing values), scaling (i.e., the type of measurement), distributional (i.e., the shape of the frequency distribution) and interpretational (i.e., meaning) characteristics. The intent behind this evaluative dimension is to ascertain that the individual measures are statistically as well as contextually useful in the context of the informational needs-dictated ensuing analyses.

A single variable can be deemed analytically appropriate if it meets the following criteria:

- The proportion of its values that are missing are within the objectively allowed limits—i.e., the previously discussed norm of not exceeding about one-third of the total number of observations.
- Its measurement scale is appropriate given its anticipated usage—specifically, the measure is continuous, if necessary (remember that continuous variables can always be re-coded into discrete values, but discrete values cannot ever be converted into continuous ones).
- Its frequency distribution meets the requirements of the statistical techniques to be used, or the desired properties can be brought about through an appropriate transformation (see Chapter 7 for details).

At the same time the said variable can be deemed informationally appropriate if it meets the following criteria:

- Its meaning falls within the general scope of the stated informational needs.
- Recency-wise, it represents the most up-to-date level of insight.

MECE CONSIDERATIONS

The individual variables deemed appropriate based on the above criteria need to contribute incrementally to the explanatory power of the entire set. Operationally, this means that individually, measures should not exhibit excessive cross-variable collinearity.[5] This is a somewhat tricky area because it imposes seemingly contradictory statistical and informational requirements. In a statistical sense, non-trivial variable correlations (usually expressed as statistically significant at a chosen confidence interval, such as 95% or 99%) are a necessary prerequisite of meaningful multivariate and cross-variable analyses, simply because absent those, the resultant models will lack any explanatory and/or predictive power. In an informational sense, however, as stipulated by the earlier discussed Occam's Razor, each individual variable is expected to make an incremental (i.e., unique) contribution to the overall explanation in order for it to not be deemed informationally redundant. In other words, while some level of variable correlation is absolutely necessary to enable explanatory analyses, too high a correlation may preclude some variables from being included in the analysis. Fortunately, steps can be taken to find an acceptable middle-ground and Chapter 10 will offer an in-depth discussion of the recommended diagnostics and remedies.

Sufficiency of Coverage of the Variable Set

The adequacy of the informational content of the entire variable set should be evaluated by considering the following two characterizations:

- The availability of multiple indicators, which translates into two or more operationally distinct but informationally related metrics. The basic idea behind this requirement is to ascertain that the underlying informational constructs are measured with multiple indicators.
- Distributional properties of specific variable subsets are aligned with individual informational needs. For instance, if the stated informational needs are focused on high value customers, the individual metrics should encompass a robust number of high-value customers-attributable values and the overall distribution of values should adhere to the requirements of statistical techniques that are to be used.

The availability of multiple metrics for each of the individual informational dimensions is important for a number of reasons. First of all, it is necessary for explaining or predicting a particular phenomenon, since any such analyses are multivariate in nature (e.g., regression

analyses require two or more variables, one to be the target and one or more to be predictors). In other situations, multiple metrics are required to ascertain the construct's convergent validity, which is the degree to which multiple indicators produce similar results, lending credence to subsequent interpretations.

Assessing distributional properties of individual variable cohorts is important primarily from the standpoint of technical data analysis. Many of the commonly used statistical techniques, such as Pearson's product-moment correlation, regression or analysis of variance, require at least some of the input variables to be continuous and normally distributed. Hence it is important that the metrics available to support individual informational needs exhibit properties required by statistical techniques that are best suited to generate the desired insights.

MECE CONSIDERATIONS

As is the case with other aggregate evaluative schemas, such as SWAT analysis, there is a certain subjectivity that is inherent to the MECE framework. It is another way of saying that effort should be taken to minimize any potential analyst bias. An obvious step that can be taken in that regard is the cross-analyst triangulation, akin to the Delphi method. Multiple analysts should each come up with an independent assessment, following which they should be shown the results of other analysts' conclusions and given an opportunity to revise their own. Several iterations should produce convergence or near-convergence and in case of the latter, an open forum discussion can be used as the mean of reaching the final solution.

EDA Part II: Assess the Informational Quality of Data

It is important to look beyond "what are the available metrics," to "what would be the ideal metrics," given the demands of the informational needs at hand. In other words, how adequate is the currently available variable set?

In practice it is rare for the data currently-on-hand to leave nothing to be desired. More often than not, the informational needs pose questions that go beyond the currently available metrics, implying other data. For instance, customer acquisition and retention rates of the firm's competitors, or competing brands' production costs or sales margins are rarely, if ever, obtainable. However, other types of insights may be embedded in the currently available raw data, though not expressly listed among the current metrics. Instead, it may be derived from the already-on-hand data either by *indexing* or *combinatorial* means. In other words, some of the already available raw metrics can be used to create new raw metrics, which then can be added into the currently-on-hand dataset. In effect, the amount of data, in terms of the raw metric count as well as the informational scope that is available at any given time can be increased with the help of some carefully thought-out steps. As a matter of fact, as shown in Figure 8.4, the gap separating the current and the ideal variable sets can be narrowed through intelligent harvesting of the not-immediately-evident informational content of the available raw data.

Current variable indexing is a process of re-expressing raw metrics in more immediately usable descriptive qualities, as exemplified by value indicators (e.g., high, medium, low) derived from numeric sales data or campaign response "flags" (e.g., yes vs. no) inferred from treatment-attributable sales metrics. It is important to keep in mind that indexing frequently takes the form of discrete indicators, thus to the degree to which it might be possible, the appropriateness of the

qualitative measurement scale should be considered in the context of the contemplated analyses prior to committing to a particular measurement scale.

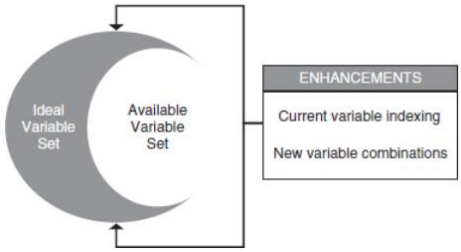


Figure 8.4 Desired vs. Available Data

On the other hand, as implied by its name, *new variable combination* simply refers to creating brand new measures through the joining of two or more of the existing ones. In the informational sense this amounts to developing higher-order insights from more disaggregate raw components, as exemplified by combining individual Likert-type survey statements (usually expressed as observable indicators of the underlying latent construct) into construct-level measures. Alternatively, somewhat dissimilar metrics can be combined into new measures yielding different insights. For instance, buyer- or household-level price elasticity measure can be computed by combining the purchase and pricing details. In contrast to the previously discussed indexing, newly created variables are typically more likely to be continually distributed, leading to fewer usage restrictions.

Graphical Examination of Data

It is important to not lose sight of the fact that not all variables in a particular dataset are equally important from the informational content standpoint. In a more technical sense, some metrics that are informationally important could lack statistical robustness. Specifically, they could manifest poor measurement precision or inadequate amount of variability, both in univariate and multivariate contexts. Thus any dataset evaluation should be guided by a dual objective: First, to differentiate between theoretically critical and the lesser theoretically impactful metrics; and second, to assess the statistical quality of the crucial pieces of data.

Always a good time-saving idea, this type of pre-analysis due diligence may be a necessity when dealing with transactional databases, often populated by a staggering number and variety of metrics. The determinants of what constitutes an "important" vs. "less important" metric are obviously highly situational, though in general any data element deemed essential to understanding behavioral or attitudinal outcomes, such as drivers of purchase/repurchase or influencers of attitudes, or those explaining cross-group differences will almost always be of critical importance. On the other hand, variables offering primarily profiling or descriptive insights, such demographics, are usually less important, largely because—as shown in later chapters—their explanatory and predictive power tends to be considerably lower.

Again, the key reason for differentiating between the more and less critical metrics is to focus the attention on the former to enable an efficient assessment of their basic statistical qualities. In other words, do these important variables exhibit the desired statistical properties, given the

informational needs at hand and the contemplated analytical methodologies? The quickest and perhaps the easiest method of answering such questions is through graphical examination of the data. Made possible (and quite easy) by the proliferation of powerful statistical analysis packages designed for personal computers, carefully selected data graphs can lead to a quick "thumbs up" or "thumbs down" assessment of the individual metrics. I should point out that graphical description of data is certainly not limited to univariate (i.e., single variable) depictions, but it is arguably most potent in those cases, as higher-order contrasts (i.e., multivariable relationships) are more effectively evaluated with the help of mathematical tests.

The starting point in assessment of any variable should be the characterization of the shape of its distribution, because as discussed later, many key statistical procedures are built around certain distributional assumptions, most frequently, the ubiquitous standard normal distribution. The easiest approach to assessing univariate distribution is to simply graph it. The cleanest (and likely the easiest) way of doing so is through the use of a histogram, which represents the frequencies of occurrences of data values within categories. Figure 8.5 shows an example of a histogram. The individual bars represent frequency counts, thus the taller the bar, the higher the count of that particular value.

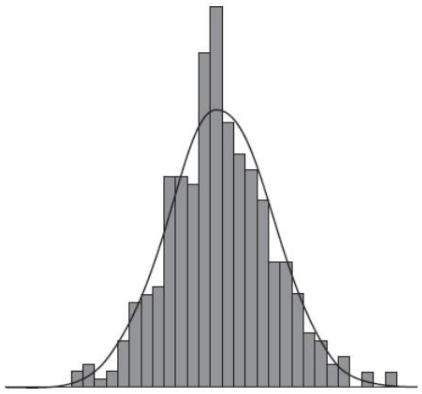


Figure 8.5 Sample Histogram

To aid in the assessment of the "normality" of a particular univariate distribution, it is helpful to add to the barred histogram an outline of the normal distribution, shown as the darker-colored bell-shape curve in Figure 8.5. Technically, the normal curve represents the expected distribution, to be contrasted with observed valued depicted by the individual bars.[7] In the above example, the distribution appears to be approximately normal. The operative term here is "approximately." It is rare to come across a distribution that has exactly the shape shown by the aforementioned bell curve and, more importantly, it is not statistically necessary. Most of the commonly used techniques, such as regression and other GLM[8] methods discussed in subsequent chapters are relatively robust with regard to some departure from normality.

Of course, as the difference between the ideal and actual distributions grows, the robustness of the analytic results will tend to diminish, which means that it is important to single out a relatively "hard" threshold beyond which a particular variable should be no longer considered normally distributed. Unfortunately, a histogram-based visual distributional assessment leaves that determination to the more-or-less qualitative judgment of an analyst. However, given the

importance of accurately discerning the distributional qualities of at least the key database metrics, if a particular distribution raises doubts regarding its shape, one of the available numeric goodness-of-fit tests should be used to assess its level of normality. The x^2 (chi squared) test is among the easiest to implement univariate normality diagnostics.

The chi squared test measures the goodness-of-fit between the *observed* and *expected* (under the normal distribution) frequencies, as depicted in Figure 8.5. In other words, for every bar (observed), which represents a single data category, and line (expected) intersection, the test determines if any differences represent persistent dissimilarity or are mere random, non-systematic fluctuations. The tests requires the ability to compute the cumulative distribution function, which is the probability that the variable takes on a value that is less than or equal to the expected value. Its computation, however, is relatively straightforward:

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where,

O_i is the observed frequency for category i
E_i is the expected frequency for category i

The *observed* frequency reflects the actual counts contained in the data: On the other hand, the *expected* frequency is computed as follows:

$$E_i = N(F(Y_u) - F(Y_l))$$

where,

F is the cumulative distribution function for the distribution being tested
Y_u is the upper limit for class i
Y_l is the lower limit for class i
N is the sample size

The results of the χ^2 test are interpreted in the familiar *p*-value expressed significance test format, as the test is defined for the following general hypotheses:

H_o: The data follow normal distribution.
H_a: The data do not follow normal distribution.

A statistically significant result leads to the rejection of the null (H_o) hypothesis and a corresponding acceptance of the alternative (H_a) hypothesis, which ultimately leads to the conclusion that the variable of interest is not normally distributed.

It should be noted that there are multiple other tests that have been developed and used to detect departures from normality, the best known of which include the Kolmogorov-Smirnov and Anderson-Darling operationalizations. In general, though, these tests adhere to a more-or-less similar evaluative logic. In a statistical sense, the latter, which is used frequently in the financial sector, is a modification of the former, giving more weight to the tail-end of the distribution than its Kolmogorov-Smirnov parent. Unlike the chi squared test discussed above, however, these two

tests are restricted to continuous distributions, thus not appropriate to use as a numerical extension of histogram-based univariate distribution evaluation.

EDA Part III: Describe–Assess–Explain

Although data visualization is gaining in popularity, the bulk of the more in-depth data exploration and virtually all of the more complex hypotheses testing remains numeric. All too often, however, the numerical data analyses do not go far enough in translating the usually somewhat abstract results into more meaningful insights. Many analysts find it difficult to look beyond *how* a particular analysis was carried out to *what* the results mean to the organization. The appropriateness and robustness of employed methodologies will certainly always be of critical importance, but to users, the clarity of results in terms of business implications will carry a lot more meaning. In short, in order for results of analyses to yield highly impactful and advantageous knowledge, analytic findings cannot be confounded with analytic methodologies in a way that impedes users' understanding.

These considerations are particularly important in the context of the actual (exploratory) data analytical continuum of *describe–assess–explain* described next. Although throughout this section, as well as the rest of this book, a number of computational formulas are presented, the overall objective is to develop an intuitive level of understanding of the subject matter being discussed, to enable one to move beyond technical details in presenting their findings. Hence the scattering of technical details is geared toward the identification of the most effective methods of solving a particular problem, with the ultimate goal of bringing to bear the most insightful and advantageous knowledge. An added benefit of this approach is that the ensuing discussion will not get bogged down in an encyclopedic delineation of every conceivable approach available, but instead will be focused on the few that have been shown (in practical business applications) to generate the most valid and reliable outcomes.

Unlike the previously discussed (two) steps of the EDA process which focused on preparatory steps, this part of the EDA process is concerned with the exploration of the available data. In terms of the *describe–assess–explain* continuum: The *describe* part is focused on revealing the basic facts about the data, such as the average customer spending level or the frequency of store visit. The second element, *assess*, captures the goal of developing a fundamental understanding of patterns of relationship in the data. Stated in different terms, it conveys more informationally meaningful insights as it represents the first step of evolving beyond merely detailing of *what-is*, toward *why-it-is*. For instance, does promotional spending coincide with sales gains? Or , is the customer spending related to geography or household characteristics?

Lastly, the *explain* part of the continuum is the informationally richest source of insights as it brings to bear the final building blocks that are needed to transform the more-or-less generic information into unique and thus far more valuable knowledge. In practical business terms, it objectively addresses the reasons behind either the observed failure or the success of a particular action, which fosters a more effective future deployment of organizational resources.

Describe: Univariate Distributional Properties

A part of the Data Compilation and Evaluation process discussed in Chapter 7 was examining—and correcting when appropriate—individual variables' *skewness* and *kurtosis*. Additionally, the initial preparation steps also involved filling-in missing values, as necessary. The end objective of

these steps was to arrive at a fully populated (i.e., no missing values) dataset, which also exhibits desired distributional properties.

Data exploration necessitates examining the resultant cleansed data in the context of their informational content. Doing so entails computing several basic univariate descriptors, all of which is necessary to validly describe the informational content and analytic usability of the individual metrics. The choice of the descriptive metrics is primarily a function of the individual variable's measurement scale, specifically, whether it is discrete or continuous.

Discretely coded data, often referred to as *categorical* or *qualitative* (the latter not to be confused with qualitative research or conclusions) assumes only values that represent distinct categories, such as integers, as exemplified by the number of children in a household, the number of UPCs making up a brand, or the number of brand purchases. For instance, a person can have any integer-expressed number of children including 0, but cannot have 2.5 or 4.1 children, just as a brand can have a varying number of specific UPCs (e.g., individual size/flavoring soft-drink alternatives), so long as that number is not fractional. In the vast majority of applied business situations, discrete variables have a finite number of values, as implied by the above examples (although that is not theoretically required).

On the other hand, a *continuous variable* is one that takes on an infinite number of possible values, usually bounded by two extremes.[9] Thus in contrast to categorical metrics where only integer-based values are permissible, here, any value is possible, so long as it falls between the range-defining end points, usually referred to as the maximum and minimum values. Common examples include sale revenues; individuals' age, weight or wealth; or even derived metrics, such as a product repurchase propensity or offer response likelihood. In a practical sense, these otherwise advantageous (from the standpoint of data analyses) basic properties of continuous variables can potentially complicate the interpretation of results, particularly in the sense of differentiating between statistically and practically significant findings (an in-depth discussion of this topic is presented below in the *Beware of Significance Tests* section).

In view of the differences between their respective distributional properties, it is intuitively obvious that the two cannot be subjected to the same types of mathematical manipulations, such as division or multiplication. As a result, the discretely coded data presents far fewer analytic options than does its continuously coded counterpart, which naturally limits its informational value. (Keep in mind that, as previously mentioned, continuous metrics can always be re-coded into categorical ones, but the reverse is not possible. It is important to remember, however, that any continuous-to-categorical conversion will lead to an irreversible loss of information as a direct result of diminished variability of the resultant data.) The practical consequences of the impermissibility of some basic algebraic operations, such as division required to compute average values, can be a source of significant impediments to knowledge creation. That said, careful due diligence is urged when contemplating data capture or variable coding and when possible, preference should be given to continuous measurement.

In view of this fundamental computational schism separating the discretely and continuously coded metrics, different statistical procedures are available to analyze descriptive characteristics. In general, discrete variables are described in terms of *counts* and *frequencies*, while continuous variables are best characterized with the help of measures of *central tendency*, a departure from the average and the range, which measures the spread between the smallest and the largest values.

Discrete Variables

Considering that discrete variables are comprised of a finite—and usually manageable—number of distinct categories, the most effective univariate analytical method involves numeric or graphical (such as the histogram depicted in Figure 8.5) frequency distribution review. However, just because the number of categories is finite, does not mean that it is small. For instance, a brand is typically a categorical (specifically, nominal)metric, yet in many instances, the number of brands in a dataset could be quite large (e.g., ready-to-eat breakfast cereal). Either numeric or graphical frequency distribution will yield a large number of categories making the output somewhat difficult to digest.

An easy corrective step to take is to create groupings of categories. Using the breakfast cereal example cited above, it would be advisable to lump a number of the smaller brands into somewhat homogenous categories, while keeping the largest (i.e., biggest-selling) brands as separate entities. An added benefit of taking that step is that it will help to limit the amount of noise in the data, since low frequency categories contribute disproportionately more to the unexplained variability in the dataset then they do to the explanatory power of most statistical models. In addition, as discussed in Chapters 6 and 7, some statistical techniques may require that the categorical variables be re-coded into *dummy variables*, which in effect converts a single, multi-category metric into multiple binary ones, where the number of resultant metrics is equal to the original number of categories. It is not only tedious, but likely will diminish the parsimony of the resultant statistical model.

Continuous Variables

Continuous variables, being informationally richer, offer more univariate analytical depth. However, the most important insights can be gleaned from the analysis of their central tendencies, variability and dispersion.

MEASURES OF CENTRAL TENDENCY

There are three commonly used categories of measures of univariate central tendency : the average, the median and the mode.[10]

The *average* depicts a typical value in a particular distribution. Computationally, there are multiple ways of calculating averages, with the *mean* being the most commonly used one. Technically, mean can be arithmetic, geometric, harmonic or weighted. The *arithmetic mean*, also referred to as a "straight" or a "simple" mean, is obtained by summing two or more values and dividing the resultant by the number of items. The *geometric mean* is defined as the n^{th} root of the product of all values in a set of numerical data, where n is the number of values in the dataset. Although somewhat more computationally involved, the geometric mean is more resistant to extreme values than the more interpretationally straightforward arithmetic mean. The *harmonic mean* is yet another method of computing the average while minimizing the influence of extreme values and it is defined as the quotient of n divided by the sum of the reciprocals of all the values in a set of numerical data, where n is the number of values in the dataset.

Most statistical applications rely on a simple arithmetic mean, computed as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where,

x_i = an individual observation
n = number of cases

Median and *mode* are the two alternatives to the mean as the measure of average or typical values. A *median* is the middle number in a set of ordered data, usually found with a simple formula:

$$(n + 1) / 2$$

where,

n = number of records

In the event a sample contains an odd number of records, the median of that sample will be one of the actual values contained herein; otherwise, if there happens to be an even number of records in a particular sample, the resultant median will be computed as the mean of the two middle values.

The last of the three measures of central tendencies is the *mode*, which is simply the actually observed value that happens to appear in the largest number of records, or put differently, it is an observed value with the highest frequency of occurrence. In contrast to both the mean and the median, the mode is likely to not be unique.

MEASURES OF VARIABILITY

In a statistical sense, the best way to think about any central tendency metric is to consider it to be the expected value of a random variable. A particular variable will, theoretically, at random take on different values and if one were to guess the most likely value that is to be assumed by a random variable, the "best guess" would be the previously discussed measure of central tendency—the mean (hence in statistical analyses, the mean is often characterized as the expected value of a random variable). Given that, if one were to compute the difference between the mean of a particular variable (i.e., the expected value) and each actual value, the difference (or the residual, expressed as actual value-mean value), aggregated across all records would yield the measure of variability contained in the data, in relation to the variable of interest. Hence, the amount of variability in the data can be simply expressed as:

$$\text{Variability} = \sum_{i=1}^{n} (observed - mean)$$

However, there is a caveat here: Given that the mean will fall more-or-less in the middle of the distribution, roughly half the values will be larger than it and half smaller, thus if added, the positive and negative deviations from the mean will cancel each other out, all of which would result in zero (0) variability (or a very small value close to 0). Simply squaring both the positive and negative values will eliminate the canceling effect and lead to a real number-based

variability estimate, which offers a convenient way of circumventing the apparent computational flaw. The resultant measure is called *variance*, which is denoted as s^2 and computed as follows[11]:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

where,

x_i = an individual observation
x = mean
n = number of cases

The most practically compelling interpretation of variance is that it is indicative of the explanatory power of a particular variable: The larger the variance, the more likely is the variable to contribute to the resultant explanation. That is because large variance is indicative of significant cross-record differences, which is a necessary prerequisite for any metric to have strong predictive power. In addition, variance also gives rise to another useful statistic—the *standard deviation*—which is a more effective way of comparing the levels of variability across variables. Methodologically, standard deviation is simply the square root of variance, denoted by s[9] and computed as follows[12]:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

where,

x_i = an individual observation
x = mean
n = number of cases

Standard deviation is an important metric because it illustrates the amount of variability contained in a particular variable—the larger the standard deviation, the more variance there is in a particular variable, across cases. Variables with very small standard deviation are informationally poor because they contribute little to the cross case differentiation (this is in contrast to parameter estimates, where small standard deviation is desirable as it is an indication of stable estimates). As was shown in Chapter 7, the low amount of variance can significantly reduce the predictive or explanatory power of otherwise intuitively important metrics. However, because its computation involves squaring of the residual ($x_i - x$), both the standard deviation as well as its "parent," the variance is particularly affected by extreme values, which can be a cause for concern, but also underscores the importance of solid data due diligence efforts.

MEASURES OF DISPERSION

The last key univariate distributional characteristic is the range, which is simply the spread between the smallest and largest values. Again, because it measures the absolute distance between

the extremes, it can be more informative to use an interval range, such as the frequently used quartiles. Quartiles are points that divide an ordered distribution into four parts, each containing one quarter of all scores. Examining the percentage of all cases falling into each of the four quartiles sheds additional light on the magnitude of spread depicted by the range. A large range accompanied by a condensed quartile distribution (i.e., very few cases falling into quartiles 1 and 4) is informationally poorer than an equally large range accompanied by a more evenly balanced quartile distribution. Any univariate analysis usually requires a relatively tedious variable-by-variable examination of the descriptive characteristics of data, such as the average or the range of values, which tends to narrow the focus to a relatively small subset of all available data. As previously suggested the choice of variables to be included in the analysis subset is largely subjective and driven by the analyst's level of experience. When it comes to transactional databases, there tends to be a short list of candidate metrics, including current customers' spending levels or profitability, product mix, price and promotional elasticity and repurchase frequency. The good news is that most of the important metrics will usually be included in the analysis. The bad news, however, is that it is difficult to come across new, competitively advantageous insights with everyone glued to the same set of metrics.

Even more importantly, univariate data analysis is outcome-oriented, which means it can lead to a virtual avalanche of status quo tidbits while offering little-to-no explanatory diagnostics. As a matter of fact, its implicit assumption of individual metric orthogonality, or independence, means that cross-variable comparisons can be potentially misleading as otherwise spurious associations are interpreted as cause and effect. For instance, a side-by-side comparison of seemingly related factors, such as promotional spending and sales, can suggest that an increase in the latter was caused by the former, where in fact sales gains could have been driven by a different—and unaccounted for—set of factors, such as pricing or distribution changes. This means that while univariate analysis can shed light on *what-is*, it contributes relatively little to the creation of a competitively advantageous knowledge base because it cannot offer reliable insights into the reasons behind the observed outcomes.

Assess: Bivariate Relationships

Potential relationships first suggested by univariate profiling should be initially investigated with the help of bivariate analyses, which assess the persistence and generalizability of pairwise relationships. For instance, a concurrent increase in the levels of sales and promotional spending may be indicative of certain promotions' sales impact, or the observed increase in the frequency of product repurchase seemingly accompanying coupon usage may be indicative of an apparent relationship between the two. In other words, analyses of individual variables may reveal apparent relationships between pairs of variables.

Although methodologically straightforward, analyses of associations can be riddled with hidden traps, such as the *Simpson's paradox*, according to which the direction or strength of the relationship changes when data is aggregated across natural groupings that should otherwise be treated separately. In other words, one should be cautious when searching for "globally valid" relationships—it might be better to assess multiple associations framed in the context of a more homogenized population. Furthermore, few if any databases offer anything other than a subset of all possible data that might be related to a particular phenomenon, in addition to which, the available data may vary in terms of its accuracy. Furthermore, a threat that is particularly potent in the context of bivariate analyses is the potential presence of *intervening* or *moderating factors*,

where an outside variable moderates a particular bivariate or conditional relationship (where the strength or direction of the relationship changes across the values of a third variable).

Keeping an eye on these and other potential traps, there are two general approaches to quantifying bivariate relationships:

1. Simple cross-tabulation.
2. Correlation analysis.

When both variables are discrete, cross tabulation—or cross tab for short—is usually the easiest method of assessing the relationship between the two. On the other hand, when both measures are continuous, quantifying their correlation is the most effective method of assessing their relationship. In fact, a correlation coefficient can be computed for virtually all random variables, regardless of their scaling properties. However, even with the proliferation of powerful analytic software applications, correlating discrete variables can be tricky, in view of the multiplicity of esoteric and rarely used tests, so much so that a robust cross tabulation can be quite a viable alternative.

Cross Tabulating Discrete Variables

Bivariate χ^2 (chi square) analysis is the simplest approach to quantifying bivariate relationships with the help of cross tabulation. Since cross tabs involve the construction of matrices, where variables are usually expressed as columns and individual categories as rows, it is obviously advantageous to keep the number of categories relatively low[13] (at this point, we are only concerned with bivariate, or 2 × 2 designs, thus the number of variables would obviously always remain low). The goal of the test is simple: To determine if there are non-spurious associations between the specified variables. It is important to point out that the test is binary—i.e., it can confirm or reject the presence of relationships, but it yields no information regarding the strength of the association. However, as shown below, there are supplemental methods of discerning that information.

In a statistical sense, χ^2 is a nonparametric test, which means that it places no distributional requirements on the sample data. That said, however, in order to yield unbiased estimates, the test requires the sample to be random, the data to be reported as raw frequencies (not percentages), the individual categories to be mutually exclusive and exhaustive and the observed frequencies cannot be too small (the often-cited rule of thumb calls for a minimum cell[14] size of 5 frequencies—in practice, however, a minimum sample size of 50 is more reasonable to assure the robustness of business analyses). The test itself is relatively simple—it compares the difference between the observed and expected frequencies in each cell to determine if significant patterns of similarities (i.e., a relationship) exist between the two variables; it is computed as follows:

$$x^2 = \sum \frac{(O - E)^2}{E}$$

where,

O is the observed value
E is the expected value

The χ^2 test will determine if any two variables are related—however, it is not indicative of the strength of the relationship. A separate measure—Cramer's V—was developed for cross

tabulations larger than 2 × 2 to quantify the strength of χ^2-significant relationships. That measure is computed as follows:

$$\text{Cramer's V} = \sqrt{\frac{x^2}{N(\kappa - 1)}}$$

where,

χ^2 is the value of the previously computed chi square statistic
N is the total number of observations
κ is the smaller of the number of rows or columns

Thus in situations where there are more than five cells (i.e., at least one of the variables has more than two categories) the strength of the relationship can be computed. The resultant Cramer's V coefficient is interpreted the same way as the commonly used Pearson's product-moment correlation coefficient, discussed next. Of course, since its computation involves the taking of a square root, the phi values will always be positive, ranging from 0 to +1.

Bivariate Analyses of Mixed Variables

As detailed above, the informationally richer continuous data yields more breadth of bivariate analytical insights. A somewhat more methodologically challenging situation is when one of the variables is continuous while the other is discrete. There are different methods available for assessing bivariate relationships of continuously measured and mixed (continuous + discrete) pairwise variable comparisons, all of which fall under the general umbrella of *correlation analysis* (not to be confused with the earlier mentioned Pearson product moment correlation, which is a special case of bivariate correlation where both variables are continuous).

The term "correlation" has a rather broad usage. Aside from bivariate correlation mentioned above, there are several other expressions of correlation-based associations. Regression analysis, discussed in Chapter 8, produces a measure of *multiple correlations*, which is the correlation of multiple independent variables with a single dependent variable. In addition, there is the *partial correlation*, which is the measure of association of two variables controlling for the impact of other variables; there is also a similarly named *part correlation*, which is similar to partial correlation except that the impact of other variables is only controlled for one of the two correlation measures. Lastly, there is also the *canonical correlation*, discussed in more detail below. All of these are special purpose variants of the general bivariate correlation and will be discussed throughout this book as appropriate. However, in this chapter the focus is on the general discussion of the bivariate correlation.

Technically, *correlation* is a concurrent change in value of two numeric variables, which represents the degree of relationship between those variables. As it is intuitively obvious, the correlation-expressed relationship can be either positive, when an increase in one is accompanied by an increase in the other, or negative, which is when an increase in one is accompanied by a decrease in the other. A numeric result of correlation analyses is a correlation coefficient, which is a metric expressing the strength of the relationship between the two variables of interest, ranging from +1 (perfectly positive correlation) to –1 (perfectly negative correlation) and centered on 0, which denotes a lack of relationship. In the context of knowledge creation, correlation represents

an informational improvement over univariate analyses because it begins to explain the phenomenon of interest, which goes a step beyond just summarizing the observed status quo.

In spite of its widespread usage, there is a healthy amount of confusion surrounding the notion of correlation analysis. The bulk of the misunderstanding centers on the scope, or more specifically, the number of items being correlated; a close second is the choice of a specific formulation—i.e., a correlation coefficient.

Overall, correlation is bivariate, which means that it can only be computed for two entities at a time, as a simultaneous assessment of multivariable (i.e., 2+) relationships would be methodologically complex and practically limiting, primarily because it would necessitate the use of conditional expressions and/or interaction terms.[15] That said, it is important to point out that correlations can be computed for a set of two individual variables, as well as for two sets of variables (i.e., for correlation purposes, an entity can have very specific operationalization or it can represent a summary). Both cases, conceptually speaking, will result in bivariate relationships as there are ultimately only two entities involved, yet in the methodological sense there are two distinctly different approaches that need to be employed.

In the case of two individual variables being correlated, an approach generally described as the *bivariate correlation* should be used, while a different methodology known as the *canonical correlation* should be employed with summary-based (i.e., the previously mentioned sets of metrics) operationalizations. It should be noted, however, that in practice the use of canonical correlations is relatively infrequent in business analyses and virtually non-existent in analyses of large transactional databases. The reason for that is that since the canonical correlation is primarily of value in quantifying relationships between two sets of metrics, where each set is intended to measure the same underlying (and usually unobservable, in and of itself) construct,[16] this type of analysis will obviously be of little value to database analytics focused primarily on transactional data. Thus beyond this brief mention, canonical correlation will not be discussed here any further and the term "bivariate correlation" will denote a relationship between two individual variables only.

In terms of specific formulations there are multiple methods of computing correlation coefficients, the bulk of which were devised to address specific data requirements. Table 8.2 offers an enumeration of the different coefficient types.

Largely due to computational convenience, bivariate correlation coefficients are typically computed for multiple pairs of relationships and presented in a matrix format, yielding an efficient and succinct presentation format and one that encourages further explorations and comparisons. In addition, unlike covariances, which are expressed in the original units of measurement, correlations are standardized, i.e., the original units of measurement are replaced with mean = 0 and standard deviation = 1, which makes coefficients directly comparable in spite of any original scale differences.[17]

Figure 8.6 shows an example of a simple correlation matrix. The intersection of a specific row and a column pinpoints a correlation coefficient computed for that particular pair of variables. For instance, the value for factor_3 and factor_4 is θ.535, which is a moderately strong positive correlation. As noted above, the same variables appear in the rows and columns of the matrix, which means that the diagonal values represent correlations of individual variables with themselves, which will always be equal to 1 since a variable is perfectly correlated with itself.[18] However, this means that a half of the matrix shown in Figure 8.6 is redundant because correlations are non-directional, i.e., factor_1–factor_2 correlation is functionally the same as factor_2–factor_1 correlation. In other words, it would suffice to only show values above or below

the diagonal, as depicted in Figure 8.7, which makes the matrix appear less busy thus making it easier to visually examine it, particularly when a larger number of variables are included.

Table 8.2 Coefficients of Correlation

Correlation Measure	*Description*	*Application*
Pearson's product-moment (r)	Both variables are continuous and normally distributed; relationship is linear	The most commonly used formulation
Spearman's rank (rho)	Both variables ordinal or one ordinal and one interval	The most commonly used substitute for Pearson's r
Kendall's rank (tau)	Both variables ordinal or one ordinal and one interval	A less frequently used alternative to Spearman's rho
Polyserial	Interval and ordinal variables (3+ categories) and the latter reflects underlying continuum; bivariate normality required	The preferred method used to correlate interval and multichotomous ordinal variables
Polychoric	Both variables are dichotomous or ordinal, but reflect underlying continuous variables; bivariate normality required	The preferred method to correlte two dichotomous or ordinal variables
Biserial	Same as polyserial, but the discrete variable is dichotomous	Mostly in theoretical research employing structural equation modeling
Rank biserial	An ordinal variable is related to a truly dichotomous variable (no underlying continuity)	Rarely used in practical research
Point biserial	An interval is correlated with a truly dichotomous (no underlying continuity)	Can use Pearson's r formular
Phi	Both variables are dichotomous	A substitute for Pearson's r used with dichotomies

	factor_1	factor_2	factor_3	factor_4	factor_5	factor_6	factor_7
factor_1	1	.987	−.166	−.260	−.072	−.312	.797
factor_2	.987	1	−.153	−.242	−.054	−.329	.798
factor_3	−.166	−.153	1	.535	.199	−.139	−.113
factor_4	−.260	−.242	.535	1	.268	−.142	−.213
factor_5	−.072	−.054	.199	.268	1	−.022	.013
factor_6	−.312	−.329	−.139	−.142	−.022	1	−.362
factor_7	.797	.798	−.113	−.213	.013	−.362	1

Figure 8.6 A Correlation Matrix

	factor_1	factor_2	factor_3	factor_4	factor_5	factor_6	factor_7
factor_1	1						
factor_2	.987	1					
factor_3	−.166	−.153	1				
factor_4	−.260	−.242	.535	1			
factor_5	−.072	−.054	.199	.268	1		
factor_6	−.312	−.329	−.139	−.142	−.022	1	
factor_7	.797	.798	−.113	−.213	.013	−.362	1

Figure 8.7 A Correlation Matrix: Non-Redundant Elements Only

However, as pointed out earlier, there is more to computing a correlation coefficient than choosing between a bivariate and a canonical correlation. The proliferation of the "point and click" computing capabilities has the unfortunate side effect of fuzzying the distinctiveness among the different correlation coefficient formulations. By far the most commonly used formulation—Person's product-moment correlation coefficient—tends to be the default in popular statistical packages, such as SAS or SPSS, but it is certainly not the only formulation and even more importantly, it carries specific data distributional (normal vs. non-normal) and relationship type (i.e., linear vs. non-linear) requirements, the violation of which will significantly limit the reliability of the resultant statistic. The two other bivariate correlation coefficients—Spearman's and Kendall's rank correlations—do not make specific data or relationship type requirements, which makes them suitable substitutes under certain circumstances.

Somewhat complicating the picture are the mixed-scale correlations, particularly where one variable is measured on a metric scale (i.e., interval or ratio) while the other one is measured on a non-metric scale (i.e., nominal or ordinal). There are two approaches to dealing with such situations:

1. Re-code the metric into a non-metric variable and use Spearman's rank correlation coefficient if the result is an ordinally scaled variable and use the polychoric correlation with dichotomies (see Table 8.2). This takes advantage of the fact that continuously measured variables are informationally richer, which means they can always be reduced into categorical ones, simply by breaking out their continuous values into discrete ranges. Of course, the re-coding process tends to be arbitrary since most continuous scales do not have natural discrete break points.
2. The second approach is to replace the product-moment correlation with amended formulations which account for scale differences. As shown in Table 8.2, there are multiple coefficients available: the *biserial, polyserial, polychoric, point biserial, phi*, etc. In general, the choice of the appropriate formulation is primarily a function of the type of measurement scale and its constancy between the variables being correlated. Specifically, different computational methods should be used when both variables have the same scale characteristics—such as both are ordinal or nominal—versus when their measurement scales are different, such as one is

ordinal and the other is nominal. Figure 8.8 below offers a simple decision rule to be used when choosing among the available correlation formulations.

To select an appropriate formulation, start out by identifying the measurement scale of each of the variables to be correlated. As discussed earlier, a random variable can be nominal, ordinal, interval or ratio. Nominally scaled variables carry no ordering or magnitudinal information whatsoever—they are simply labels intended primarily for convenience. Although their informational value is quite limited, the biserial correlation coefficient can be used to quantify their relations to a metrically measured variable. An ordinal scale is informationally richer as here data points are rank-ordered, although it is still limited insofar as it contains no information about the cross-category spacing (i.e., spacing is not assumed to be equal or have any other numerical properties). An even richer source of information is the interval ratio, which in addition to being rank-ordered is also assumed to be equally spaced (i.e., the distance between adjoining pairs of values is constant across the entire value continuum, which means that the measurement distance between values 1 and 2 is the same as the distance between values 2 and 3, 4 and 5, etc.). Lastly, the ratio scale contains all of the informational characteristics of the other three scales, in addition to which it also has a rational point of origin, such as age or income.

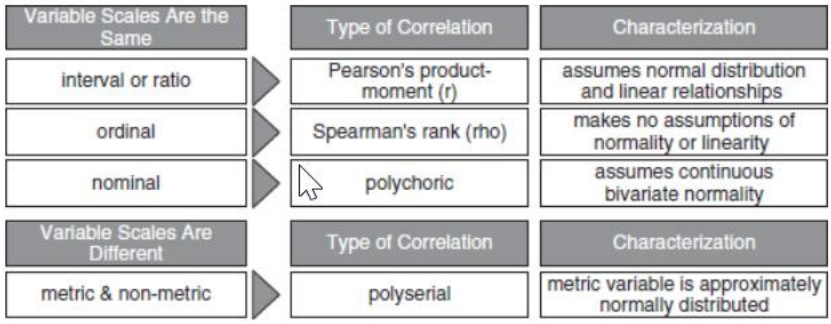


Figure 8.8 Correlation Coefficient Types

If both variables are measured with either an interval or a ratio scale (they could both be the same, or one interval *and* the other ratio) and their distributions are approximately normal *and* their relationship is assumed to be more-or-less linear, Pearson's product-moment correlation will yield the most robust estimate of the their relationship. If, on the other hand, either of those conditions is not met—i.e., the variables are not either interval or ratio, at least one is not normally distributed or their relationship is believed to not be linear—Spearman's rank correlation is the appropriate choice.[19] An example of a typical output (generated with the help of SPSS) depicting a correlation matrix, using Pearson's product-moment method is shown below.

As shown in Figure 8.9, aside from the correlation coefficient itself there are several other pieces of information included in the output, all playing a distinct though somewhat different role in the evaluation of the correlation results.

The first is the effective sample size. An *effective sample size* is the actual number of cases used in the particular analysis, which is contrasted with a *nominal sample size*, which is the total number of cases in the dataset. Under certain circumstances, most notably a persistent missing value problem, the effective sample size can be quite smaller than the nominal one, which at some point may diminish the robustness of the findings. What then is the minimum acceptable sample size?

There is no single concrete minimum, as that is usually dependent on multiple factors, most importantly the amount of variation in the data. That said, the best general guideline to minimum sample sizing can be derived from the Central Limit Theorem, which states that whenever a random sample is taken from any distribution, the sample means will be approximately normally distributed, which seems to imply that beyond a certain point, sample size expansion may not be necessary. The proverbial $64,000 question is, of course, what is that threshold? As a general rule of thumb, it is believed that fewer than about 30 observations calls for nonparametric analysis, while more than 30 but fewer than 50 observations should be treated with caution. In other words, a sample size of as few as 50 records may be sufficient. As previously discussed, normal distribution is a requirement of Pearson's product-moment correlation; hence attaining an appropriately sized sample is important to the validity of the statistic.

Correlations

		factor_1	factor_2	factor_3	factor_4	factor_5	factor_6	factor_7
factor_1	Pearson Correlation	1	.987**	−.166*	−.260**	−.072	−.312**	.797**
	Sig. (2-tailed)		.000	.011	.000	.272	.000	.000
	N	235	235	235	235	235	235	235
factor_2	Pearson Correlation	.987**	1	−.153*	−.242**	−.054	−.329**	.798**
	Sig. (2-tailed)	.000		.019	.000	.408	.000	.000
	N	235	235	235	235	235	235	235
factor_3	Pearson Correlation	−.166*	−.153*	1	.535**	.199**	−.139*	−.113
	Sig. (2-tailed)	.011	.019		.000	.002	.033	.084
	N	235	235	235	235	235	235	235
factor_4	Pearson Correlation	−.260**	−.242**	.535**	1	.268**	−.142*	−.213**
	Sig. (2-tailed)	.000	.000	.000		.000	.029	.001
	N	235	235	235	235	235	235	235
factor_5	Pearson Correlation	−.072	−.054	.199**	.268**	1	−.022	.013
	Sig. (2-tailed)	.272	.408	.002	.000		.737	.839
	N	235	235	235	235	235	235	235
factor_6	Pearson Correlation	−.312**	−.329**	−.139*	−.142*	−.022	1	−.362**
	Sig. (2-tailed)	.000	.000	.033	.029	.737		.000
	N	235	235	235	235	235	235	235
factor_7	Pearson Correlation	.797**	.798**	−.113	−.213**	.013	−.362**	1
	Sig. (2-tailed)	.000	.000	.084	.001	.839	.000	
	N	235	235	235	235	235	235	235

**Correlation is significant at the 0.01 level (2-tailed).

*Correlation is significant at the 0.05 level (2-tailed).

Figure 8.9 Pearson's Two-Tailed Correlations with Significance Tests

In practice, however, prohibitively small sample sizes are rare in database analytics, given the size of most databases. Interestingly, the sample size "over-abundance" is a more likely challenge as too large a sample size can lead to an artificial inflation of statistical significance, a commonly used though controversial measure of the nonspuriousness of correlation and other coefficients. This is an important consideration and as such is discussed in more detail in the subsequent section.

Lastly, a proper assessment of a correlation coefficient also involves a choice between a one- and a two-tail test. A *one-tail test* is used to identify events that are different only in one direction in reference to the average—such as customer spending levels that are considerably above the average. In that sense, a one-tail test would not differentiate between the average and extremely small values, as it is focused on detecting only abnormally large values. A *two-tail test*, on the other hand, can be used to identify values that are either significantly greater or smaller than the expected or average values. Naturally, the two-tail test is informationally richer because it can detect unexpected events on both ends of the continuum—those significantly larger as well as significantly smaller than the average or expected values.

In the past, it was also necessary to specify the so-called level of *statistical significance*. Technically, the significance level of a test is the maximum probability of incorrectly rejecting a

true null hypothesis, which is also known as the Type I error.[20] Since the null hypothesis typically stipulates that there are no differences between the entities being tested, such as two mean product repurchase rates, the concept of statistical significance is in fact a measure of the amount of risk an analyst is willing to accept in concluding that noteworthy differences exist where in fact there are none. In the context of correlation analyses, significance testing is used to assess the degree to which the reported bivariate correlations are manifestations of enduring relationships or a mere product of random chance. However, statistical significance testing suffers from some severe deficiencies, which are particularly evident in the context of database analytics. Given the pivotal role of significance testing in virtually all sample-based analyses, the limitations of significance testing deserve a more exhaustive treatment, presented in the next section.

Beware of Significance Tests!

Statistical significance testing (SST) is a hypothesis testing tool, the purpose of which is to identify universally true effects. SST's secondary and closely related objective is that of generalizing sample-based insights onto a larger population. Although principally a theory development method, significance testing has in recent years been adopted to promotional program measurement where it gained quick acceptance as the impact validation standard.

Operationally, SST utilizes any of the known distribution statistical difference tests, such as *F*, *t*, or χ^2 to compare observed effects to expected effects with the purpose of distinguishing between spurious and persistent relationships, as shown below in Figure 8.10.

While the statistics utilized in significance testing (i.e., the above referenced *F*, *t*, or χ^2) are themselves methodologically sound, their program measurement applications tend to outstretch their usability limits leading to misapplications and misinterpretations. Some of it is due to simple user error, but a considerable share of SSTs misuse can be attributed to fundamental lack of fit between *theory testing* and typical *business objectives*.

Significance Tests	Definition	What Does the Result Mean?
Chi-square test	A test of statistical significance based on a comparison of the observed cell frequencies of cross-tabulation of two variables that would be expected under the null hypothesis of no relationship .	❖ If p-value is less than the chosen threshold (for example, $\alpha = 0.05$), conclude that there is a relationship between two variables with 95% statistical significance .
t-test	A test of significance for continuous variables where the population variance is unknown and the sample is assumed to have been drawn from a normally distributed population.	❖ If p-value is less than the chosen threshold (for example, $\alpha = 0.05$), conclude that there is a relationship between two variables with 95% statistical significance .
F-test	A statistical test of the difference of means for two or more groups (**Analysis of variance:** Sample is distributed normally).	❖ If p-value is less than the chosen threshold (for example, $\alpha = 0.05$), conclude that the means of two or more populations we test are different with 95% statistical significance ❖ If the test result is significant, perform post hoc test to determine specific differences.

Figure 8.10 Types of Statistical Significance Tests

Although rarely compared "side-by-side," theory testing and applied knowledge creation processes differ on some very important dimensions. Perhaps most importantly, theory testing aims to uncover universally true knowledge claims, while marketing analytics focus on the

identification of sustainable competitive advantage. It follows that significance testing is used as a *sample-to-population* generalization tool for scientific theory building purposes, and as a *now-to-future* or longitudinal replicability tool for applied knowledge creation. This is a critical distinction as it gives rise to one of the more common SST application errors discussed later in this text.

Another common SST misapplication stems from its dependence on sample size. Sample size and the likelihood of detecting statistical significance are highly correlated, so much so that at a moderately large sample size even inordinately trivial differences can become statistically significant, while not being statistically significant at a smaller sample size (everything else being the same). For a variety of reasons that are not important at this moment, theory testing research typically utilizes small sample sizes leading to limited sample size distortion. The opposite, however, is true for most applied business endeavors which depend on large scale (i.e., large sample size) for business viability, resulting in a considerable sample size distortion.

Expected precision of estimates is yet another (albeit more subtle) theory testing vs. applied business knowledge-creation distinction. In short, while theory development is primarily concerned with the identification of universally true relationships and less so with the exact quantification of the magnitude of effects, business analyses are almost single-mindedly focused on quantifying program-specific incrementality. It is a matter of pragmatism: The goal of business actions, such as promotions, is to benefit a particular organization only; hence it is of little concern to business analyses if a particular relationship is not generalizable to other users. In fact, from the standpoint of a particular organization, the lack of cross-user generalizability is actually a preferred outcome.

Putting the above pieces together suggests that when applied to a large-scale database analytical initiatives, statistical significance testing is of questionable value for three key reasons: First, extremely small treated vs. control differences are likely to be found statistically significant even if their magnitude renders them practically inconsequential, which will then give rise to the previously discussed statistical vs. practical significance divergence, ultimately leading to SST misapplication. Second, significance testing does not support future replicability generalizations, which means that we cannot use the results from today's test as basis for forming expectations regarding tomorrow's rollout; again, an issue of central importance to promotional program measurement. Third, treatment attributable incrementality cannot be expressed as an exact quantity, which although not a show-stopper is still less than ideal, particularly when the range of effects encompasses both positive and negative values.

Those are not trivial differences. Significance tests are computationally relatively straightforward and highly suggestive of normative applicability limits. At the same time, the goals of the theory building and practical applications-focused analyses are oftentimes quite different. The interaction between the significance tests' applicability limits and the different (i.e., theoretical vs. practical) applications of those tests are sufficient to question the wisdom of unqualified significance testing usage in business applications. SST's sample size dependence (i.e., the likelihood of a given relationship being deemed "statistically" significant increases as the sample size gets larger, everything else being the same), inability to support longitudinal conclusions (i.e., offering an objective quantification of the probability of future replicability of current relationships) or the basic incommensurability of scientific and business objectives (i.e., seeking universally true generalizations vs. future replicability) all highlight the dangers of blind SST reliance by business analysts.

Faced with these shortcomings of an otherwise key methodological element, analysts grew accustomed to drawing a line of demarcation between the *statistical* and practical *significance*. In effect, it has become a commonplace in applied marketing analytics to expressly differentiate between the "statistically significant results we accept" (i.e., the results that are deemed both statis e "statistically significant result we do not accep own below:

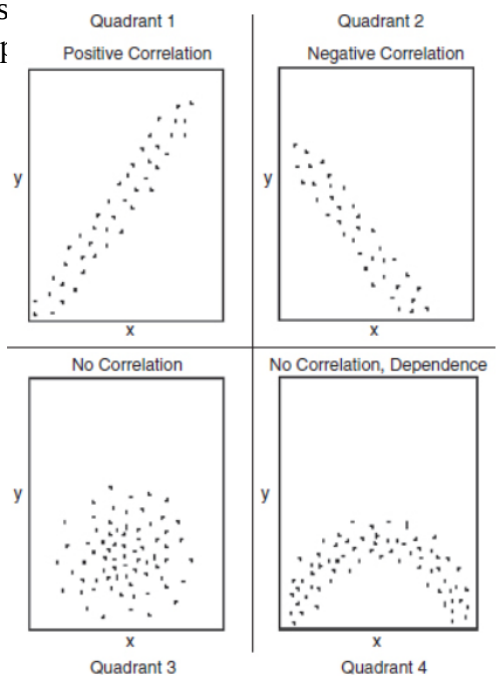


Figure 8.11 Distinguishing between Correlation and Dependence

$$y = \alpha + \beta x_1 + \beta x_1^2 + \varepsilon$$

where,

α is an intercept term (practically interpretable only when the _ terms equal 0)
βx_1 is a test of linear dependence of X on Y
βx_1^2 is a test of curvilinear dependence of X on Y
ε is an error term, or the unexplained residual

Another common analytical data exploratory challenge involves describing interactions taking place in the context of dependence analysis. As mentioned earlier, a correlation between X and Y can be a result of both X and Y being impacted by yet another factor, Z. For instance, product sales (Y) can be dependent to some degree on promotions (X), as well as specific pricing decisions (Z). Assessing an interaction between two predictors (X and Z) involves computing a new variable that is a multiplicative combination of X and Z. As a result, to test the dependence of Y on X and Z, as well as any potential interactions between X and Z, we can again use a multiple regression model to specify the following set of dependencies:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

where,

α is an intercept term (practically interpretable only when the _ term equals 0)
$\beta_1 x_1$ is linear dependence of X on Y
$\beta_2 x_2$ is linear dependence of Z on Y
$\beta_{12} x_1 x_2$ is multiplicative dependence of X and Z on Y (X–Z interaction)
ε is an error term, or the unexplained residual

The above assumes that X and Z, as well Y, are all continuous (i.e., measured with either interval or ratio scales describe earlier) variables. In fact, however, these or any other metrics could be discrete. For instance, promotional spending could be dichotomously coded as "high" or "low" and product pricing could be coded as "regular" and "discounted."

Factorial analysis of variance (ANOVA), a special case of regression analysis, is the best exploratory analytical tool for tackling that problem. The inner-working of ANOVA is somewhat different from that of the above outlined regression. In essence, ANOVA carries out a number of tests, where the means of the dependent variable, such as product sales, are compared across a number of different factors (individual independent variables, such as the discretely coded promotional spending and product price) and their levels (the individual values of independent variables, such as "high" promotional spending, or "regular" pricing). The end objective is to pinpoint statistically significant[21] interactions between specific factor-level conjoints and the dependent variable of interest.

The so-operationalized pursuit of maximally complete explanation typically brings to the forefront the notion of informational domain specification. In other words, is the available data sufficient to generate a complete explanation of the phenomenon of interest?

The *informational domain specification* (IDS) is a relatively complex—and in the opinion of some—an overly academic consideration. In the most general sense, it is a conceptual "blueprint" of the entire explanatory model, spelling out indicator-construct assignments as well as the entire web of the dependence and interdependence relationships. It can certainly take on that appearance, but at the same time, it is crucially important to the creation of competitively advantageous knowledge. IDS is the only objective way of assessing the adequacy of the available raw data to support the stated informational objectives, as tackled with the above outlined statistical methodologies. Conceptually and methodologically, it is a more general form of what is commonly known as *model specification*, which is a critical component of sound theory testing research.[22] Given all that, IDS is discussed in more depth in the next section.

Informational Domain Specification

Throughout the database analytical process exemplified in Figure 1.4, there are several "transitions" where the objective, well-codified science of data analyses intersects the subjective and rarely codified art—or perhaps more correctly stated, intuition—of the analyst. Correctly specifying the informational domain is likely the pinnacle of that intersection.

In everyday terms, informational domain specification is the process by means of which analysts select and arrange specific (raw) data elements to form a conceptual model, based upon which specific (statistical) analyses and tests will be carried out, all with the goal of answering the questions posed by the stated informational needs. *Informational domain specification* represents

an intersection of several competing considerations: First, the scope, in terms of the selected variables, needs to be sufficient. At the same time, or second, the selected variable list needs to be non-redundant and operationally comparable (i.e., variables that are to be related to each other on the same plain of abstraction need to expressed at a comparable level of aggregation). Third, the model needs to exhibit a certain level of parsimony—in other words, throwing every conceivable metric into the mix is undesirable from both the statistical (introduction of numerous, albeit spurious correlations tends to detract from finding a clear solution) and interpretational (describing an outcome in terms of an excessively large number of "important" factors in some way defeats the purpose of conducting the analysis) points of view.

In terms of the outcomes of the IDS process, the informational domain can be just-, over- or under-specified. Ideally, an informational domain is *just-specified*, which is accomplished when a sufficient number of non-trivial explanatory variables are available. Of course, knowing whether or not that is the case is in many regards "half the battle." First of all, no single, objective appropriate number of variables-type benchmark exists, largely because the number of metrics is not important per se, so long as the resultant solution is manageable and maximally explanatory. In other words, the number of explanatory variables is sufficiently but not excessively large when it yields a statistically exhaustive explanation, i.e., the model explains the vast majority of the variability in the data, while at the same time it is small enough to be parsimonious[23] and practically actionable. Given the obvious difficulty of balancing the number of variables and the amount of explanation contained in the model, more often than not a domain is either under- or over-specified.

An *under-specified informational domain* is one that yields too few non-trivial explanatory variables, which translates into an insufficiently small amount of the variability in the data being explained by the model (which means that any predictions made based on such a model are likely to be unstable as well as inaccurate). On the other hand, an *over-specified domain* is one that depends on an excessively large number of trivial, yet correlated explanatory factors, for the explanation of the variability in the data. This is typically an indication of either a relative scarcity of truly explanatory data, or a poor data management strategy, such as making use of too many disaggregate metrics.

In practice, informational under-specification diminishes the explanatory power and reliability of information, simply because under-specified explanations are spurious, or chance-driven. Over-specified explanations, on the other hand, are interpretationally cumbersome because they employ an excessively large number of practically unimportant factors. Under-specification is most often a function of data scarcity, which in many instances is hard to remedy. Over-specification is usually a function of flawed variable retention rationale, such as the use of too many disaggregate metrics or over-reliance on statistical significance testing (SST),[24] which is used frequently as a variable inclusion or retention standard. Excessive dependence on statistical significance tests increases the likelihood that variables of negligible importance will be included alongside highly explanatory factors, particularly as the sample size increases. As large sample size analyses are becoming a commonplace due to the proliferation of large databases coupled with rapid gains in data processing technologies, the frequency of the significance-testing-induced informational over-specification increases, so much so that a more in-depth discussion of the, well, significant limitations of SST seems warranted.

Although much of the informational domain specification is situational, as it depends of the specific characteristics of data, there are a number of general steps that can be taken to increase

the likelihood of the informational domain being just-specified. These include metric aggregation, indexing and variable transformations.

METRIC AGGREGATION

Transactional and other databases usually are made up of variables exhibiting various degrees of aggregation, or specificity. As detailed in the *Data Basics* chapter, some of the marketing data is a by-product of operations (e.g., point-of-sales), other data is purposefully acquired (e.g., consumer satisfaction surveys), and still other represent third-party estimates (e.g., geodemographics), resulting in considerable amount of informational properties invariance. Homogenizing the individual metrics' levels of aggregation can be a relatively complex task, as it may entail computing summary measures for detailed, indicator-level metrics (usually with the help of factor analysis discussed in the next chapter), as well as householding of more disaggregate purchase details. Nonetheless, combining detailed metrics into more aggregate, summary-level variables will have the desirable effect of reducing the number of explanatory variables while retaining the bulk of the original metrics' informational content.

INDEXING

As it is used in the database analytical process outlined in this book, *indexing* refers to the assigning of predetermined labels to database records with the goal of delineating distinct and non-overlapping categories of database records (e.g., households, customers, etc.). In some regards, indexing leads to the creation of "shadow" variables, which are typically used as the basis for record grouping and homogenization of analytical subsets, such as identifying and subsequently selecting (for analysis) only high-value customers. The resultant metrics are almost always categorical which, on the one hand may limit the usability of such metrics in certain contexts, while at the same time may expand the data file's informational domain by creating new predictors.

TRANSFORMATIONS

The *Analytic File Creation* chapter outlined a number of potential data transformations that can be used to correct undesirable distributional properties, such as skewness or kurtosis. Data transformations can also be a useful informational domain specification tool—specifically, the underlying measurement characteristics of data can be re-coded by reversing negatively coded consumer opinions into positively coding ones, or by standardizing items measured on magnitudinally dissimilar scales. Carefully transforming selected metrics will increase the availability of "eligible" variables, thus diminishing the possibility of under-specification.

Data Reduction

The last general area of the data exploration involves possible data reduction steps. In contrast to the previously described data exploration process, which is an integral part of any database exploratory analytical endeavor, data reduction may be a desired component of that process, but not necessarily a required one. Whether or not it should be considered depends on the type of data, which is meant in both an informational as well as a technical sense.

Informationally, a database can contain a number of disaggregate metrics, many of which might be indicators of a more general (and meaningful) higher-order construct. For instance, consumer survey data might contain a number of substantively similar, yet distinct variables, simply because the psychometric measurement theory dictates that latent constructs (such as product interest or brand attitude) should be assessed with multiple indicators, because a single (observable) indicator is rarely, if ever, a perfect predictor of the underlying (unobservable) construct. In the end, a database housing the results of consumer or other surveys is likely to contain a number of metrics which in a singular form are informationally trivial and which should be combined into more meaningful, higher-order aggregates.

At the same time, it is important to consider the technical aspects of such metrics, most notably, their measurement scales. Ideally, these disaggregate metrics are continuous (i.e., measured with either an interval or ratio scales), as that would offer the maximum amount of analytical flexibility. As previously noted, continuous variables are informationally richer than their discrete counterparts, making them analyzable with a wide array of statistical techniques. This is important in the context of data reduction considerations because the most commonly used data reduction technique—*factor* or *principal component analysis*[25]—requires the input variables to be continuously distributed.

However, data reduction is not limited to surveys. Some databases are "cluttered" with so many metrics that basic exploratory analyses or even simple reporting become very cumbersome. Should that be the case, it might be worthwhile to consider extracting the most usable subset of the data into a more easily analyzable sub-environment—in effect setting aside a smaller *data mart*, while keeping everything else in a larger *data warehouse* (see the *Data Basics* chapter for a more in-depth discussion of the differences between data marts and data warehouses). In doing so, it is important to keep in mind that *what is being done* is as important as *how it is being done*. In that sense, selecting a subset of a large database can be a daunting task and making arbitrary choices between what to keep vs. what not to keep would obviously be counterproductive. Hence, at least in some instances, rather than selecting a smaller subset of data, it might be more desirable to compress large quantities of detailed (and individually informationally trivial) metrics into a far smaller number of more aggregate variables. In analytical jargon, this amounts to data reduction analyses.

In general, there are two distinctly different statistical approaches to data reduction:

1. Factor analysis.
2. Correspondence analysis.

Factor Analysis

Although sometimes used to denote a single statistical technique, *factor analysis* is in fact a generic name referring to a class of multivariate statistical methodologies tasked with defining the underlying structure of the data and extracting sets of common underlying dimensions, known as factors. In essence, the analysis evaluates the pattern of cross-variable correlations and identifies interrelationships in a way that pinpoints the inherent variable groupings, giving rise to an objective summing up of multiple disaggregate metrics into a single higher-order variable.[26] Since the underlying analyses are based on inter-variable product-moment correlations, factor analysis requires the input data to be continuous.

Perhaps more than most other multivariate techniques, factor analysis is an iterative method that is built around the loop of *input–analysis–outcome evaluation–input changes–analysis*, etc. The basic reason for the repetitive iteration is the weeding out of specific metrics that may only be

spuriously correlated with others—i.e., could not be included in a robust higher-order metric aggregation. Again, because the method simultaneously considers the entire correlation matrix (i.e., all inter-variable correlations), finding a robust, reliable solution requires the identification and elimination of random distractions.

Aside from the data, it is also important to make several analysis-related decisions, as discussed below. The first two—the extraction and rotation methods—are input decisions, while the third consideration—retention criteria—is the output consideration.

Extraction Method

There are several different approaches to identifying the underlying variable groupings, including the principle component analysis, generalized least squares, maximum likelihood, alpha factoring, etc. Naturally, these decisions should be made in the context of data specifics, but in general, the *principal component analysis* was found to yield stable and valid results, across situations. However, given its inner-workings (i.e., it is built around the extraction of a single factor—the principal component—followed by a redistribution of factor membership), this extraction method necessitates a careful selection of a factor rotation method, discussed next. It is important to keep in mind that the extraction order is important, as the amount of the total variance explained by individual factors is a function of their extraction order. Thus the first factor will always explain more variability than the second, which in turn will always explain more than the third, etc.[27] It is also important to keep in mind that the resultant factors are based on an assumed linearity—in other words; non-linear cross-variable interrelationships are likely to go unrecognized.

Rotation Method

Since factor analysis essentially leads to a grouping of variables in a multidimensional space typically represented by the axes of Cartesian coordinates, a decision needs to be made regarding the stipulated interrelationship among those classifying axes. In general, that relationship can either allow for some factor correlations (a provision statistically known as *oblique* rotation) or it can assume full factor independence (statistically known as *orthogonal* rotation). Once that decision is made, the axes themselves need to be mathematically rotated to find the best fit for the previously extracted variable groupings. Although there are a number of specific axis rotation algorithms that have been developed, they can all be categorized as either *orthogonal* or *oblique*, since the underlying factors can either be assumed to be correlated or uncorrelated. Orthogonal extraction methods include varimax, equimax and quartimax, while the oblique extraction methods include oblimin, promax and orthoblique techniques. Although subtle computational differences separate the individual orthogonal and oblique rotation algorithms, in practice they tend to produce results that are not interpretationally different.

Retention Criteria

The key output consideration is how many factors to retain. In that sense, factor analysis could be viewed on a continuum ranging from the *number of factors = the number of input variables*, to the *number of factors = 1* (i.e., all input data is summarized into a single factor). Clearly, neither of the two extremes is particularly appealing or usable, which is why it is important to identify an objective decision rule. There are typically two options: 1. a pre-existing knowledge of the number of factors that are to be expected, or 2. eigenvalue[28] = 1 rule. Although an analyst might certainly

have an expectation of the number of factors based on previous research or theoretical considerations, it is far more common for that not to be the case. The frequently used "eigenvalue = 1" rule of thumb simply postulates that a combination of several variables (i.e., a factor) that cannot explain the amount of variance that is an equivalent of a single variable has trivial informational value and should be treated as a set of spurious correlations, rather than enduring cross-variable relationships.

The end result of the three-part consideration set outlined above should be a rotated factor matrix, exemplified in Figure 8.12.

As shown in the figure, there were 5 separate factors extracted in this particular analysis. The intersection of rows and columns identify variable-factor memberships and the individual coefficients spell out factor loadings, which represent the strength of the relationship (i.e., correlation) between the underlying factor and its individual components. The factor loadings are standardized, ranging from –1 to +1 and the coefficients smaller than .4 were not shown to make the visual output interpretation easier. It is worth pointing out that several variables, not shown, were excluded from the analysis during earlier iterations due to being split-loaded, which is a condition of a variable exhibiting strong correlation to more than a single factor (typically, it exhibits itself in two or more approximately equal-sized loadings).

The end result of the analysis depicted in Figure 8.12 was the compression of a set of 15 disaggregate input variables into 5 higher-order factors, which represents a 3-to-1 reduction in a number of variables—a rather moderate reduction. Depending on the type of data, an average ratio of 10-to-1 or so may be plausible.

To account for cross-factor differences in the number of constituent variables (as previously implied, the first extracted factor typically pulls the highest number of input variables, followed by the second extracted factor, etc.), the resultant factor composites are also weighted by the number of inputs variables, as shown below:

$$factor_{\kappa} = (variable_1 + variable_2 + variablen)/_n$$

The resultant summed and weighted factor scores are themselves continuously distributed metrics that can be used as inputs into subsequent analysis, such as cluster analysis discussed in the *Segmentation* chapter, or regression analysis outlined in the *Behavioral Predictions* chapter.

Beyond the factor structure shown above, it is also important to consider the amount of the total input dataset's variance explained by the analysis. Figure 8.13 shows the typical evaluative output.

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
CompanyAge_ Transformed					.831
NumberofEmployees_ Transformed	.854				
AuditFeesOther_ Transformed	.556				
AuditFeesTax_ Transformed	.591				
NumberofBoard MeetingsLastYear_ Transformed				-.781	
DominantShareholder Percentage_Transformed			.897		
InsiderControl Percentage_Transformed			.801		
InsidersPlus5Owners Percentage_Transformed			.791		
SharePrice52WkHi_ Transformed		.979			
SharePrice52WkLo_ Transformed		.939			
SharePriceCurrent_ Transformed		.971			
SharesOutstanding_ Transformed	.828				
Revenues_Transformed	.880				
PreviousCEOTenure_ Transformed				.697	
CEOAllOtherComp_ Transformed					.504

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Figure 8.12 Rotated Factor Matrix

Total Variance Explained

Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	3.137	20.911	20.911
2	3.073	20.484	41.396
3	2.291	15.272	56.668
4	1.307	8.714	65.382
5	1.242	8.280	73.662

Extraction Method: Principal Component Analysis.

Figure 8.13 Explained Variance

As shown above, the 5 factors (called *components* above, due to the *principal component analysis* being used as the extraction method) cumulatively explain nearly three-quarters of all variability in the input dataset, which reflects a good fit of the final factor structure to the data. It means that the initial set of 15 variables could be replaced with a much smaller set of 5 factors, an 80% reduction in the number of variables achieved at the cost of loss of only about 24% of the potential explanatory power, which can be a very attractive tradeoff, particularly when the starting point is a dataset containing several hundreds or even thousands of individual metrics.

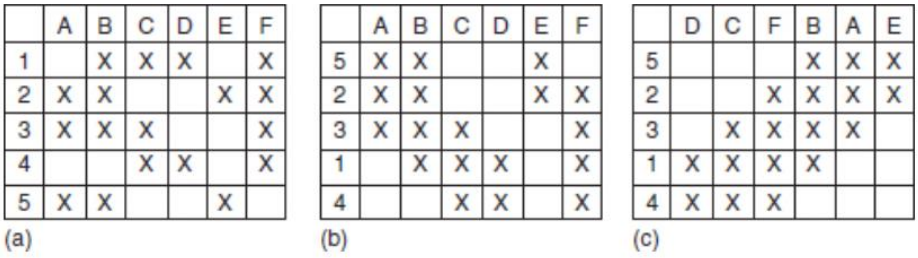
An important, yet often overlooked constraint placed on factor analysis is the ratio of the number of records to the number of variables. It is usually recommended that, at minimum, the input data matrix has at least 4-to-5 times as many usable records as it has variables; e.g., a 100 variable dataset should have at least 400–500 usable observations. Significant departures from this number of records-to-number of variables ratio can make factor loading estimates unstable.

Correspondence Analysis

Unlike factor analysis, the use of which is relatively widespread, particularly in survey and academic research, correspondence analysis (CA) is a relatively little-known technique, at least in the U.S. The general goal of this methodology, as implied by its name, is to describe the relationships between discrete (nominal or ordinal) variables in a low-dimensional space (i.e., a relatively small number of categories), while at the same time describing the relationships between each variable's categories. Similar to cross-tabulation contrasts, χ^2 tests are used as bases for making conclusive determinations, but CA decomposes the χ^2 measures of association into components, in a manner resembling the above-described factor analysis (specifically, the principal component extraction). Thus in a way, correspondence analysis accomplishes the goal of factor analysis for categorical data, effectively offering a complementary methodology for non-continuous data reduction. In a methodological sense, CA projects estimates for one variable on an underlying "factor" to a category estimate for the other variable, thus making possible the arranging of the individual metrics based on their similarity to each other and relative to axes in multidimensional space. Ultimately, it makes possible the conversion of frequency table data into graphical displays.

A simple illustration of the correspondence analysis process steps is illustrated in Figure 8.14.

Shown in part (A) are two categorical variables: one with categories A through F and the other one with categories 1 through 5. By rearranging first the column values A–F for (variable1), as shown in part (B), followed by rearranging of the row values 1–5 (variable2), a clear pattern emerges, as shown in part (C).



	A	B	C	D	E	F
1		X	X	X		X
2	X	X			X	X
3	X	X	X			X
4			X	X		X
5	X	X			X	

(a)

	A	B	C	D	E	F
5	X	X			X	
2	X	X			X	X
3	X	X	X			X
1		X	X	X		X
4			X	X		X

(b)

	D	C	F	B	A	E
5				X	X	X
2			X	X	X	X
3		X	X	X	X	
1	X	X	X	X		
4	X	X	X			

(c)

Figure 8.14 The Process of Correspondence Analysis

The two key questions typically raised in conjunction with the assessment of the results of correspondence analysis are those of external and internal stability. *External stability* is of most interest in the context of theory-testing-oriented research, as it points to the concepts of statistical significance and confidence interval estimation. In other words, are the sample-based results indeed representative of a larger population? Depending on the specific goal of the data analytical endeavor, external stability of CA results may be of varying concern to the analyst. Quite often, in applied analysis, the goal is to quantify effects that are true only of the particular sample only. For instance, when investigating purchase trends of a particular brand's buyers, an analyst is in fact only interested in effects that are attributable just to that particular sample of shoppers, as generalizing those results to all shoppers would be of no practical value. Should that be the case, external stability of CA results would not be of great concern.

On the other hand, *internal stability*, which is a reflection of the degree to which specific results provide a good summary of the dataset on which they are based, will always be of keen interest to analysts. In other words, how valid or robust (in a statistical sense) are the results vis-à-vis the underlying dataset? In terms of the underlying data, internal validity is about determining if results were unduly influenced by outlying observations or overly influential variable categories, so much so that the shown results are not likely to be replicated with another dataset.

Both external and internal stability of correspondence analysis can be assessed with *bootstrapping*, which is a technique of sequential simulated re-sampling with replacement from the dataset at hand. The effect of bootstrapping is an empirical examination of the replicability of CA solution by comparing results across multiple, simulated "new" samples.

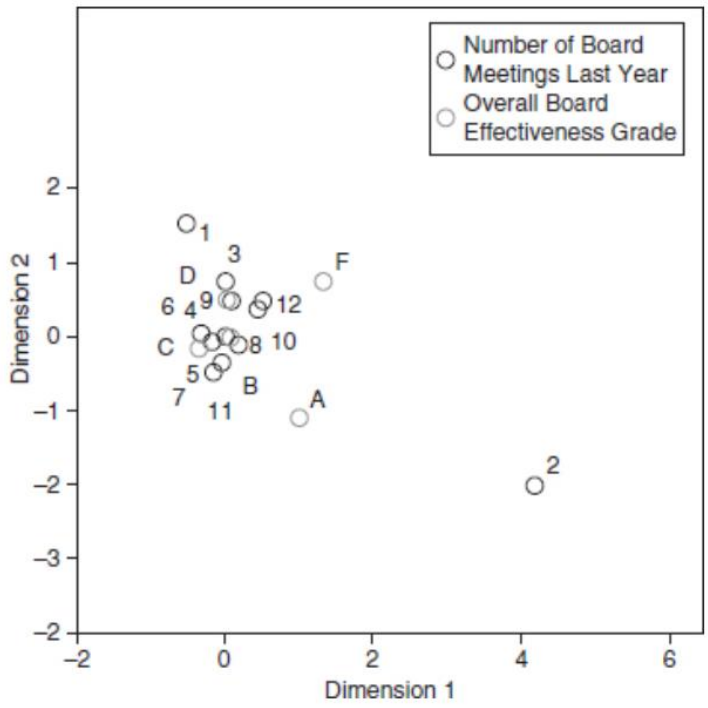


Figure 8.15 Correspondence Analysis

The plot in Figure 8.15 shows a clustering of two discrete metrics: the *number of board (of directors) meetings last year* and *the overall board effectiveness grade*.[29] The former is measured on a scale ranging from monthly (12) to once a year (1), while the latter is measured on a traditional school scale of A–F. The scatterplot reveals close interactions between the two factors as well as the individual factor levels.

Although visually compelling, the results of correspondence analysis are considerably more subjective than outputs of factor analysis. Nonetheless, this technique offers a tool that could be quite helpful in compressing the otherwise hard-to-manage number of non-continuous metrics.