Chapter 9
Test Design and Analysis

Abstract Another cornerstone of database marketing is testing. Testing provides transparent evidence of whether the program prescribed by sophisticated data analyses actually is successful in the marketplace. Much of the testing in database marketing is extremely simple – select 20,000 customers, randomly divide them in half, run the program for one group and not the other, compare results. However, there are several issues in designing and analyzing database marketing tests; we discuss these in this chapter.

9.1 The Importance of Testing

Capital One may be the one of the most successful credit card companies today (Cohen 2001). The secret to the success is its test-and-learn management philosophy that Capital One calls its Information Based Strategy (IBS). Capital One conducted 45,000 tests in the year 2000, which on average is 120 per day. For example, once Capital One comes up with an idea for a new product offering, it attempts to find a target population by testing the new product with various promotional campaigns to various samples of customers. Based on the test results, Capital One identifies what types of customers are most receptive to the new product and what should be the corresponding promotional campaign. It sometimes even conducts additional tests to fine-tune the strategy. Capital One always makes important marketing decisions (e.g., customized pricing, promotion, and packaging) through a series of tests.

Database marketers should not invest a large amount of company resources unless its expected benefit is greater than the costs. Frequently it may not be easy to calculate the expected benefit because the future is uncertain. Unless you are absolutely sure that it will succeed, you should conduct tests to make an informed decision. The objective of testing is to obtain more information before committing a large amount of resources and, hence, reduce the risk of possible failure. The field of database marketing is particularly amenable to tests because companies have addressable customer databases and hence can

randomly assign its customers to various treatment conditions, and observe the results.

While Capital One is the acknowledged leader in database marketing tests and is known for extensive use of testing, most database marketers consider testing an integral part of the way they do business. Database marketers test various decisions including media choice, the development of promotional campaign, the selection of mailing lists, choice of message format, and so on. Moreover, the decision-making process is really "closed-loop." A campaign is revised based on a test, the modified campaign is tested, then implemented, and then the results are used to suggest further tests, and so on. That is, information learned from a test or from full-scale campaigns become inputs to the next tests, which in turn feed the next round of testing and full campaign roll-outs.

9.2 To Test or Not to Test

Probably the first question that should be asked before conducting a test is the most basic – should a test be conducted? As discussed, testing provides information to aid in making correct management decisions. However, information is usually obtained at a cost. Testing costs may include the cost of time delay as well as its administrative cost. For example, to assess the benefit of a loyalty program or a churn management program, one really should run the test for about a year. This is typically not practical. The database marketer must think through whether useful information can be gleaned from a 1 or 2-month test. Hence the decision to collect information or data can be analyzed to see if the expected benefit of the information exceeds its collection costs.

We discuss two approaches for deciding whether to run a test. The first is based on decision analysis and is called the "Value of Information." This potentially quantifies how much the database marketer should be willing to spend on a test. The second approach, "Assessing Mistargeting Costs," is more conceptual, but provides a framework for thinking about whether or not to conduct a test.

9.2.1 Value of Information

Testing provides information. In this section we discuss the fundamental concepts in quantifying the value of information. We first study a decision tree that is very useful for understanding complex decision-making problems. Using the decision tree, we show how to calculate the "value of perfect information" and then extend to the problem of computing the "value of imperfect information."

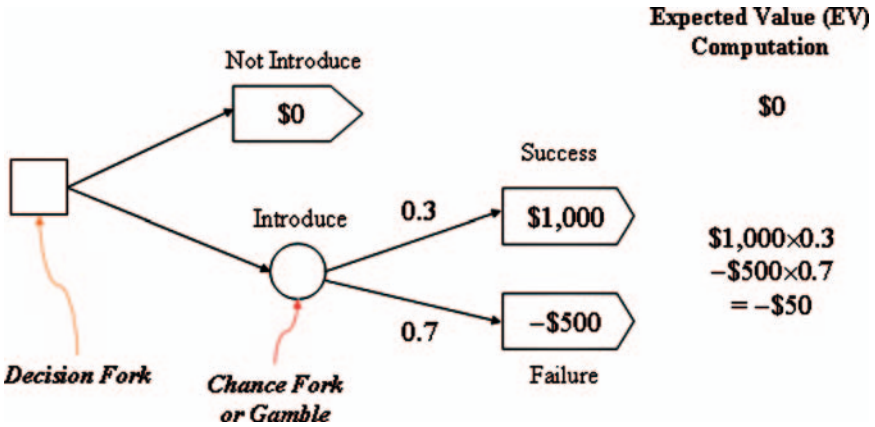

Fig. 9.1 Decision tree for calculating the expected value of a new product launch.

Consider a problem of a new product introduction. The average probability of new product success is known to be 30%. That is, without collecting any additional information on the new product, the success probability of the new product is 30%, and its failure probability is 70%. Suppose that a firm would make $1,000 if the new product succeeds and lose $500 if it fails. Should the firm introduce the new product? If the firm does not introduce the new product, the payoff is $0. On the other hand, if the firm decides to introduce the new product, it will succeed with probability 0.30 and gain a payoff of $1,000 and fail with probability 0.70 and obtain a payoff of −$500. As a result, the expected value or payoff for the new product introduction is $-\$50\ (= \$1,000\ \times\ 0.3 - \$500\ \times\ 0.7)$. Therefore, the firm should not introduce the new product. The decision tree shown in Fig. 9.1 summarizes these calculations.

Decision trees are a graphical way to organize the probabilistic computations leading to the best decision. We draw the decision tree starting with a decision. Should the firm introduce the new product? The decision fork shown as the square box in Fig. 9.1 has two arrows (or alternatives) coming out: introduce or not introduce. We now evaluate the payoffs from each alternative. The outcome of the first alternative or "not introduce" is $0. The payoff from the second branch is more complicated to calculate. If the firm decides to introduce the new product, the payoffs will be determined by chance. We represent this as a circle – called the chance fork – distinguished from the decision fork. Two possible outcomes branching from "introduce" are "success" or "failure." The new product will succeed by 30% of the time and fail 70% of the time. The payoff given "success" is $1,000 and the payoff given "failure" is −$500. Hence the "expected value" or payoff from introducing the new product is $-\$50\ (= \$1,000 \times 0.3 - \$500 \times 0.7)$. Since the expected payoff of "not introduce" ($0) is larger than that of "introduce" (−$50), the firm should not introduce the new product.

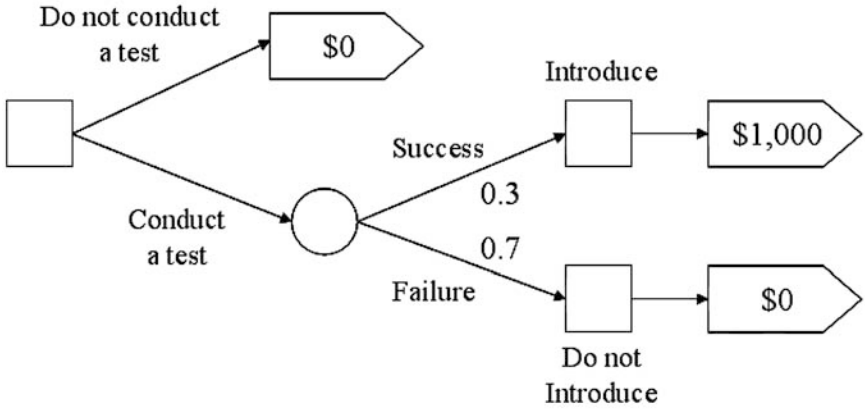

Fig. 9.2 Decision tree for assessing the value of perfect information.

9.2.1.1 Value of Perfect Information

We next consider a case of conducting a test to aid in making decision on introducing a new product. We first consider the value of the perfect test (or information). The perfect test can forecast with 100% accuracy whether the new product will succeed or fail. Figure 9.2 shows the decision tree to determine whether we conduct a test.

If we do not conduct the test, we will not introduce the new product (because the expected value of launching the product is −$50 as calculated above) so that the corresponding payoffs will be $0. However, if we decide to conduct a test, the payoffs will be determined by chance. There is a 30% chance that the new product will actually be a success and 70% chance it will be a failure. Assume the test can perfectly predict whether the new product will succeed or fail. If the new product is forecasted to succeed in the test, it will actually succeed. The firm should then introduce the new product and the resulting payoff will be $1,000. Alternatively, if the new product is predicted to fail in the test, it will actually fail. The firm should then not introduce the new product, and the corresponding payoff will be $0. Therefore, the expected payoff becomes $300($= \$1,000 \times 0.3 - \$0 \times 0.7$). The value of perfect information (or the perfect test) is $300 since the payoff increases from $0 to $300 by conducting the test. In other words, the firm should conduct the perfect test unless its cost is greater than $300.

9.2.1.2 Value of Imperfect Information

Information provided by a test is rarely perfect. The test cannot provide perfect information for several reasons including small sample size, measurement errors, and so on. Going back to the problem of new product introduction, we assume that the test provides imperfect information. Assume the test correctly forecasts 90% of the time when the new product will actually succeed. So the test will say "failure" 10% of time for the would-be successful new product. In addition, if the new product will actually fail, the test is assumed to predict that the new product will fail 80% of the time, and wrongly forecast that it will succeed 20% of times. What is the value of information provided by this imperfect test?

Before we proceed into the decision tree for imperfect information, let us briefly calculate some important preliminary probabilities. We are able to calculate the joint probability of test results ("Success" or "Failure") *and* actual results (Success or Failure) by multiplying these two probabilities. For example, the (joint) probability that the test says "success" and the new product will actually succeed as P(Product is a Success & Test says "Success") = P(Success & "Success") = P(Product is a Success) × P(Test says "Success," given that the product actually is a success) = $(0.3) \times (0.9) = 0.27$. Similarly, P(Product is a Success & Test says "Failure") is $(0.3) \times (0.1) = 0.03$ while P(Failure & "Success") = 0.14 and P(Failure & "Failure") = 0.56.

From these four joint probabilities, we can calculate the probability that the test says that the new product is a "success" or "failure." The probability the test says the product will succeed, P("Success"), equals P(Product is a Success & Tests says "Success") + P (Product is a Failure & Test says "Success") $= 0.27 + 0.14 = 0.41$. The test will say 41% of the time that the new product is a "Success." When the test says the new product will be a "Success," about 66% of the time ($= 0.27 \div 0.41$) the product will actually succeed, but 34% of the time ($= 0.14 \div 0.41$), it will fail. Similarly, P(Test says "Failure") is 0.59 ($= 0.03 + 0.56$). The test will say 59% of the time that the new product is a "Failure." And when the test says "Failure," about 5% of the time ($= 0.03 \div 0.59$) it will instead succeed and 95% of the time ($= 0.56 \div 0.59$) it will fail.

Now we are ready to draw the decision tree for imperfect information. Similar to the case of perfect information, the firm has a decision making problem of whether to conduct a test. The payoffs will be \$0 if the firm does not conduct a test. Note that the firm should not introduce the new product without additional information provided by the test. However, if the firm decides to conduct a test, the payoffs will be determined by chance. Figure 9.3 summarizes the decision tree to determine whether we conduct a test.

Given conducting a test, there is a chance fork where P("Success") = 0.41 and P("Failure") = 0.59. That is, the test will say 41% of times that the new

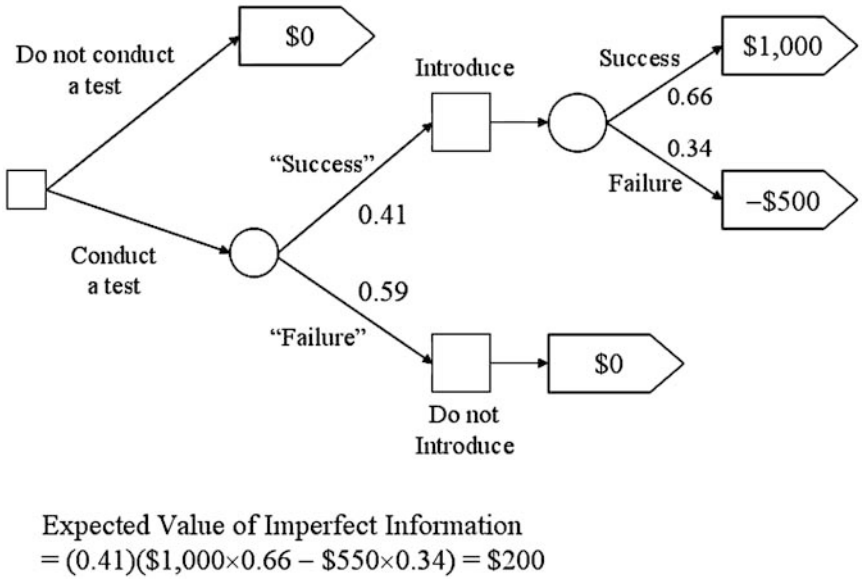

Fig. 9.3 Decision tree for assessing the value of imperfect information.

product is a "Success" and 59% of times that the new product is a "Failure." If the test predicts "Success," the firm will face another decision making problem of whether to introduce the new product or not. As computed before, if the test predicts "Success," there is a 66% ($= 0.27 \div 0.41$) chance it will actually succeed, but a 34% ($= 0.14 \div 0.41$) chance it will fail. The expected value of introducing the new product if the test says "Success" is \$488 ($= \$1,000 \times 0.66 - \$500 \times 0.34$). As a result, the firm should introduce the new product when the test predicts "Success." Similarly, if the test predicts "Failure," the firm will face a decision making problem of whether to introduce the new product or not. If the test predicts "Failure," there is a 5% chance ($= 0.03 \div 0.59$) it will instead succeed and a 95% chance ($= 0.56 \div 0.59$) it will actually fail. Hence, the expected value of introducing the new product if the test says "Failure" is −\$425 ($= \$1,000 \times 0.05 - \$500 \times 0.95$). As a result, the firm should not introduce the new product when the test predicts "Failure."

Combining the above results, if the test says "Success," the firm should introduce the new product and its expected profit is \$488. Alternatively, if the test forecasts "Failure," the firm should not introduce the new product and its expected value is \$0. In addition, P(Test says "Success") $= 0.41$ and P(Test says "Failure") $= 0.59$. Hence, the expected profit from conducting the imperfect test is \$200 ($= \$488 \times 0.41 + \$0 \times 0.59$). That is, the payoff increases from \$0 to \$200 by conducting the imperfect test. Note that it is less than the value of perfect test (\$300). The firm should conduct the imperfect test unless its cost is greater than \$200.

The above illustrates a decision-theoretic technique for deciding whether the company should launch a product. The test might be a direct mail offer announcing the product, so this is very relevant to database marketing. The tree-approach is logical and provides a nice "picture" of the decision-making problem. However, it requires key inputs, for example, the probability that the test will say "Success" if indeed the product will succeed, etc. These probabilities typically must be assessed judgmentally. This may seem a bit disconcerting, but the decision-theoretic viewpoint is that managers internally weigh these chances anyway when deciding whether to conduct a test. The value of information approach merely asks the manager to write down those assumptions explicitly and then "play out" rigorously the implications of those assumptions on what the manager should decide.

9.2.2 Assessing Mistargeting Costs

Another way to view the question of whether to test is as follows: There is a correct or optimal decision to make. However, we may not make that optimal decision for two reasons: (1) We decide to conduct a test and the test involves wrong decisions for some or all of the customers involved in the test. This is called the mistargeting costs of the test, or MT_{test}. (2) We roll out what we think is the optimal action on our entire customer base and it turns out to be the wrong decision. This is what we call mistargeting costs of the rollout, or $MT_{rollout}$. We therefore have the following formula:

$$\Pi = Optimal\ Profit - DC_{test} - MT_{test} - MT_{rollout} \tag{9.1}$$

where:

Π = Total profit
$Optimal\ Profit$ = Profit if the company takes the correct action
DC_{test} = Direct costs of the test
MT_{test} = Mistargeting costs of the test
$MT_{rollout}$ = Mistargeting costs of the rollout

For example, a company may need to decide whether to cross-sell Product A or Product B to its customers. There is a correct decision – Product A, B, or neither – but we don't know which is correct. The direct cost of a test would include administrative costs, the cost of delaying actions that may allow competitors to move faster, and the cost of contacting people for the test, etc. $MT_{rollout}$ would be the deviation we get from optimality because we cross-sell the wrong product or cross-sell when neither of the products are profitable. MT_{test} would be the cost we would incur by taking wrong

actions during the test. For example, we might randomly select three groups of customers, each of sample size n, and cross-sell Product A to Group I, Product B to Group II, and neither to Group III. For one of these groups, we've made the right decision, but for two groups we have made the wrong decision. The mistargeting costs occur because for two of the groups, we've wasted resources and may not be able to cross-sell these customers again for this particular campaign (e.g., if you contact the customer for Product A, you can't go back to them later to cross-sell product B).[1]

The level of mistargeting costs will be lower if (1) we have good prior knowledge on the correct course of action, (2) there is low variation in the possible value of a response if the customer responds, and (3) there is low variation in the possible response rates that might be obtained. That is, if there are only a limited number of possible values for the value of response and the response rate, and we have a good prior on it anyway, mistargeting costs will be low. In addition, $MS_{rollout}$ will be lower to the extent that we've conducted a large test, i.e., have a large sample size, because then we're more likely to learn the correct action and won't mistarget on the test. We can summarize these thoughts in the following extension of Equation 9.1:

$$\begin{aligned}\Pi = &\textit{Optimal Profits} - DC_{test}(n) - f(Priors, \sigma_V, \sigma_p) \\ &\times n - g(Priors, \sigma_V, \sigma_p, n) \times (N - n)\end{aligned} \tag{9.2}$$

where:

$f(\bullet)$ = Mistargeting cost per participant in test
$g(\bullet)$ = Mistargeting cost per customer in rollout
N = Total size of customer base
n = Total number of customers participating in the test

Given our discussion above, $f(\bullet)$ and $g(\bullet)$ will both decline as a function of strong priors on the correct action, but increase if there is wide variation in the possible value of a response or the response rate itself.

Equation 9.2 provides the following insights:

- The purpose of a test is to transfer mistargeting costs from the full rollout to a test. The mistargeting costs in the test are incurred on a smaller subset of customers ($n << N$), but we learn from the test ($g(\bullet)$ is decreasing in

[1] We are implicitly assuming that the test "destroys" the experimental units. If the customers in a test could be included in the full rollout, MT_{test} would be much smaller. But this is often not the case. Consider a credit card test where Groups A and B are randomly selected to receive two different cards. The optimal credit card might turn out to be the Group B card. But to then go back to Group A and offer them that card would present problems. First, some of them would have signed up for the Group A card. Second, the Group B card may be perceived differently by the Group A customers because they saw the Group A card first. Third, the company may wish to avoid "cluttering" their customers, so may rule out tested customers from the full rollout.

n) so that mistargeting costs in the rollout are minimized, and these lower costs are multiplied by a large number ($N - n$).

- If we have strong prior information on the right course of action, we need not test because we're pretty sure of the right answer. So why incur the direct costs of testing plus the mistargeting cost of taking the wrong action with one of our experimental groups in the test (MT_{test})?
- If there is wide variation in possible values of either the value of a response or the response rate, then we should test, because there is then a huge plus or minus around the mistargeting costs we could incur with a rollout.
- Too little testing (e.g., small n or not many treatment groups) can hurt because we don't learn enough from the test to decrease mistargeting costs for the rollout. On the other hand, too much testing (e.g., high n or too many treatment groups) can also hurt because we'll incur a lot of mistargeting costs on the test and even though we'll probably learn the correct course of action, we won't be able to apply the lower mistargeting costs on the rollout to enough customers ($N - n$ will not be large enough). So there is probably a middle-ground to be taken with regard to testing.

Hypothetically, Equation 9.2 could be quantified, but we see its value as a framework for providing guidance of whether or not to test. The above bullet points highlight the insights generated from the framework. Generally, a test should be run if (1) prior information on the correct course of action is not available or not reliable, (2) there is wide variation in the possible value of a response, (3) there is wide potential variation in the response rate, (4) the direct costs of running the test, in terms of time, administrative, and contact costs, are low, and (5) the number of customers and treatments needed to learn much from the test is not a significant fraction of our total customer base. For example, if our total customer base is 30,000 and we are thinking of response rates in the range of 1%, we may want to test three groups of 5,000. But that means 15,000 customers are involved in the test, and that is a significant proportion of our total customer base. We could calculate some scenarios depending on potential response value or response rate, but testing half a company's customer base means the direct costs are probably high, and we may only be able to apply our learning to the untested half of our customer base.

9.3 Sampling Techniques

Once we define the population and sampling units for a test, we draw one or more samples from the population. Broadly classifying, there are two types of sampling techniques: probability sampling and nonprobability sampling.

9.3.1 Probability Versus Nonprobability Sampling

A probability sample is where customers ("sampling units") are selected by chance, and, hence every customer in the population has a known chance of being selected for the sample (Boyd et al. 1981). A probability sample can be implemented objectively since customers are selected strictly at random. This probabilistic selection allows us to measure sampling error and consequently make statistical inferences based on the results.

On the other hand, a nonprobability sample is where samples are not selected randomly. Here one selects customers based on the researcher's judgment, convenience, or other nonrandom process. Since subjectivity is involved in the sampling process, we cannot determine the probability of each customer being included in the sample. As a result, we cannot measure sampling error and there is a high risk that statistical inference based on a nonprobability sample will be biased.

There are various types of nonprobability samples including convenience sampling, judgmental sampling, quota sampling, snowball sampling, etc. These samples are frequently used in survey research due to lower sampling costs and faster sample collection, even though they are statistically inferior to a probability sample. On the other hand, most database marketers use the probability samples. Typical database marketers have customer information files and, hence, are able to select random samples quickly and cheaply.

9.3.2 Simple Random Sampling

We focus now on the probability sample. Several kinds of probability samples are in common use. Varying in terms of efficiency, they include simple random sampling, systematic sampling, stratified sampling, cluster sampling and others. High efficiency means that, for the same sample size, a parameter is estimated more accurately, i.e., the standard error of its estimate is lower. Generally, sampling efficiency is positively related to sampling cost. Given the sampling budget, database marketers should select the most efficient sampling technique.

Simple random sampling is the most popular probability sampling technique. Most statistical inference assumes that observations are collected by simple random sampling. With an accurate list of all the firm's customers or prospects, it is cheap and easy to implement. In simple random sampling, every items/names has an equal chance of being included in the sample. That is, simple random sampling is similar to a lottery system. If we sample n items without replacement from the population of size N, this probability is n/N.

Let us explain how simple random sampling works from an illustration. Suppose a database marketer has 10 customers in her customer information file. She wants to select two customers by simple random sampling. To draw a

simple random sample, each of ten customers (in the population) is assigned a unique identification number, one through ten for example. Next a random number (r_1) is generated from a 0–1 uniform distribution. If $0 \leq r_1 < 0.1$, then we select the first customer. We select the second customer for $0.1 \leq r_1 < 0.2$, the third for $0.2 \leq r_1 < 0.3$, and so on. After we select the first customer for the sample, another random number (r_2) is generated from a 0–1 uniform distribution. A customer is selected among the remaining nine customers. If $0 \leq r_2 < 1/9$, then we select the first customer. We select the second customer for $1/9 \leq r_2 < 2/9$, the third for $2/9 \leq r_2 < 3/9$, and so on.

Sample selection of size n from the population of size N can be similarly done. Fortunately to database marketers, most commercial software such as SAS have a built-in function of implementing simple random sampling. All database marketers need to do is specify a simple command for performing simple random sampling.

9.3.3 Systematic Random Sampling

Even though simple random sampling will be representative *on average*, there is still a chance it could yield an un-representative sample, especially if the sample size is small. Hence, many database marketers prefer employing other sampling techniques that provide higher statistical efficiency (without incurring not much additional costs) than simple random sampling. An alternative sampling technique frequently used by database marketers is systematic sampling. Systematic sampling provides an easy way to implement a simple random sampling. Moreover, it is often more efficient than simple random sampling, as explained below.

Let us illustrate systematic sampling by an example. Suppose we want to sample n out of the population size N. First, we determine the sampling interval (k) by rounding N/n to the nearest integer. Next, we randomly select a starting point and select every kth item successively in the target population. For example, if N is 1,000 and n is 100, then the sampling interval k should be 10.[2] Then an item between 1 and 10 is randomly selected. If this number is 8, the sample of 100 customers will consist of customer 8, 18, 28, and up to 998.

Systematic sampling is statistically more efficient than simple random sampling when the ordering of elements in target population is related to the variable of interest. For example, if the customer information file is ordered with respect to their cumulative purchase amounts, systematic sampling will evenly select customers with various purchase amounts. It increases the sample representativeness. On the other hand, a simple random sampling may be unrepresentative because it may sample only heavy users or

[2] Systematic sampling is often called a Nth name sampling in direct marketing applications. Here N represents a sampling interval k.

a disproportionate number of heavy users. However, if the population is ordered in a way unrelated to the variable of interest – for example, customers ordered alphabetically – systematic sampling will provide almost identical sampling error to simple random sampling (Malhotra 1993).

Systematic sampling yields a probability sample in that every element in the target population has a known and equal chance of being included in the sample. It is the most popular sampling technique among database marketers since it is frequently more efficient than simple random sampling without incurring additional costs.

9.3.4 Other Sampling Techniques

There are other probability sampling techniques such as cluster sampling, stratified sampling, area sampling, sequential sampling, etc. These are not very popular among database marketers, but among survey researchers.

Researchers often use cluster sampling rather than simple random sampling to save on survey costs. In cluster sampling, samples are selected in groups. For example, suppose a researcher needs a sample of 1,000 representative US customers for in-depth personal interviews. Simple random samples will provide 1,000 customers who live all over the country. It is not economically sensible to interview 3 customers in New York, 5 in Los Angeles, and so on. In cluster sampling, the USA is divided into several clusters or blocks – using zip codes, for example. And randomly select a manageable number of clusters, say 10, and select 100 customers for each of the selected cluster. Cluster sampling will significantly reduce the sampling costs by selecting a small number of clusters in the first stage, but there is a danger of sample misrepresentation. Customers in a block or a cluster tend to be similar in demographic characteristics. Hence, if clusters covering big metropolitan cities are only selected in the first stage, customers selected from those clusters in the second stage may not be representative of average US customers.

The goal of stratified sampling is to increase statistical efficiency by increasing the sample representativeness. Database marketers are beginning to use this sampling more frequently today. Stratified sampling first divides the target population into several segments with respect to one or more common characteristics and then randomly selects customers from each one of these segments. For example, the customer base might be segmented several groups based on profitability, a segment below $200, a segment of $200–300, and so on. Stratified sampling guarantees more representative samples with respect to the criterion used to segment the target population. Statistical efficiency will be greater when the customers within each segment are more homogeneous. There are several strategies of stratified sampling. The most popular is the proportional allocation which uses a sampling fraction in each of the strata proportional to that of the total population. For example, if

the population consists of 70% in the female stratum and 30% in the male stratum, the relative size of the two samples should reflect this proportion. See Lehmann et al. (1998) for more details on stratified as well as cluster sampling.

9.4 Determining the Sample Size

Determining the test sample size is not an easy task. Marketers are interested in knowing the true parameter value (e.g., response rate for the direct mail offer) for the target population. Considering the cost of testing, they take small samples from the target population and attempt to estimate the true parameter. The larger the sample size, the closer the estimate will be to the true parameter value. However, the larger sample size increases the cost of testing (see Sect. 9.2.2). There is a trade-off between the accuracy of the test results and the cost of conducting tests.

In order to determine the optimal sample size, marketers need to consider various qualitative and quantitative factors including the cost/benefit of the correct decision-making, the level of prior knowledge for the true parameter value, the (expected) incident/response rates, the desired level of precision, etc. For example, larger sample size will be preferred if the benefit of correct decision-making is great.

Considering the complexity of determining the sample size, several authors have often provided some practical guidelines. For example, Schmid (1995) has suggested a rule of thumb, so-called the Rule of 100. It says that one should have a minimum of 100 responses for each cell. According to this rule, if you expect a 2% response, the sample size should be at least 5,000 to get 100 responses. Alternatively, Levin and Zahavi (1996) suggest that the sample size should be around 10% of the size of the population.

Although these heuristic rules are practically simple in determining the sample size, there is no compelling evidence on why these rules correctly yield an optimal sample size. In essence, these rules intentionally ignore various factors influencing the optimal sample size in order to provide a simple guidance to practitioners.

9.4.1 Statistical Approach

A more formal method to determine the sample size is based on traditional statistical inference. Most marketing research textbooks provide the statistical formula to determine the sample size required to achieve a given level of precision at a desired level of confidence (Tull and Hawkins 1993). The formula generally is provided in two forms: one for the estimation of means (e.g., mean order amount) and the other for proportions (e.g., response

probability). The sample size for mean can be derived from the following:

$$z = (\overline{X} - \mu)/\sigma_{\overline{X}} = D/\sigma_{\overline{X}} = \frac{D}{\sigma/\sqrt{n}} \tag{9.3a}$$

where z is the "z-value" from the standard normal distribution corresponding to the desired level of confidence, $\overline{X}$ is the sample mean, μ is the population mean, $\sigma_{\overline{X}}$ is the standard error of the sample mean, σ is the population standard error, D is the level of precision, and n is the sample size. Similarly, in the case of proportions, we use the following formula:

$$z = (p - \pi)/\sigma_p = D/\sigma_p = \frac{D}{\sqrt{\pi(1-\pi)/n}} \tag{9.3b}$$

where p is the sample proportion, π is the population proportion, and σ_p is the standard error of the sample proportion. From Equations 9.3a, b, we can solve for the sample size for the sample mean and the sample proportion to achieve a given level of precision (D) at a desired level of confidence (z):

$$\text{Sample size for estimating means: } n = \sigma^2 z^2/D^2 \tag{9.4a}$$

$$\text{Sample size for estimating proportions: } n = \pi(1-\pi)z^2/D^2 \tag{9.4b}$$

For example, the population (or true) response rate for the catalog (π) is 1%. And the cataloger wants 95% confidence (hence, its z-value is 1.96) and allows the error (of the estimate) to be within 20% of the population response rate. Then, the optimal sample size should be about $\{(0.01)(0.99)(1.96)^2\}/\{(0.2)(0.01)\}^2 \approx 9,508$.[3]

If more than 10% of the population is included in the sample, the finite population corrections to the above formula are often applied. That is, the correction factor should be incorporated into Equations 9.3a, b as in the following.

$$z = \frac{D}{\sigma\sqrt{(N-n)/(N-1)}/\sqrt{n}} \tag{9.5a}$$

$$z = \frac{D}{\sqrt{\pi(1-\pi)}\sqrt{(N-n)/(N-1)}/\sqrt{n}} \tag{9.5b}$$

where N is the size of the population. Solving Equations 9.5a, b with respect to n, we have:

$$n_c = \frac{nN}{N+n-1} \tag{9.6}$$

[3] Note Equation 9.4b, the case of proportions, is somewhat paradoxical because it says we need to know the true response rate, π, in order to figure out the sample size we need to estimate π! However, often managers have some idea what to expect for a response rate. For example, if one were trying to estimate the response rate to a direct mailing, and management was willing to assume the response rate will be approximately 1%, the value $\pi^o = 0.01$ would be inserted in Equation 9.4b, where π^o is the *a priori* "guesstimate" of the true proportion π.

Table 9.1 Optimal sample sizes for the sample proportion p and the precision D

	$D = x$ % of π^{oa}			
π^{oa}	$x = 5\%$	$x = 10\%$	$x = 20\%$	$x = 30\%$
0.01	152,127[b]	38,031	9,508	4,226
0.05	29,196	7,299	1,825	811
0.10	13,830	3,457	864	384
0.20	6,147	1,537	384	171

[a] π^{o} represents an *a priori* estimate for the true proportion π, to be estimated by the sample proportion p.

[b] The samples sizes are calculated assuming $z = 1.96$, or 95% confidence.

where n_c is the adjusted sample size and n is the unadjusted sample size in Equations 9.3a, b. Note that the population size is very large, no correction is required since $n_c \approx n$.

One needs to determine three unknown values to determine the optimal sample size statistically: the population variance, the degree of confidence, and the desired level of precision. Estimates of the population variance, σ^2 or $\pi(1 - \pi)$, sometimes are available from similar previous studies (for the case of proportions, see footnote 3). If there is no secondary source, one may conduct a pilot study or simply rely on researcher's judgment. The two other unknowns are determined based on the researcher's subjective judgment. That is, we need to specify the level of precision (D) that is the maximum permissible difference between the sample mean/proportion and the population mean/proportion. We also need to specify the z value associated with the confidence level. For example, for a 95% confidence level, the probability that the difference between the population mean/proportion and the sample mean/proportion will be within the specified precision is 95%. The corresponding z value is 1.96. Table 9.1 shows the samples sizes required to estimate the population proportion π at a level of precision D.

In various database marketing applications, the level of confidence is typically assumed to be 95% (corresponding to $z = 1.96$). The true proportion π depends on application, but a response probability of 1% is not unusual in direct mail solicitations. The level of precision D may be acceptable if it is within 20% of the actual proportion. For example, given the true response rate of 10%, the estimated response rate of 8–12% is acceptable. Table 9.1 indicates that the optimal sample sizes for a typical database marketing application should be in the 1,000s, not 100s.

Summarizing, the statistical way of determining the sample size is theoretically sound. However, it is not very practical in that three unknown parameters should be specified quite subjectively to determine the sample size.

9.4.2 Decision Theoretic Approach

Considering both the statistical properties of the test samples and the economic factors, Pfeifer (1998) has proposed a practical method to determine

the optimal sample size. His approach is decision-theoretic in that the optimal sample size is considered a business decision and, hence, the economic trade-offs should be carefully evaluated for the increase of sample size. Here we briefly describe Pfeifer's approach to determining the sample size. Even though Pfeifer applied the approach to a (direct) test-mailing problem, it can easily be applied to other database marketing situations.

9.4.2.1 Problem Definition

A direct marketer needs to decide the number of names to mail in a test (the sample size = n) from a total of N names (the size of population = N). The fixed cost for the test mailing is A and its unit variable cost is C. Let r be the number of responses from the test mailing and V be the net present value to the firm for a given response. Once r is observed, the direct marketer will decide whether to send the mail to the remaining $N - n$ names in the population. The fixed and variable cost for the rollout mailing is reasonably assumed to be the same as in the test mailing. Let r_R be the number of responses to the rollout mailing and V_R be the net present value to the firm for the corresponding rollout response.

9.4.2.2 Prior Response Probability

The key parameter of Pfeifer's model is the uncertain population response rate, π. The population response rate is the unknown probability that a randomly chosen name will respond to the offer. From her experience, the direct marketer is assumed to have a prior distribution for π. More specifically, Pfeifer assumes that the prior distribution of π follows a beta distribution with parameters $a = n_0\pi_0$ and $b = n_0\ (1 - \pi_0)$.

$$f(\pi) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)}, \quad 0 \le \pi \le 1, a > 0, b > 0 \tag{9.7}$$

The parameters a and b, or n_0 and π_0, are the means by which the direct marketer expresses her prior knowledge on the population response rate π. These two parameters are the required inputs to determine the optimal sample size in Pfeifer's model. The parameter π_0 may be interpreted as the direct marketer's best guess for the population response rate and the parameter n_0 as the level of uncertainty in her guess.[4] The prior information incorporated in π_0 and n_0 is equivalent to the information that can be obtained from $\pi_0 n_0$ responses out of the (test) mailing to n_0 customers. For example, suppose

[4] The mean of a beta distribution is $a(a+b)^{-1} = \pi_0$. Hence, π_0 can be regarded as the direct marketer's best guess for the population response rate. Similarly, the variance of a beta distribution is $ab(a+b)^{-2}(a+b+1)^{-1} = \pi_0\ (1-\pi_0)(n_0+1)^{-1}$. Hence, n_0 measures the level of uncertainty.

that 1,000 customers receive catalogs and 20 customers respond. Then the population response rate is estimated to be $\pi_0 = 0.05$ and the variance of π_0 is $\pi_o(1-\pi_o)/n_o = (0.05)(0.95)/1,000$. Therefore, values of $\pi_0 = 0.05$ and $n_0 = 1,000$ would mean that the decision-maker was 95% confident the true response rate was somewhere within $\pm 1.96\sqrt{(0.05 \times 0.95)/1,000} = \pm 0.018$ of $\pi_0 = 0.05$.

Given the prior distribution on π, the direct marketer sends test mailings to n names and gets r responses. Observing the test mailing results, the direct marketer updates her estimate on π. Her updated probability forecast of π or the posterior distribution of π can be written as

$$f(\pi|n,r) = \frac{\pi^{r+a-1}(1-\pi)^{n-r+b-1}}{B(r+a, n-r+b)} \tag{9.8}$$

9.4.2.3 Calculating the Expected Rollout Profit

The direct marketer will decide whether to roll out to the remaining $N-n$ names in the population after she gets the test mailing results. The profit from the rollout is

$$\text{Profit}_R = V_R r_R - (N-n)C \tag{9.9}^5$$

A risk-neutral marketer will roll out if the expected rollout profit, $E(\text{Profit}_R)$, is greater than zero. Hence, we compute the expected number of responses from the rollout mailing, $E(r_R)$. Noticing that r_R is distributed as a beta-binomial, its mean can be written as (Johnson and Kotz 1969)

$$E(r_R) = (N-n)\frac{r+a}{(r+a)+(n-r+b)} = (N-n)\frac{r+n_0\pi_0}{n_0+n} \tag{9.10}$$

Therefore, the expected rollout profit is

$$E(\text{Profit}_R) = V_R E(r_R) - (N-n)C = (N-n)\left[V_R\frac{n_0\pi_0 + r}{n_0+n} - C\right] \tag{9.11}$$

As mentioned, the direct marketer should roll out the population if $E(\text{Profit}_R) > 0$. Hence, the direct marketer should roll out the list if

$$r > C(N-n)(n_0+n)[V_R(N-n)]^{-1} - n_0\pi_0 \tag{9.12}$$

Let r^* be the smallest integer that satisfies the Equation 9.11. Then the direct marketer should roll out if the number of responses from the test mailing is greater than and equal to r^* and should not if the number is less than r^*.

[5] Pfeifer (1998) included the fixed cost term for test mailing (A) in Equation 9.9. However, we delete it to simplify our exposition. Our key results do not change without the fixed cost. In addition, Pfeifer himself mentioned that fixed costs can be negligible if the test mailing is included as part of a regular mailing.

9.4.2.4 Selecting the Optimal Sample Size

In order to determine the optimal test sample size, let us consider the expected profit including both the test and the rollout mailing. If $r < r^*$ for the test mailing result, the direct marketer will not roll out and, hence, the profit (from the test) becomes $\text{Profit}_T = Vr - nC$. Alternatively, if $r \geq r^*$ for the test mailing result, the direct marketer will roll out. And the resulting profit (from both the test and the rollout) becomes $\text{Profit}_T + \text{Profit}_R = Vr - nC + V_R E(r_R) - (N - n)C$. That is, the total profit is a function of r. Since the probability distribution of r is the beta-binomial, the expected (total) profit becomes

$$\begin{aligned} & E(\text{Profit}_T + \text{Profit}_R | n) \\ & = \sum_{r=0}^{r^*-1} g(r|n)(Vr - nC) + \sum_{r=r^*}^{n} g(r|n)[Vr - nC + V_R E(r_R) - (N-n)C] \\ & = Vn\pi_0 - nC + \sum_{r=r^*}^{n} g(r|n)[V_R(N-n)(n_0\pi_0 + r)(n_0 + n)^{-1} - (N-n)C] \end{aligned} \tag{9.13}$$

where $g(r|n)$ is the beta-binomial density for r.

9.4.2.5 Illustrative Example

Given the test sample size n, the direct marketer can calculate the expected profit from Equation 9.13. To find the optimal sample size, the direct marketer evaluates Equation 9.12 for various candidate values of n and selects the one that maximizes the expected profit. Let us provide an illustrative example given by Pfeifer (1998). A house list consists of 50,000 customers and it costs \$1 to mail. The direct marketer's best guess for the population response rate (π_0) is 2% and the corresponding level of uncertainty (n_0) is assumed to be 200. With this prior specification, Pfeifer implicitly assumes that the response rate of a given house list is a randomly drawn number from a beta distribution with $\pi_0 = 2\%$ and $n_0 = 200$. The responses to the mailing are worth \$50 each. So mailing to all 50,000 customers in the list will result in \$0 expected profit. It will cost $\$50,000$ ($= \$1 \times 50,000$) which is equal to the expected revenue of $\$50,000$ ($= \$50 \times 2\% \times 50,000$). Without a test mailing, the direct marketer will be indifferent between mailing to all 50,000 customers and doing nothing.

The values of the parameters in Equation 9.12 are all determined. The appropriate parameter values are: $N = 50,000, C = \$1, V = V_R = \$50, \pi_0 = 0.02$ and $n_0 = 200$. Figure 9.4 shows the plots of expected profits as a function of the test sample sizes n. It indicates that a test sample size of about 2,000 maximizes the expected profit. The expected profit with the test sample of

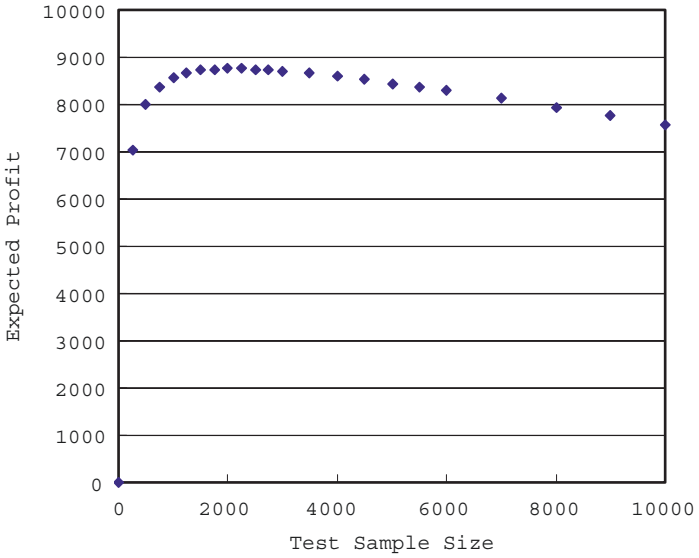

Fig. 9.4 Test sample size versus expected profit using Pfeifer (1998) Model (From Pfeifer 1998).

0 is \$0, because the expected response rate of a house list is the breakeven response rate. Also, the expected profit is \$0 if the test sample size $N = 50,000$, because there would then be no rollout and the expected profits for the test would just depend on the expected response rate which is at breakeven. For any test sample sizes between 0 to 50,000, we have the option of rolling it or not based on the sample information from the testing results. If the test indicates the full roll-out will not be unprofitable, the manager will lose money on the test but avoid the even larger loss of an incorrect roll-out. However, if the test indicates the full roll-out will be profitable, the manager makes money on the test and then makes even more on the full launch. That is, there is 50% chance that the test results suggest the full roll-out and 50% chance that the test results suggest no roll-out. But we lose some money when the test results are bad, whereas we can make big money when the test results are good. As a result, the manager on average makes money through testing, and the optimal test size is about 2,000 out of a population of 50,000, or 4%.

It also is important to note that the expected profit numbers that emerge from this analysis combine subjective and objective information. In a classical statistical sense, the expected profits could be a biased estimate of the true expected profits if the manager's prior for the response rate does not on average equal the true response rate. For example, if the manager's prior is overly optimistic, and overly confident in that prior, the test information will have little impact on the updated response rate and the expected profits will be overly optimistic. However, it can be argued that managers learn the

average response rates across lists, and so their prior is not overly optimistic, and if they are not confident in the performance of the particular list to be tested, they can indicate low confidence through a small value for n_0.[6]

9.4.2.6 Extending Pfeifer's Model

This approach is very promising but could be extended in several ways. Foremost would be the incorporation of a control group. Note that the method assumes a rollout should occur if the expected profit from the rollout is greater than zero. However, this assumes that if no action is taken, no profit is generated. This may be the case for a direct marketer who is thinking of a program of contacting people who are not current customers, which is the orientation of Pfeifer's paper. But if the company has current customers, there will be profits from those customers even if the action is not taken. For example, if Product A is not cross-sold to the customer, the customer may buy it anyway through a different channel. These profits are uncertain as are the profits that might accrue from directly contacting the customer. Therefore there is uncertainty if the action is taken or not taken. This necessitates a control group. The question then becomes, what should be the size of the test group and the control group. Control groups are often used in testing and it seems the above approach could be extended to this situation.

Another extension would be to incorporate uncertainty in the value of the customer, V. If the decision is whether to send a catalog, V represents customer expenditure given the customer responds. This number will also be uncertain. An extension would be to incorporate priors on this quantity as well. Obviously, the more diffuse those priors are, the higher sample size will be needed.

In summary, Pfeifer's approach is a practical tool for deciding sample size, directly applicable to customer acquisition tests. Extending the method as discussed above would provide important and interesting avenues for future research. It also should be noted that the usefulness of the model hinges on the validity of the manager's prior. If the manager states a highly optimistic prior with great certainty, he or she is likely to calculate positive expected profits from a roll-out no matter what the test results, and lose money. The key point is that the expected profit calculations are a combination of objective evidence from the test and subjective judgment encompassed in the prior, and therefore essentially a subjective judgment of expected profits. Having noted this limitation, the model is still valuable because managers use judgments all the time in deciding whether to undertake a full roll-out. The model merely captures those judgments rigorously and calculates the implications for profitability.

[6] The authors thank Phil Pfeifer for helpful insights on presenting and discussing this model.

9.5 Test Designs

Experimental research generally consists of three phases: the experimental or planning phase, the design phase, and the analysis phase. In this section we focus our attention on the design phase and somewhat on the analysis phase. Once the objective of the research is set in the planning phase, the research problem should be expressed in terms of a testable hypothesis. For example, a cataloger would like to know whether the new catalog design increases the response rates among current customers. A mobile telecommunication service provider wants to know whether churn rates are higher among customers aged below 25. It is then time to design the experiment. In this section we study the test designs that are most popular among database marketers.

9.5.1 Single Factor Experiments

Single factor design is the simplest test design and is fundamental for understanding more complex designs. This section discusses single factor experiments in which no restrictions are placed on randomization. Randomization refers to the random assignment of sample units to experimental (or control) groups by using random numbers. Treatment conditions are also randomly assigned to experimental groups. For example, a credit card company is contemplating whether to make a promotional offer to increase card usage. It comes up with an idea of offering coupons on gas purchases. Ten randomly selected customers are given $5 coupons and another ten randomly selected customers are offered $10 coupon on gas purchase. It randomly selects an additional ten customers who do not receive any promotional offers. This is called a control group.[7] The card usages of 30 customers for a month after the experiment are shown in Table 9.2. Then single factor model becomes

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij} \tag{9.14}$$

where Y_{ij} represents the ith observation ($i = 1, 2, \ldots, n_j$) on the jth treatment ($j = 1, 2, \ldots, k$). For example, the third observation in control condition (Y_{31}) is 600 in Table 9.2. μ is a common effect for the whole experiment, τ_j is the treatment effect of jth condition, and ε_{ij} is a random error.

We usually assume that the error term ε_{ij} is distributed as *i.i.d.* normal with zero mean and the common variance. That is, $\varepsilon_{ij} \sim N(0, \sigma^2)$. It is also assumed that the sum of all treatment effects is zero, or $\sum_{j=1}^{k} \tau_j = 0$. To

[7] In practice, a control group is defined as the group that receives the current level of marketing activity or receives no treatment at all. A control group is included to ascertain the true *incremental* effect of the treatments versus no treatment. For example, customers still use credit cards without promotional coupons. Hence, the true experimental effect of promotional coupon offers is the incremental card usage from the promotional coupon over the credit card usages among the control group customers.

Table 9.2 Credit card usage data for single factor experiment

Customer Id	Treatment conditions	Card usage
1	Control	$500
2	Control	$550
3	Control	$600
4	Control	$450
5	Control	$500
6	Control	$400
7	Control	$450
8	Control	$550
9	Control	$550
10	Control	$500
11	$5 coupon	$550
12	$5 coupon	$600
13	$5 coupon	$700
14	$5 coupon	$650
15	$5 coupon	$700
16	$5 coupon	$550
17	$5 coupon	$750
18	$5 coupon	$650
19	$5 coupon	$600
20	$5 coupon	$700
21	$10 coupon	$700
22	$10 coupon	$750
23	$10 coupon	$700
24	$10 coupon	$800
25	$10 coupon	$600
26	$10 coupon	$700
27	$10 coupon	$750
28	$10 coupon	$800
29	$10 coupon	$700
30	$10 coupon	$750

describe the basics of a one-way analysis of variance (ANOVA), we rewrite Equation 9.14:

$$Y_{ij} = \mu + (\mu_{.j} - \mu) + (Y_{ij} - \mu_{.j}) \text{ or } Y_{ij} - \mu = (\mu_{.j} - \mu) + (Y_{ij} - \mu_{.j}) \quad (9.15)$$

where $\mu_{.j}$ is the expected value of Y_{ij} given the customer receives treatment j. Comparing Equations 9.14 and 9.15, the jth treatment effect τ_j can be represented by $\mu_{.j} - \mu$.

Since the means in Equation 9.15 are not known, they are estimated from the n_j observations for each treatment condition j. These observations can be used to estimate the grand mean μ and the treatment means $\mu_{.j}$. Restating Equation 9.15 in terms of "sample means," we obtain:

$$Y_{ij} - \overline{Y}_{..} = (\overline{Y}_{.j} - \overline{Y}_{..}) + (Y_{ij} - \overline{Y}_{.j}) \quad (9.16)$$

where $\overline{Y}_{..}$ is the sample (grand) mean over all observations, $\overline{Y}_{.j}$ is the sample mean over the observations with treatment condition j. The equation says

that the deviation of each observation from the overall mean consists of two parts: the deviation of the treatment mean from the overall mean and its deviation from its own treatment mean.

Taking squares and summations of Equation 9.16, we have:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{..})^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{Y}_{.j}-\overline{Y}_{..})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{.j})^2 \qquad (9.17)$$

The term in the left is called the total sum of squares while the first term in the right is called the between groups/treatments sum of squares and the second term is called the within groups sum of squares or the error sum of squares. Equation 9.17 says that the total sum of squares is equal to the between groups sum of squares plus the error sum of squares.

The main interest in single factor experiment is to test whether there are treatment effects. That is, we conduct a one-way analysis of variance test where the hypothesis to be tested is $H_{0:}\ \tau_j = 0$ for all j. If the hypothesis is accepted, we conclude that there are no treatment effects and all the variations in the dependent variable Y_{ij} are explained by the grand mean μ and the random error ε_{ij}.

Going back to the Equation 9.17, it can be shown that the between-group sum of squares divided by its degree of $(k-1)$, called mean squares, is distributed as chi-square. Similarly, the error sum of squares divided by its degree of freedom $\sum_{j=1}^{k}(n_j-1)=(N-k)$ is also distributed as chi-square. And since these two chi-squares are independent, their ratio can be shown to be distributed as F with degrees of freedom, $(k-1)$ and $(N-k)$. Therefore, if H_0 is true, we can test the hypothesis by evaluating the following quantity.

$$F_{k-1,N-k}=\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{Y}_{.j}-\overline{Y}_{..})^2/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{.j})^2/(N-k)} \qquad (9.18)$$

The quantity in the numerator becomes larger when the deviations of the treatment means from the grand mean become larger. Hence, we reject the null hypothesis if the quantity in the Equation 9.18 is larger than the critical region $F_{1-\alpha}$, where α is the designated significance level.

An one-way analysis of variance is applied to the credit usage data in Table 9.2 and its results are summarized in Table 9.3. The test statistic for the hypothesis $H_{0:}\ \tau_j = 0$ for all $j = 1, 2, 3$ is $F^* = 124,000/3,916.7 \approx 31.7$ that is larger than the critical value $F_{2,27} = 5.45$ at the significant level $\alpha = 0.01$.[8] Hence, we reject the hypothesis and conclude that there are statistically

[8] Most of statistical software can handle an one-way analysis of variance and provide the ANOVA summary table similar to Table 9.3. See "PROC ANOVA" in SAS, "ANOVA Single" under Data Analysis & Tools in EXCEL, and "one-way ANOVA" under Compare Means & Statistics in SPSS.

Table 9.3 One-way ANOVA for credit card usage data

Source of variation	Degree of freedom	Sum of squares	Mean squares
Between groups	2	248,000	124,000
Within groups	27	105,750	3,916.7
Total	29	353,750	–

$F^* = 124,000/3,916.7 \approx 31.7 > F_{2,27} = 5.45$ at the significant level $\alpha = 0.01$

significant differences in credit card usages among different amount of coupon offers on gas purchases.

9.5.2 Multifactor Experiments: Full Factorials

Database marketers often need information on a wide variety of strategic issues. For example, a credit card company manager attempts to devise an optimal promotional package to increase credit card usages among current customers. She has more than one tactic of interest. For example, there might be three tactics, or three factors to test in the experiment: (1) the use of coupons for gas purchases, (2) the use of cash rebates, and (3) the use of "affinity" cards. The manager is interested in the impact of all three of these marketing strategies on credit card usage rate. One method is to hold all other factors constant except one and observe the effects over several levels of this chosen factor. Alternatively, one can perform a full-factorial experiment in which all levels of a given factor are combined with all levels of every other factor.

A full-factorial experiment is superior to the one-at-a-time experiment in several aspects (Hicks 1982). A factorial experiment will provide greater statistical efficiency since all data are used in computing the effect of each factor. In addition, we can evaluate interactions among factors with a factorial experiment. This design component is particularly important because we frequently observe synergistic effects among marketing variables.

To see how a full-factorial experiment works, let us again consider an example of a credit card company that is considering a promotional offer to increase the card usage. Now the manager wants to look at the effects of two promotional variables on card usage: coupons for gas purchases and a cash rebate for credit card usage. Three levels of coupon amount ($0, $5 and $10) and two levels of cash rebate (0% and 1%) are considered. That is, it is a 3×2 factorial experiment, yielding six possible combinations of coupon amount and rebate level. Five customers are randomly selected at each of these six treatment conditions. Table 9.4 shows the (monthly) credit card usage of these 30 customers after the experiment.

The mathematical model for this experiment can be written as

$$Y_{ijk} = \mu + \tau_{jk} + \varepsilon_{i(jk)} \tag{9.19}$$

Table 9.4 Credit card usage data for factorial experiment

Customer Id	1st treatment conditions	2nd treatment conditions	Card usage
1	Control	Control	$450
2	Control	Control	$500
3	Control	Control	$450
4	Control	Control	$400
5	Control	Control	$450
6	Control	$5 coupon	$500
7	Control	$5 coupon	$500
8	Control	$5 coupon	$600
9	Control	$5 coupon	$400
10	Control	$5 coupon	$500
11	Control	$10 coupon	$500
12	Control	$10 coupon	$550
13	Control	$10 coupon	$550
14	Control	$10 coupon	$500
15	Control	$10 coupon	$500
16	1% cash rebate	Control	$500
17	1% cash rebate	Control	$450
18	1% cash rebate	Control	$500
19	1% cash rebate	Control	$450
20	1% cash rebate	Control	$470
21	1% cash rebate	$5 coupon	$650
22	1% cash rebate	$5 coupon	$700
23	1% cash rebate	$5 coupon	$700
24	1% cash rebate	$5 coupon	$650
25	1% cash rebate	$5 coupon	$600
26	1% cash rebate	$10 coupon	$800
27	1% cash rebate	$10 coupon	$850
28	1% cash rebate	$10 coupon	$900
29	1% cash rebate	$10 coupon	$800
30	1% cash rebate	$10 coupon	$950

where the subscript $j(j = 1, 2, 3)$ represents the levels of coupon amounts, the subscript $k(k = 1, 2)$ represents the levels of cash rebates and the subscript $i(i = 1, 2, 3, 4, 5)$ represents the number of observations/customers for each treatment condition j and k. For example, Y_{422} is 650 in Table 9.4. Similar to the single-factor experiment, μ is the grand mean for the whole experiment, τ_{jk} is the treatment effect for jth coupon condition and kth rebate condition, and $\varepsilon_{i(jk)}$ is a random error.

Treating each treatment condition as unique, the model in Equation 9.19 does not consider the factorial or multifactor nature of the experiment. That is, we apply a one-way ANOVA to the data in Table 9.4 and summarize the results in Table 9.5a.

The test statistic for the hypothesis $H_{0:}\ \tau_{jk} = 0$ for all $j = 1, 2, 3$ and $k = 1, 2$ is $F^* \approx 54.0\ (= 122,687.5/2,270.8)$ that is statistically significant at the significance level of 1%. Hence, we reject the hypothesis and conclude that there are statistically significant differences in credit card usages among different coupon offers and cash rebates.

Table 9.5a One-way ANOVA for factorial data

Source of variation	Degree of freedom	Sum of squares	Mean squares
Between groups	5	613,437.5	122,687.5
Within groups	24	54,500	2,270.8
Total	29	353,750	–

We can slightly modify Equation 9.19 to represent the multi-factor nature of the factorial experiments. Decomposing τ_{jk} into the main effect of coupon treatment condition (C_j), the main effect of cash rebate treatment condition (R_k), and their interactions (CR_{jk}), Equation 9.19 can be rewritten as

$$Y_{ijk} = \mu + C_j + R_k + CR_{jk} + \varepsilon_{i(jk)} \tag{9.20}$$

The two-way ANOVA is the appropriate tool to analyze the model in Equation 9.20, the general model for a two-way factorial experiment. The main interests in two-way factorial experiment are three tests: (1) whether there is a main treatment effect of coupon ($H_0 : C_j = 0$ for all $j = 1, 2, 3$), (2) whether there is a main treatment effect of cash rebate ($H_0 : R_k = 0$ for all $k = 1, 2$), and (3) whether there is an interaction effect between coupon and cash rebate ($H_0 : CR_{jk} = 0$ for all $j = 1, 2, 3$ and $k = 1, 2$). If the hypothesis $H_0 : C_j = 0$ is accepted, we conclude that coupon amounts on gas purchase will not affect on credit card usage. Similar conclusions will be derived from other tests.

We apply a two-way ANOVA to the data in Table 9.4 and summarize the results in Table 9.5b.

The between group sum of squares in Table 9.5a (613,437.5) is now decomposed into three sums of squares in Table 9.5b: between coupons sum of squares (258,875), between rebates sum of squares (229,687.5) and the coupons × rebates interaction sum of squares (124,875). Table 9.5b also shows that each of the two main effects and the interaction effect are statistically significant at the 1% level.

The significant coupons × rebates interaction implies that a change in one factor produces a different change in the response variable at one level of the other factor than at the other levels of this factor. The interaction can be more clearly seen in Fig. 9.5 where the mean card usages (over each of five

Table 9.5b Two-way ANOVA for factorial data

Source of variation	Degree of freedom	Sum of squares	Mean squares	F
Between coupons	2	258,875	129,437.5	57.0
Between rebates	1	229,687.5	229,687.5	101.1
Coupons × rebates	2	124,875	62,437.5	27.5
Errors	24	54,500	2,270.8	–
Total	29	667,937.5	–	–

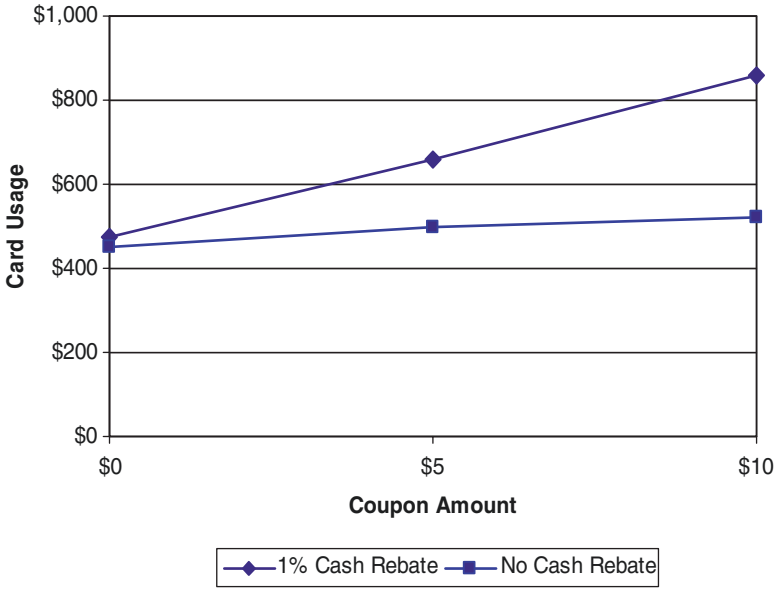

Fig. 9.5 Interaction in a factorial experiment.

customers) for each treatment condition are graphed. Given no cash rebates, the mean card usages are \$450 for no coupons, \$500 for \$5 coupons, \$520 for \$10 coupons. On the other hand, with 1% cash rebates, the mean card usages are \$475 for no coupons, \$660 for \$5 coupons, \$860 for \$10 coupons. That is, there exist positive synergies between coupons and cash rebates. Or coupons are more effective when they are used with 1% cash rebates. If there are no coupons × rebates interactions, the card usage plot of no cash rebates should be parallel to the card usage plot of 1% cash rebates.

9.5.3 Multifactor Experiments: Orthogonal Designs

A full-factorial experiment is very useful to database marketers since several factors are simultaneously considered and, hence, all interaction effects can be identified. However, as the number of factors considered in a factorial experiment increases, the number of treatment conditions increases very rapidly. For example, it is not unusual for database marketers to consider 5 factors where each factor has three levels. The number of treatment conditions for this factorial experiment is 245 ($= 3^5$). It is not economical – sometimes, it is not even feasible – to assign customers to each of 245 treatment conditions. In order to overcome this problem, researchers use "fractional factorial" designs, where only a fraction of all possible treatment combinations is selected for

Table 9.6 Orthogonal array for 2^9 factorial design

Combination	Factors and levels								
	A	B	C	D	E	F	G	H	I
1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	1	2	2
3	1	1	1	2	1	2	2	1	2
4	1	1	1	2	2	1	2	2	1
5	1	2	2	1	1	1	1	2	2
6	1	2	2	1	2	2	1	1	1
7	1	2	2	2	1	2	2	2	1
8	1	2	2	2	2	1	2	1	2
9	2	1	2	1	1	1	2	1	2
10	2	1	2	1	2	2	2	2	1
11	2	1	2	2	1	2	1	1	1
12	2	1	2	2	2	1	1	2	2
13	2	2	1	1	1	1	2	2	1
14	2	2	1	1	2	2	2	1	2
15	2	2	1	2	1	2	1	2	2
16	2	2	1	2	2	1	1	1	1

testing (Hicks 1982). In fractional factorial designs one willingly gives up the measurement of all possible interaction effects and obtains a smaller number of treatment conditions.

A special class of fractional factorial designs, called orthogonal arrays, is a highly fractional design in which all main effects can be identified although all interaction effects are assumed to be negligible (Green 1974). Widely used in conjoint analysis experiments, orthogonal arrays are known to be the most parsimonious set of designs (in the sense of the lowest number of treatment conditions) available for estimating main-effect parameters. For example, we consider an experiment with 9 factors where each factor has two levels. Hence, the number of treatment conditions for the full factorial experiment is 512 ($= 2^9$). Table 9.6 shows an orthogonal array for this experiment that was provided by Addelman (1962). In the case of orthogonal arrays a necessary and sufficient condition for the main effects of any two factors be uncorrelated (unconfounded) is that each level of one factor occurs with each level of another factor with proportional frequency.

Assuming that all interaction effects can be neglected, we can reduce the number of treatment conditions from 512 to 16. With relatively few treatment combinations, orthogonal arrays allow us to estimate all main effects on an unconfounded basis for a dozen or more factors, each at two or three levels.

The concept of "confounding" is important in fractional designs and merits some elaboration. Assume we have a three-factor experiment where we want to test three treatments, each at two levels. Let's say the experiment is for a credit card and the factors are coupon (yes or no), rebate (yes or no), and

Table 9.7 Two potential fractional designs for a 2^3 experiment

	Design 1			Design 2		
Treatment	Coupon	Rebate	Affinity	Coupon	Rebate	Affinity
1	Yes	Yes	Yes	Yes	Yes	Yes
2	Yes	Yes	No	Yes	No	No
3	No	No	Yes	No	Yes	No
4	No	No	No	No	No	Yes

affinity card (yes or no). A full factorial experiment would have $2^3 = 8$ combinations. Let's assume eight combinations are impractical for the researcher so we want to design an experiment with just four combinations. Table 9.7 shows two possible designs. Which is the better design? Experiment 1 runs into the problem of "confounding." The confounding is between the Coupon and Rebate. With this experiment, we will not be able to differentiate the impact of the coupon from that of the rebate, because every treatment group that gets a coupon also gets a rebate, and every group that does not get a coupon does not get a rebate. If card usage is higher for groups 1 and 2, we don't know if it due to the coupon or the rebate. There is no way to differentiate these effects. In contrast, Experiment 2 has no confounds between any pairs of the three factors. If treatment groups 1 and 2 have higher usage rates, that can be interpreted as due to the coupon, because treatment groups 1 and 2 both always have coupons but sometimes have rebates or affinities and sometimes not. Similarly, groups 3 and 4 never have a coupon but sometimes have rebates or affinities and sometimes not. Experiment 2 is the preferred design.

Lists of orthogonal designs (e.g., Table 9.6) are provided by Plackett and Burman (1946) and Addelman (1962), making it easier to develop orthogonal arrays. SPSS also provides a routine for creating orthogonal arrays.

The price of the orthogonal array is that it assumes there are no interaction effects, whereas there may be interaction effects as we saw in the credit card example. Another way to state the assumption is that the orthogonal array cannot differentiate between a main effect and various interactions, so we just assume there are no interactions and that the main effects we estimate just reflect main effects and nothing else. This is somewhat troublesome but often main effects are clear and important, and interactions are indeed secondary. There are in fact intermediate-type fractional factorial designs where a fraction of all possible combinations are selected so that at least some of the interactions can be estimated (see Winer 1971 for a thorough treatment). These fractional factorials of course will require more treatments.

9.5.4 Quasi-Experiments

As the name implies, a quasi-experiment is almost a true experiment. A quasi-experiment is where we are unable to fully manipulate the scheduling

or assignment of treatments to test units (Malhotra 1993). There are many types of quasi-experiments, but their common feature is that the assignment of treatments to customers is not controlled by the researcher.

Quasi-experiments are therefore used in database marketing when it is difficult to randomly assign customers to treatment conditions. For example, we may want to evaluate the impact of a customers' participation in a reward program on their purchase frequencies. We offer the reward program to all customers and let customers decide whether they participate the program or not. Suppose that 40% of customers participate and the rest do not. Monthly purchase dollars before and after launching the rewards program are measured. Program participants increase their purchase dollars from \$100 to \$120 as a result of the rewards program. Purchase dollars of non-participants are also increased from \$90 to \$100. Non-participants in this quasi-experiment serve as the control set. Hence, we may conclude that customers increase their purchase dollars by (\$120–100) − (\$100–90) = \$10 due to their program participations. However, this conclusion is misleading since customers were not randomly assigned between program participants and non-participants. There could be a self-selection bias whereby the customers who self-selected into the rewards program were pre-disposed to buy from the company anyway (see Chapter 11, Statistical Issues in Predictive Modeling). A true experimental design would approach this situation by dividing customers randomly into participants and non-participants. The random assignment would eliminate concerns for selection bias.

One way of reducing the selection bias in quasi-experiments is to introduce covariates in analyzing the experimental effect. This is called the analysis of covariance (ANCOVA). ANCOVA tries to control statistically for factors that influence purchase frequency besides membership in a rewards program. One can also develop a formal selectivity model (see Chapter 11, Statistical Issues in Predictive Modeling).

In summary, quasi-experiments are in general less preferred because one loses the randomization of the true experiment. Randomization rules out other factors as causes (on average) and particularly addresses selection bias. However, in the real world, one may not have the luxury of randomizing. In that case, the researcher at a minimum should use an analysis of covariance framework, and consider formulating a formal selectivity model.