Chapter 4

Probability and Statistics Summary

Chapter Outline

What You Will Learn in This Chapter

- How is it possible to make decisions when so many events are random?
- What is randomness and what rules does it follow?
- How are probabilities handled when events are not independent?
- How can probability concepts be generalized to for handling new situations?
- How are statistics used in data mining to find interesting results?

Human Biases

In general, people make many mistakes when evaluating data. One particular problem is known as "confirmation bias," where people seek information that confirms what they already believe. Business people and even professional researchers are prone to this problem. When we create theories, we like to "prove" that they are right. A psychology study with nearly 8,000 participants concluded that people are twice as likely to seek information that confirms their beliefs instead of seeking out conflicting data to disprove them. Psychologist Scott Lilienfeld of Emory University noted that "We're all mentally lazy. It's simply easier to focus our attention on data that supports our hypothesis, rather than to seek out evidence that might disprove it." Consequently, people are reluctant to change their opinions and models, and instead rationalize why things might have gone wrong. And gathering more data makes the problem worse. If you are focusing on confirming beliefs, then new data reinforces those beliefs, reducing the value and diversity of the information available. [Zweig 2009]

Try to collect unbiased data first then analyze it with an open mind. Keep notes during the early stages to record options and uncertainties. Return to those notes later to explore alternatives. Involve other people and look for options and criticisms.

Jason Zweig, "How to Ignore the Yes-Man in Your Head," *The Wall Street Journal*, November 13, 2009. http://online.wsj.com/article/SB10001424052748703811604574533680037778184.html?mod=WSJ_hps_LEADNewsCollection

Introduction

How is it possible to make decisions when so many events are random? Randomness exists in many aspects of the world and business. Many events are subject to fluctuations and occurrences that cannot be defined with certainty. The business world is not deterministic. Instead, the same action taken at two points in time can lead to different results. One manager can make a decision and the results can work out well. A second manager facing the same problem can make the same decision and have everything fail. Of course, the first manager gets a huge bonus and a writes a book proclaiming the brilliance of the decision, and the second manager looks for a new job. But, perhaps instead of brilliant, the first manager was merely lucky. So how is it possible to make decisions, and evaluate decision makers, when the world is random?

The science of probability and statistics was developed specifically to deal with questions of decision making under uncertainty. It defines the concepts of randomness and describes mathematical rules that determine relationships among events. Statistical tools developed over centuries of research form the foundations of data mining and decision making.

The field of probability and statistics is large and complex. It is based largely on mathematics, and some of the most powerful theorems were found with the use of advanced mathematical concepts. Fortunately, it is possible to use the results of this work without needing to understand the heavy mathematics. Yet, to understand the results of some of the tools, it is important to know some of the basic foundations and definitions. This chapter presents the basic concepts needed to use and understand the results of common data mining tools. It begins with a summary of some key probability concepts. It also explains some of the critical concepts in statistics that are commonly used in decision making problems and business intelligence. The chapter will be easier to read if you have already had an introductory course in probability and statistics, but the text does define the fundamental concepts needed for the rest of the book. The main goal of the chapter is to develop a statistical perspective and improve your "intuition" or understanding of how probability and statistics are understood and evaluated in making decisions.

Probability Basics

What is randomness and what rules does it follow? Randomness can arise from many sources, but those details are not critical yet. For the moment, **random** simply means that some events or outcomes cannot be predicted with complete certainty. Instead, an outcome is said to have some probability of arising. The concept of probability is critical to understanding randomness and statistics. Technically, two good definitions of probability exist. Ultimately, the two definitions lead to similar conclusions and applications, but it is sometimes useful to look at the world one way or the other. The oldest and most common definition of probability is known as **relative frequency**: The **probability** of an event arising is equivalent to the number of times the event can arise versus the total number of events that can occur. The simplest example is a coin flip with the events of a head or tail appearing on top. With two total events possible (staying on an edge is considered a bad toss and ignored), the relative frequency or probability of either event is 1/2. Rolling a six-sided die is another common example, where the total number of possibilities is 6 and each side has the same probability of arising, so the chance of any single number appearing is 1/6.

Customer Gender	Car Model	Customer Age	Car Weight	Car Price
Male	Focus	21	2588	15520
Female	TT	42	2965	35200
Unknown	Suburban	36	5607	40370

Figure 4.1

Examples of discrete and continuous data. Gender and Model are discrete because the values are specifically defined and can be counted. Weight, Price, and even Age are continuous data. Although the values displayed are truncated, the attributes could take on any value.

In probability calculations, the list of total events has to be complete, so the probability of any event must always lie between 0 and 1. Writing probability as P, and using a generic outcome A, the probability of any event cannot be less than zero or greater than one:

$$0 \leq P(A) \leq 1$$

The function P(A) is read as "the probability of A" where A is some event. For instance, the probability of tails occurring from a coin toss is 1 out of 2 or 1/2. Probabilities of all outcomes must add to one. Sometimes it is useful to refer to all other outcomes as the negation, or not A. It is often written P(A') and read as "the probability of not A." So, $P(A) + P(A') = 1$.

The second way to define probability is known as the **Bayesian** approach for reasons that will be clear later. In this method, probability is subjective and defined as the degree of belief of some event happening. The interesting twist is that any individual could have a different belief about the probability of an event. The probability must still fall between zero and one, but it is **subjective**. For instance, perhaps the person flipping the coin knows how to flip it so that heads appears more often than half the time. Alternatively, perhaps the probability belief by one person, or entire organization, is biased, or even wrong. As a child with limited experience, perhaps you believed that when flipping a coin, the head could appear more often than tails. Probability theory reveals that someone with this belief would likely be wrong many times, but the beliefs would change with experience or increased information. Subjective probability is an interesting and useful way to approach some problems, and some types of data mining are based on this approach. Instead of looking at coin flips, consider the question of predicting the level of sales for next year. Each range of outcomes has some probability of occurring that is unknown and could be subjective. As more information is collected (say daily or weekly sales), these probabilities could be adjusted to provide a more accurate forecast.

The issue of subjective probability highlights a second key aspect of probability. The actual probability of an event might be unknown. Some events are simple enough to enumerate or count and the probabilities can be defined. With more complex events, estimating the probability is an important step in forecasting and statistics.

Discrete and Continuous Data

Two types of data often exist in problems: **discrete** and **continuous** measures. Discrete events can be counted—there might be an infinite number of options, but each one is unique and can be assigned a number. Examples of discrete data include the coin flip, the sides of a die, the gender of a customer, whether a product is defective, a car model, and the types of products produced in a bakery.

Continuous data is measured in real values instead of integers, and it cannot be counted as simple groups. Examples include the gas mileage of a car, the height of a customer, the temperature of a food item, and the weight of a bicycle. Some measures might first appear to be discrete, but the measurement might be truncated, and it could take on any value. For instance, the age of a person is commonly truncated to years; but if the date of birth is available, the age could be measured in days. In fact, the age of a person is continuous data because time could be divided into any level of measurement. Similarly, the price of an automobile would be continuous data—particularly measured in cents. Figure 4.1 shows some simple examples of discrete and continuous data. The Customer Gender and Car Model are clearly discrete because the options are well-defined. Car Weight, Car Price, and Customer Age are continuous data. The values displayed are truncated to integer values, but given appropriate measuring tools, the attributes could take on any value. Car price might (or might not) be limited to cents, which might appear to make it a discrete measure. But, because cents is such a small fraction of the overall value, prices are typically treated as continuous data.

Continuous and discrete data are treated differently in statistics and so the data mining tools are different as well. Interpreting results and making decisions is somewhat different as well. Because some tools work only with discrete data, it will sometimes be beneficial to convert continuous data into discrete groups. **Discretizing** continuous data is an important step in some of the procedures. For example, the weight of a car could be classified into three discrete categories: (1) Light where weight < 2000 pounds, (2) heavy where weight > 4000 pounds, and (3) medium with weights between 2000 and 4000 pounds. It can be difficult to determine the number of categories and the cutoff values. Sometimes they are established by experts or tradition. In other cases, some tools exist to help identify appropriate categories.

The differences between discrete and continuous data appear in many chapters, and the statistical implications are covered in more detail in the statistical sections of this chapter. For the moment, basic probability concepts are easier to understand with discrete events instead of continuous.

Counting and Combinations

Look again at the basic definition of probability: the number of ways an event can occur divided by the number of total events. For discrete data, this definition requires knowing how to count. Sure, that seems easy; but it means knowing how to count the number of ways a specific event can occur. For simple problems, such as the coin flip or defective/not-defective, the number of cases is easy to specify and counting to two is easy. More complex problems with multiple outcomes and interactions can be harder to count. Also, it can take a while to list thousands or millions of possible outcomes. Mathematicians have developed tools to help count outcomes for certain cases that arise quite often.

Counting when Order Matters

A marketing manager wants to use standard capital letters to assign codes to products. The firm currently has 250 different products, and is planning to introduce another 30 per year. The manager has suggested using a three-letter code (e.g., AAB, ABC, DFG) to identify product types. Is this code large enough to handle all of the products? This problem has two important characteristics: (1) Order of the letters matters, and (2) letters can be repeated or reused. For an example of (1), ABC is different from CBA even though both use the same letters. To illustrate (2), AAB and DDD are both legitimate codes even though letters have been used more than once. This type of problem is one of the simplest to count. With three letter positions available, each one is different, and each can hold any of the 26 letters in the English alphabet. So, a three-letter code has the following number of possibilities: 26*26*26 = 17,576. This number is clearly large enough to handle well over 100 years of new products (100*30=3000). Would a two-letter code be large enough? Maybe, because 26*26=676, but subtract the original 250 products, and the remaining codes would last for a little over 14 years. In a similar question, assume a state assigns car license plates using three letters and three digits. How many unique plates can be made? The answer is 26^3 10^3 or about 17.5 million. That seems like a large number, but go back and talk to New Jersey officials in the 1990s—when the state ran out of numbers. The car problem is complicated if the state wants to avoid reissuing numbers for a couple of years.

Counting becomes more complicated as constraints are added to the problem. Consider an example of a presentation where the top officers of the company will be seated on a platform in a row. How many ways can these people CEO, CFO, HRM, CIO, and CMO be arranged? Notice that the list consists of unique items, so duplicates do not exist. Simple **arrangements** are relatively easy to count. Start with the first location. With n items, there are n possible choices for the first spot. After one of those is selected, there are n-1 possibilities for the second slot, n-2 for the third and so on down to one person for the last position.

Number of Arrangements = n!

The exclamation point means factorial, or n*(n-1)*(n-2)*…*1. Technically, 0! is defined as equal to one. So, the presentation would have 5! or 120 arrangements. But, change the problem slightly: The CEO has to sit in the middle seat. How many arrangements exist now? Many counting problems are exercises in logic and definitions. This one is straightforward to solve once you realize the problem now involves arranging only four people. So the answer is 4! or 24 possibilities. This number is much smaller than the 5!, so factorials grow rapidly. Perhaps with only 24 arrangements, the remaining executives will be able to select their preferred seats in less time.

A variation of arrangements is to add a new constraint. A **permutation** is an arrangement of items but some of them are not displayed. Consider a situation where a company has a showroom or Web site for products. A retail Web site wants to showcase its daily sales items on its main page. The company puts 20 products on sale each day and wants to build a display ad that showcases three products in a row, showing each product for 15 seconds before moving to the next. Each time a person opens the main page the display should feature a different group of three products or at least in a different order. How many permutations exist? The count of the number of permutations of n items taken k at a time is:

Number of permutations P(n,k) = n! / (n-k)!

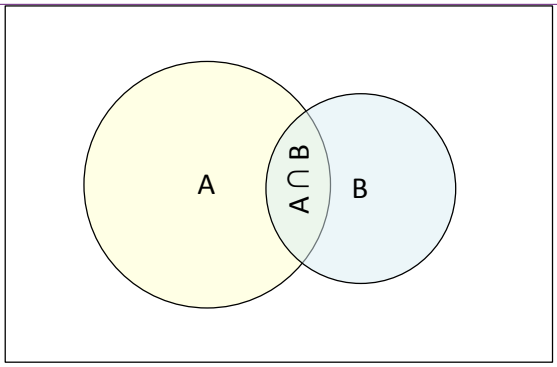


Figure 4.2

Venn diagram of two events A and B. If events A and B can overlap—both happen together, then the circles overlap and the overlap is the portion of space where both A and B occur.

This formula can be derived much the way the arrangements formula was created. Start at the first position with n items, the next has n-1 and so on. The difference is that the progression stops when you run out of spots (k). Write out the formulas for n! and (n-k)! to see that dividing creates: n (n-1) (n-2)… (n-k+1), or the numbers in sequence from n until all k spots are filled. In the example, the answer is 20! / (20-3)! = 20! / 17! = 20*19*18 = 6,840 different displays.

Counting when Order does not Matter: Combinations

Another constraint that often arises in problems is that order might not matter. Typically, in lottery games, picking the three numbers is important, but they could be chosen in any order—particularly when the numbers are unique and drawn without replacement. In business, it can be important to know which items are purchased together, but it rarely matters if one item was placed into the basket or on the counter before another one. A **combination** is a collection of items (k) out of a total set (n) where order does not matter.

Consider the shopping basket example. If a company sells 20 products, how many different shopping baskets exist that contain exactly 5 items? The order of the items is unimportant, so a basket containing items A, B, C, D, E is equivalent to one with E, D, C, B, A. The answer is to take the number of permutations (where order does matter) which is given by n! / (n-k)!, and divide by the number of possible arrangements (k!). This formula appears in several areas of mathematics (e.g., Pascal's triangle and the binomial formula) and is important enough to have its own symbol:

$$\text{Number of combinations} = C(n,k) = \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

In the example, n=20 and k=5. Excel has a COMBIN function that computes the value directly. To find the value by hand, compute the permutation then divide by k factorial to get 20*19*18*17*16 / 5*4*3*2 = 15,504 different baskets.

Other problems exist with different constraints, but most are derived from these formulas.

From Counting to Probability

The counting formulas are used for discrete data to determine both the denominator (all possible ways) and numerator for various probability computations. Consider the example in the previous section. A company has 20 products, identified as A, B, C, …, T. Assume the items are equally-likely to be purchased ($p_i = 1/20$) and the purchases are independent. What is the probability that a customer who purchases exactly 5 items buys both items A and B? The solution is to find the number of ways that A and B and three other items can be purchased and divide by the number of ways that any 5 items can be combined. The denominator is the total combination and is straightforward: C(20,5). The numerator is slightly tricky. Write it as five slots with two of them filled: A, B, __, __, __. Because A and B are fixed, the numerator asks for the number of ways to find 3 items out of 18 remaining products or C(18,3). Compute the two combination counts and divide to get: 816 / 15,504 = 0.0526. There is slightly more than a five percent chance that any two specific items will be purchased in a 5-item cart when 20 products are available. The trick is to identify exactly what needs to be counted and which rule applies based on the conditions of the problem.

Probability Rules

Specific rules have been found that make it easier to work with probabilities and probability functions. They deal with situations where multiple events can arise. The two most powerful are the addition and multiplication rules. Before defining the rule, it is important to understand how two events can interact. Figure 4.2 shows the common Venn diagram for two events. Event B was deliberately drawn with a smaller circle to indicate that the two events do not need to be equal. The point where the circles overlap represents the fact that both events could occur. If there is no overlap and the two events must occur separately, the events would be called **mutually exclusive**. Events are rarely mutually exclusive—unless a situation is defined that specifically prevents them from happening at the same time. For instance, consider two possible events: a person could attend a theater performance or go to a concert. These two might appear to be mutually exclusive, but more information is needed. Does the problem call for the two events happening at the exact same time, in remotely different locations, with no inclusion of video streaming, and require the person to attend the full performance of both events? It is these often-unstated assumptions that tend to make probability questions challenging. See (Mlodinow 1998) for some curious examples of interpreting conditions and applying probability to everyday life.

The Addition Rule: Union of Events

The first rule can be observed from the Venn diagram. The probability **addition rule** states:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

In words, the probability of A union B is equal to the sum of the probability of A plus the probability of P minus the probability of A and B both occurring. A union B is often read as the probability of A or B. In the Venn diagram, the probability of A is represented by the area of the entire A circle. Similarly, P(B) is the area of the B circle. The problem with simply adding the two probabilities is that the overlap portion is counted twice—once for A and once for B. Hence, the subtraction term in the formula removes the second in-

Last Name	FT Made	FT Missed	Total
Ariza	156	77	233
Brown	27	6	33
Bryant	657	104	761
Bynum	180	78	258
Farmar	59	37	96
Fisher	141	25	166
Gasol	443	136	579
Odom	199	124	323
Powell	44	12	56
Radmanovic	23	4	27
Vujacic	75	7	82
Walton	52	23	75
Total	2056	633	2689

Figure 4.3

Contingency table for 2009 LA Lakers free throws by players with more than 30 games. The cells contain the count of the free throws made and missed for each player. Margin totals are displayed for the rows and columns.

stance. If A and B are mutually exclusive, they cannot both occur together and there is no overlap, so P(A ∩ B) is zero and the formula reduces to P(A) + P(B). To be safe, it is best to remember the entire formula with the subtraction term.

The Multiplication Rule: Joint Events

The second rule also requires an understanding of events. It deals with the issue of two events occurring together: P(A ∩ B), the intersection of A and B which makes them **joint events**. The description of joint events is one of the more complicated aspects of probability and a critical element in much of data mining and business intelligence. This section deals with the simplest version of joint events. The following sections explore some of the complications and their implications for probabilities.

Two events are **independent** if the occurrence of one event does not affect the other. For example, with a balanced coin, the probability of tossing heads does not affect the probability of tossing heads a second time. The two tosses are independent. In business terms, the probability of a customer in New York buying a specific product is generally unrelated to the probability of a different customer in Los Angeles buying the same product. As long as the customers do not know each other, share Web comments, and so on. The simple **multiplication rule** states:

If A and B are independent, P(A ∩ B) = P(A) * P(B).

When events are independent, the joint probability is obtained by multiplying the two independent probabilities. Simple examples are easiest to see with games. Say event A is rolling a fair die and obtaining a 6. Call event B rolling the die a second time and obtaining a 6. The two events are independent—

Last Name	FT Made	FT Missed	Total
Ariza	0.058	0.029	0.087
Brown	0.010	0.002	0.012
Bryant	0.244	0.039	0.283
Bynum	0.067	0.029	0.096
Farmar	0.022	0.014	0.036
Fisher	0.052	0.009	0.062
Gasol	0.165	0.051	0.215
Odom	0.074	0.046	0.120
Powell	0.016	0.004	0.021
Radmanovic	0.009	0.001	0.010
Vujacic	0.028	0.003	0.030
Walton	0.019	0.009	0.028
Total	0.765	0.235	1.000

Figure 4.4

Contingency table written as probabilities. Each cell value was divided by the grand total.

whatever happens on the first roll has no effect on the second. A standard die has six sides and the probability of any one number appearing is 1/6. The joint probability of obtaining a six on the first roll and a six on the second roll is (1/6)(1/6) = 1/36.

As a business example, two machines (A and B) produce parts and occasionally create defects. P(defect from A) is 1/200. P(defect from B) is 1/300. Assume the machines and the defect rates are independent; for example, they are not both run by the same person. The probability of finding a defect from both A and B is 1/200 * 1/300 = 1/60,000.

A Scary Example

Try a scary example. A complex machine, perhaps a jet or a space shuttle, has 50,000 components. Each part has a 1/10,000 chance of failing so the probability of success is 1-1/10,000 or 0.9999. In business shorthand, this success level is sometimes referred to as *four nines*, and it is a relatively high success rate. Assume the chance of failure in each component is independent but if one part fails the entire system fails. The probability of system success is the joint probability that each part is successful: P(S1 ∩ S2 ∩ … ∩ S50000). From independence, these probabilities can be multiplied: P(S1)*P(S2)*…P(S50000). Each individual probability is high, but the joint probability of success is 0.9999^{50000} or 0.0067. Less than a one percent chance the system will not fail!

How can a complex machine possibly succeed? The answer lies with the same analysis: redundancy. Build the system so that each component has a backup that is independent of the original. For convenience, assume components and backups have the same failure rates (1/10,000). A component fails only if both the original

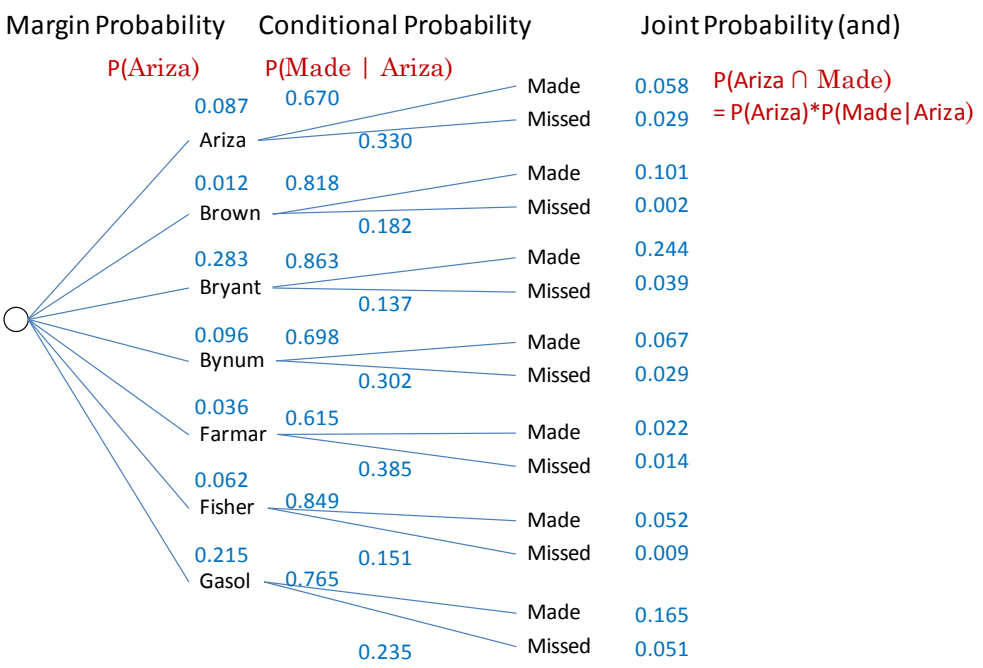
and its backup fail. So P(failure1) = P(F1a and F1b) = P(F1a)P(F1b) = 1/10,000 ^ 2 = 10^{-8}. So the probability of success for the entire system is P(Success) = (1 - 10^{-8})^50,000 = 0.9995. This success probability is much higher than the original—but it carries a huge cost because everything is doubled: size, cost, weight, and so on.

Interdependencies: Joint Probabilities

How are probabilities handled when events are not independent? Independent events are important, but business problems usually involve events that are related. Most business intelligence applications are designed to find and evaluate these interdependencies. Several aspects of probability were specifically designed to deal with joint probabilities and lay the foundation for evaluating relationships among attributes and events. This section reviews the basic concepts and defines the important terms. A detailed discussion and proofs can be found in most introductory probability books. This section relies on the use of discrete data and relatively small problems. All of the concepts and results can be applied to continuous data and more complex problems.

Figure 4.5

Partial tree diagram of the Lakers' free throws. The nodes represent values of the attributes. Here, players and whether free throws were made or missed. The numbers are essentially conditional probabilities, except for the values on the leaf nodes which are the joint probabilities. Reading from left to right, Ariza took 8.7 percent of the free throws. When he was shooting, he made 67 percent for a joint probability of 0.087*0.670 = 0.058.



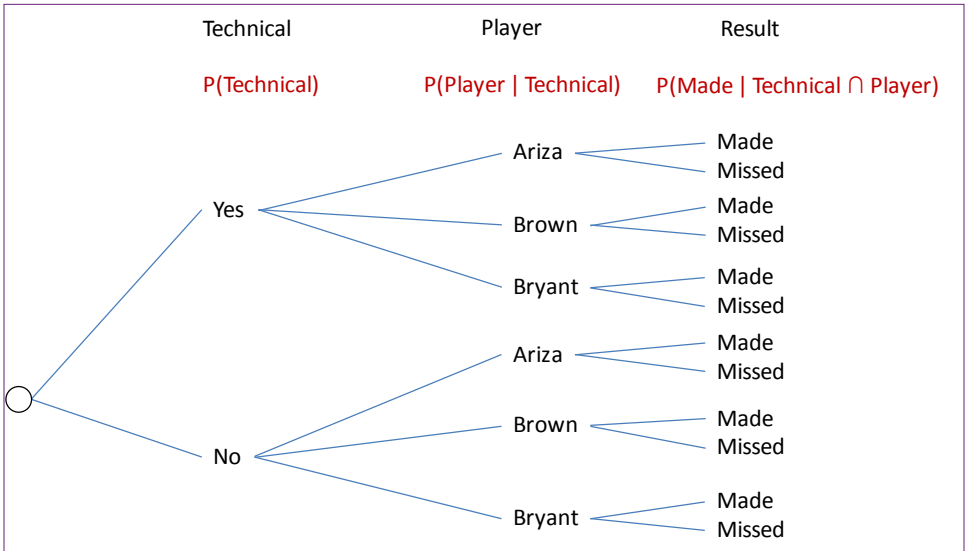


Figure 4.6

Adding an event to the tree. Trees can be organized with events in any order—whatever makes the most sense for the problem and managers.

Contingency Tables

The easiest way to understand the concepts and terms involved with interdependencies is to examine the smallest problems that involve only two attributes or events with discrete data. This type of problem is easily displayed in a **contingency table** that places one attribute as rows and the second as columns. Typically the tables contain counts of the number of observations that fall into each cell, but sometimes the tables will display the percentages.

Figure 4.3 provides a sample contingency table using the basketball database. To keep the table relatively small, the data shows the players from the 2009 Los Angeles Lakers who played more than 30 games that season. (The 2009 Lakers were the world champions.) The columns list the total free throws made and missed by each player for the entire season—including the playoff games. If you are not interested in sports, consider the problem an issue in management. A similar table could be made for any company that keeps statistics on successes for its employees, such as jobs completed on time or number of closed sales.

Each cell in the contingency table represents a **joint probability** where the specific row and column event occur. When the cells contain counts, the probability is found by dividing by the grand total. In the example, the joint probability that Player=Ariza and a free throw is made is computed as 156/2689 = 0.058. Be careful, this value is not the same as Ariza's free throw percentage or success rate. It is the probability that a player on the floor taking free throws is Ariza and that the free throw is made. The figure also shows the **margin totals** for the rows and columns. The margin totals are used to compute the percentages or probabilities within the row or column and to show how the specific row or column relates to the overall total. Figure 4.4 shows the same contingency table with the cell values and margin totals computed as probabilities by dividing by the grand total.

Notice that the row totals show the percentage of times that each player takes a free throw, not the percentage made. The column totals show the overall team's success rate. The huge value of 76.5 probably helps explain why the team won the championship—and that is a question that could be examined with later data mining tools.

A common value that is useful in making decisions is the probability of making the free throw for a specific person. For example, the coach (manager) wants to know the probability that a free throw will be made if the player is Bryant. In probability terms, this value is written P(FT Made | Player=Bryant). This expression is read: The probability the free throw will be made *given* the player is Bryant. It is known as a **conditional probability**, because it is conditioned on the information provided (Player=Bryant). In general terms, the conditional probability is found by:

$$P(A \mid B) = P(A \cap B) / P(B)$$

In the example, the probability the free throw is made given Bryant is at the line is found as P(FT Made | Bryant)/P(Bryant) = 657 / 761 = 0.863. These values are taken from the original table of counts which is closest to the way most people read the sports table. But the same result can be found using the Figure 4.4 probability table: 0.244/0.283 = 0.863. The definition of conditional probability is important and you need to remember it. The trick is to remember that you divide by the probability of the item in the condition (B).

The same formula can be rewritten as the **general multiplication rule**:

$$P(A \cap B) = P(A \mid B) * P(B)$$

This rule can always be used to find the joint probability of two events. If events A and B are independent, P(A | B) simply reduces to P(A). Independence means that the probability of A stays the same, regardless of whether event B occurred.

Tree Diagrams

Contingency tables are useful because they show all of the relevant details, but they get unwieldy when the problem gets large—with too many rows and columns. Also, they can show the values of only two attributes. Tree diagrams are another way to display relationships in dependent events. They take up space on the page, but tree viewer software typically enables analysts to zoom in and out to examine the nodes.

Figure 4.5 shows the tree for the Lakers 2009 free throw data for the first seven players. The nodes represent the values of the attributes (Player and Result). The probabilities on the lines or on the specific node are the conditional probabilities. The tree is read from left to right, so any probability to the right is conditional on everything that came before. In the example, P(Ariza)=0.087 from the starting point, which corresponds to the percentage of times that Ariza was shooting the free throw. For the next step, P(Made | Ariza) = 0.670; or P(Missed | Ariza) = 0.330. The final probability is the joint probability, and it is found by multiplying every probability gathered along the way from the node: P(Ariza ∩ Made) = P(Ariza) * P(Made | Ariza) = 0.087 * 0.670 = 0.058. The joint probabilities match those in the contingency table cells.

The tree diagram can be extended to handle more events. Each event or attribute would create a new level, where every value of the attribute gets repeated for each existing node. In the example, additional data might be found to indicate

	FT Percent	Frequency Fouled	
LastName	P(FT Made | Player)	P(Player)	multiply
Ariza	0.670	0.087	0.058
Brown	0.818	0.012	0.010
Bryant	0.863	0.283	0.244
Bynum	0.698	0.096	0.067
Farmar	0.615	0.036	0.022
Fisher	0.849	0.062	0.052
Gasol	0.765	0.215	0.165
Odom	0.616	0.120	0.074
Powell	0.786	0.021	0.016
Radmanovic	0.852	0.010	0.009
Vujacic	0.915	0.030	0.028
Walton	0.693	0.028	0.019
			0.765

Figure 4.7

Basketball data in more common format, providing limited data. The denominator in Bayes' theorem requires multiplying the two columns and computing the sum.

whether the free throw was due to a technical foul or not. This addition would double the number of end nodes. The conditional probabilities are added to the new connectors or nodes and the final joint probability is the multiplication of every value along the tree to that specific node. Figure 4.6 shows one way to extend the tree in the basketball example. Events can be organized in any order to make the tree readable by managers. In most cases, if a technical foul is called, the coach can choose the player to shoot the free throw, so it makes sense to perform the split on technical first. Learn to read decision trees; they are heavily used in Microsoft's Analysis Services.

Bayes Theorem

Thomas Bayes, a British mathematician and Presbyterian minister in the early 1700s, derived an interesting equation using the probability definitions. The proof is easy, and even the theorem is straightforward. The ultimate implications have a profound impact on statistics and on data mining.

For the simple explanation, begin with the definition of joint probability:

$$P(A \cap B) = P(A|B)P(B)$$

But, A and B are any two events, so the two terms easily could be reversed:

$$P(B \cap A) = P(B|A)P(A)$$

Put the two definitions together by setting the two right-hand sides equal:

$$P(A|B)P(B) = P(B|A)P(A)$$

Rearrange slightly to get Bayes' Theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

It looks straightforward. The key lies in understanding how it can be used.

Simple Example of Bayes

In the simplest cases, Bayes' theorem is not needed. Consider the basketball example and look at Figure 4.4 again. Initially, the contingency table was used to find the conditional probability P(FT Made | Bryant), or the probability that a free throw would be made if it is taken by Kobe Bryant. Consider a slightly different version of the question. You are watching a Lakers basketball game on TV and you notice that someone makes a free throw, but you cannot identify the person before the coverage cuts to a commercial. What is the probability the player is Bryant? Or, in a management context, you encounter a new client who praises a salesperson but does not remember the name. With the same data defined as completed sales, what is the probability the salesperson was Bryant?

If the complete contingency table or decision tree is available, it is possible to find this new conditional probability directly. In the table, find P(Bryant ∩ FT Made) = 0.244. Find the margin probability P(FT Made) = 0.765 and divide to get: P(Bryant | FT Made) = 0.244 / 0.765 = 0.320.

Even using Bayes' formula, if all of the data are available, the computation is straightforward. The formula requires P(FT Made | Bryant), P(Bryant), and the total P(FT Made). By the formula, the computation is: 0.863 * 0.283 / 0.765 = 0.320, which should be the same as the direct calculation, except for some rounding errors

Common Application of Bayes

The real strength of Bayes' theorem is that many problems do not provide complete data. Some data might be too expensive to obtain, or timing problems might make it difficult to obtain complete data. Figure 4.7 shows a more common version of the basketball data. The first column contains the free-throw percentage made by each player. In probability notation, it is P(FT Made | player). The second column is often harder to obtain, but rough values tend to exist. It shows the relative frequency of each player getting fouled—which means taking free throws. In probability terms it is P(player). With this limited data, Bayes' theorem becomes useful. Still considering the Bryant question, the numerator is simply P(FT Made | Bryant) * P(Bryant) which is found by multiplying the two columns. The denominator P(FT Made) or total probability seems to be harder to find. However, it can be rewritten as

$$P(FTMade) = \sum_i P(FTMade \mid player = i)P(player = i)$$

That is, the denominator is the sum of the probabilities for all of the players. As shown in the table, simply compute the multiple for each player row and add them up. To obtain the probability Bryant made that last free throw, take the value for Bryant and divide by the total to get: 0.244 / 0.765= 0.320. This table is easy to extend to find the probability that any of the other players made the shot, but the highest probability lies with it being Bryant.

To obtain 10 heads in 10 tosses: H H H H H H H H H H
Only one way to get it and P(H)=1/2, so P(10H)= (1/2)^10

To obtain 9 heads: H H H H H H H H H T
P(H) = 1/2 and P(T) = 1/2, so start with (1/2)^9 (1/2)
But the tail could appear in one of 10 locations, so multiply by C(10, 1)

For 8 heads: H H H H H H H H T T
Start with (1/2)^8 (1/2)^2 and multiply by C(10,2), the combination of ways of to get the two tails.

For 7 heads: (1/2)^7 (1/2)^3 C(10,3)

Total these four values to get P(7 or more heads | fair) = 0.171875

Figure 4.8

Counting heads. Computing the probability of obtaining 7 or more heads in 10 tosses of a coin is a good exercise in counting. The tosses are independent so the probabilities can be multiplied. The trick is to understand that the tail can appear in different tosses, so the values need to include the combination term.

This summation approach is a common aspect of problems that require Bayes' theorem. The point of the theorem is that data is often provided in pieces, but the formula makes it possible to determine other information based on those pieces. To look at the approach one more time, examine the decision tree in Figure 4.5 again. It clearly shows the probabilities and conditional probabilities. Multiply each line from the root to the ending leaf node. The resulting value is the joint probability. Pick the value for the chosen player (e.g., Bryant) to use as the numerator. Complete the multiplications for every node labeled FT Made, and add up the values. The resulting sum is the overall total used for the denominator.

Information Value of Bayes' Theorem

Bayes' Theorem is often used in data mining applications. The usage derives from using the theorem to look at probabilities in a different light. A slight rearrangement of the theorem yields:

$$P(A \mid B) = P(A)\frac{P(B \mid A)}{P(B)}$$

This rearrangement leads to a completely different interpretation of probability: Probability is a subjective belief that can change over time as new information is gained. In the formula, P(A) is the a priori belief in some event A. P(A|B) is the updated posterior probability—the new belief about A given that information B has been provided. The posterior belief is computed from the prior belief by multiplying by the normalized information term P(B|A) / P(B). Typically, decision makers begin with a simple naïve belief about the probability—as neutral as possible, and the addition of information is used to successively create better estimates.

In general, business managers do not need to worry about the mechanics of how the prior belief is updated to become the new probability. However, it is worth examining a simple problem just to gain some understanding of the process. A friend shows you a coin and claims that it is fair (50 percent chance of heads and tails). You mostly trust your friend, but have some skepticism (say 10 percent). You flip the coin 10 times and 7 heads appear. How should you adjust your belief that the coin is fair?

By Bayesian updating, the goal is to find:

P(Fair | 7 or more heads) = P(Fair) * P(7+heads | Fair) / P(7+heads), where

P(7+heads) = P(7+heads|Fair)*P(Fair) + P(7+heads|not Fair)*P(not Fair)

Begin with the a priori value of P(Fair) = 0.9. Most of the other values can be computed, but there is no absolute way of knowing how biased the coin might be. It appears not to be completely biased (always head), so just pick a value and guess that P(7 or more heads | not Fair) is 90 percent. The other value to calculate is the probability of obtaining 7 or more heads with a fair coin. Why "7 or more" instead of just 7? Because the coin could be unfair if it shows the 7 heads you did receive or any amount more. Computing this probability is a good exercise in counting.

Start with the probability of obtaining 10 heads (in 10 tosses) and work down to 7. Figure 4.8 shows the basic process and why it is a good exercise in counting. The case of 10 heads is easy because there is only one way to obtain all heads. Each toss is independent, so the probabilities can be multiplied. With a fair coin, the probability of 10 heads is $(1/2)^{10}$, or 0.000977, which is a small number. The case of 9 heads and 1 tail shows the counting issue. The initial probability part is straightforward. With 9 heads and 1 tail, start with $(1/2)^9 * (1/2)^1$ to get the base, but the tail can appear in any one of the 10 tosses, so this base must be multiplied by 10 which is the number of ways of arranging the one tail in the 10 trials. Essentially, the additional rule for mutually exclusive events is applied ten times to the same number.

Follow the same process for 8 and 7 heads. For 7 heads the probability is $(1/2)^7 * (1/2)^3 * C(10,3) = 0.1171875$. Technically, the arrangements of tails and heads are not combinations. With only two possibilities for the ten trials (H or T), the 10 values consist of 7 items the same (H) and 3 items the same (T). The number of ways to arrange 10 items is 10!, but because 7 items are the same, this number needs to be divided by 7!. Because the other 3 (10-7) are the same, this value also needs to be divided by 3!, which results in: 10!/(7!3!) which is equivalent to C(10,3).

Add up the values to obtain the probability of getting 7 or more heads of 0.171875. Plug this value into Bayes' formula:

P(Fair | 7+heads) = 0.9 * 0.171875 / (0.171875*0.9 + 0.9*0.1) = 0.632184

Remember that you started with a 90 percent belief that the coin was fair. After obtaining so many heads, your adjusted belief that the coin is fair has dropped to 63 percent. Information obtained from testing has improved the probability belief. This same approach is used for business problems. The example could be restated as a management question—such as whether you believe a machine is operating correctly, an employee is performing acceptably, a customer will purchase an item, and so on. Business problems quickly become more complicated, but the underlying adjustment process is the same.

Probability Distributions

How can probability concepts be generalized to make it easier to handle new situations? The fundamental aspects of probability are useful, but it would be time consuming to go back to the basic rules for every problem encountered. It turns out that many problems fall into certain categories, defined by specific assumptions and characteristics. Statisticians have developed tools to handle these specific cases. Once a new problem is categorized, it is straightforward to apply the tools and understand most of the details about the new problem. Probability distributions are one of the key statistical tools. Several standard distributions are used in many applications. In the early days, tables of these distributions were published and used to solve problems. Today, computerized functions quickly provide values for many probability functions. For convenience, the standard statistical functions are programmed into spreadsheets such as Excel. Also, distribution functions are the only way to examine probabilities with continuous data. However, discrete cases are simpler so they are examined first.

Discrete Data

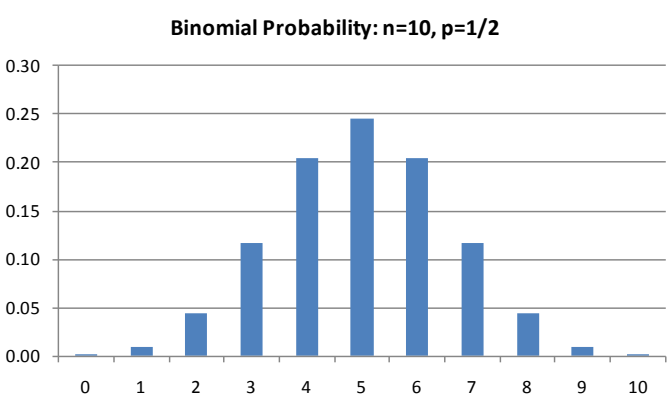
Recall that discrete data consists of values that can be counted. Simple examples include binary cases with yes/no or success/fail as the only options. Broader cases include categorical data such as gender, product category, country or state, and so on. Discrete cases also include classic games problems such as the number of heads appearing in 1,000 tosses of a coin. Discrete data can contain many values, even an infinite number, but the values are all distinct (and countable).

Return to the example from the prior section in Figure 4.8 involving the number of heads appearing in 10 flips of a coin. This type of problem is a classic example of one of the common probability distributions. Probability is defined in terms of an **experiment**, which consists of a set of events or trials and a sample space. A sample space is a set of all possible outcomes from an experiment. Ex-

Figure 4.9

Binomial distribution example. P=1/2, n=10.

Number of Heads	Probability
0	0.00098
1	0.00977
2	0.04395
3	0.11719
4	0.20508
5	0.24609
6	0.20508
7	0.11719
8	0.04395
9	0.00977
10	0.00098



periments might be clinical—where the investigator has complete control over the experiment. Or, particularly in social and business environments, experiments might consist of solely of observations on real-world activities. For instance, over a two-month interval a company collects data on customers and sales totals. The investigator might have no control over the experiment, or perhaps variables such as price or advertising were altered to see what changes occurred. In any case, it is possible to define all of the possible outcomes of the experiment.

A **random variable** is an assignment of a number to every possible outcome in the sample space. For example, if the experiment is to flip a coin ten times, a random variable could be the total number of heads that appear. In other experiments, a random variable could be defined as the weight of a group of workers, or the total sales made to a customer. Note that before the experiment is conducted or evaluated, there is no way of knowing the exact value assigned to the random variable—hence the name "random." Discrete random variables take on a limited (countable) number of values. A discrete random variable X can be assigned any value x_i from a set of values. Each possible outcome can be assigned a number called a probability $p(x_i) = P(X=x_i)$ that must meet the basic conditions:

$$p(x_i) \geq 0 \qquad \text{for all i items in the set}$$

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

The function p is the **probability function** and the pairs $(x_i, p(x_i))$ define the **probability distribution** of X. More recently, the term **probability mass function** has become popular instead of just probability function.

Binomial Distribution

The concept of distributions is easiest to see with an example. Consider the coin tossing example. A fair coin has a 1/2 probability of displaying a head on any individual toss. Throw the coin 10 times and count the number of heads in that experiment. The outcomes range from 0 to 10. Figure 4.9 shows the probability distribution along with a chart of the probabilities. The probabilities at each value are computed using the basic probability rules. The example of 7 heads and 3 tails was outlined in the prior section. Each trial flip is independent of the others, so flipping a total of 7 heads means the probability of one head AND another head and so on up to seven. Because the events are independent, $P(H1 \cap H2)$ is P(H1)*P(H2) for seven times. Generically, the probability of obtaining one success (head) is defined as *p*, which is 1/2 in this example, but will be different for other problems. In general, the probability of obtaining k successes $P(X=k) = p^k$. But, with n trials, k successes appear only if k-n failures appear, so the formula needs the probability of obtaining exactly k success and n-k failures: $p^k\ p^{n-k}$. The other catch is that it does not matter when the successes or heads appear. The experiment is only interested in the totals. The formula to this point defines one way of obtaining the specified number of successes. The base probability must be multiplied by the number of ways of arranging the outcomes. Because only two different outcomes exist, the ways of arranging them are n! / n! (n-k)!, where

k	p(k)	cdf(k)
0	0.006738	0.006738
1	0.033690	0.040428
2	0.084224	0.124652
3	0.140374	0.265026
4	0.175467	0.440493
5	0.175467	0.615961
6	0.146223	0.762183
7	0.104445	0.866628
8	0.065278	0.931906
9	0.036266	0.968172
10	0.018133	0.986305

Figure 4.10

Poisson distribution example. Mean = 5. The probability of seeing 10 or more arrivals in the time interval is 1-0.968 or about 3 percent.

the divisors handle the duplicates of the two outcomes. The result is the binomial distribution:

$$binomial(k, p, n) : P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

k: Number of observed successes.
p: Probability of observing success on a single trial.
n: Total number of trials.

Excel: BinomDist(k, n, p, false)

The binomial distribution is used for any experiment that (1) has a binary result (success or failure), (2) a repeated number of trials, and (3) a fixed probability of success on each trial. The third requirement applies to experiments where the items are draw with replacement. For problems where items are drawn from a finite set and not replaced, the probability changes with each draw and it is necessary to use the hypergeometric distribution. However, it is rarely used in data mining so it is not covered here.

What if a problem has more than two outcomes? The multinomial distribution handles cases with multiple outcomes. This probability function generalizes the binomial, where the outcomes are considered as categories (x_1, … x_k):

$$P(X_1 = x_1 \; and \ldots X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$$

k	Number of outcomes or categories
x_1 … x_k	Number of observed successes in each category
n	Total number of trials

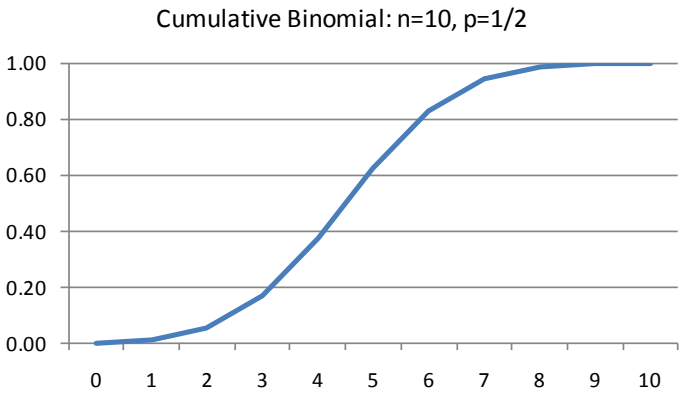


Figure 4.11

Cumulative distribution function for binomial. The s-curve is common and it must run from zero to one. $F(x) = P(X \leq x)$.

Poisson Distribution

The **Poisson distribution** is another useful discrete distribution. It is commonly used in business and operations management problems to evaluate the number of arrivals during a fixed time period. For example, retail stores and fast food restaurants need to know how many customers might arrive during a lunch hour so that a sufficient number of checkout clerks can be hired to handle the peak load. The probability function is given by:

$$P(X = k) = \frac{e^{-a} a^k}{k!}$$

In the function, α is a parameter that must be provided. It is the average number of arrivals expected during that time period. This value is generally estimated by counting the number of customer arrivals over time; or by pulling the data from the database using the time stamps on receipts. To illustrate, consider a restaurant that averages 5 customers per minute during the lunch hour time slot. What is the probability that the company is swamped with 10 or more customers during one minute? Figure 4.10 shows the probability and cdf values for the number of arrivals ranging from 0 through 10. There is about a 2 percent chance that exactly 10 customers will arrive within any minute. The cdf can be used to compute the probability of 10 or more customers because that value is 1 minus the probability of 9 or fewer customers. Hence, the answer is 1-0.968, or about 3.2 percent chance of having that many customers arrive in one minute.

Cumulative Distributions

An important aspect of probability distributions is that they are point values that return the probability of only that one discrete value. For instance, Binomial(7,1/2, 10) returns the probability of seeing exactly 7 heads in 10 coin flips. However, many problems are written so that they require cumulative values. Such as what is the probability of seeing 3 or fewer heads? The problem with point values is that

p	0.5	0.5	0.5
n success	7	70	700
trials	10	100	1000
Binomial	0.1171875	2.32E-05	5.07E-38

Figure 4.12

Binomial probability with an increasing number of trials. Observe what happens to the point probability as the number of trials increases. With continuous data, the point probability at any point is zero.

in many cases, the probability is small for seeing exactly one specific result. This point is even more critical in the next section on continuous variables.

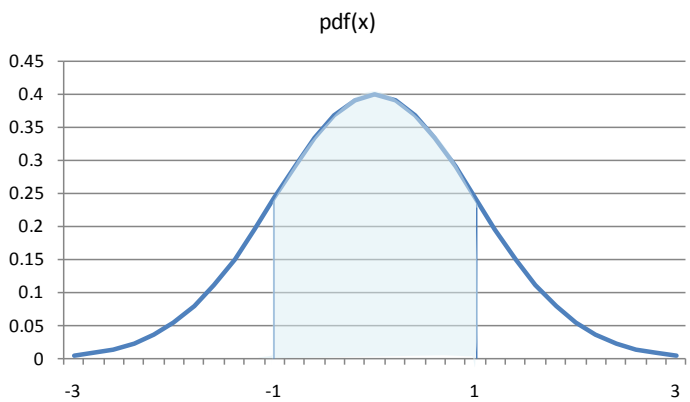
The **cumulative distribution function (cdf)** F(x) is defined as $P(X \leq x)$. With discrete random variables, it is the sum of all of the values from zero up to x.

$$F(x) = \sum_{u:x_j \leq x} p(x_j)$$

Figure 4.11 shows the cdf for the simple binomial example. Changing the parameters will alter the cdf, but the basic s-curve shape is common for most distributions. The values must range from zero to one and the function is non-decreasing. Cumulative probability tables are often used to lookup specific values. Computer functions are typically available to compute cumulative distributions. For instance, the Excel function BinomDist uses the last parameter to specify that the function should return a cumulative value (true) or just the point value (false).

Figure 4.13

Probability density function for continuous data. Probability is computed as the integral of the pdf over the specified range—which is the area under the pdf curve. It is also the difference of the cdf values: F(1) – F(-1) = 0.841-0.159 = 0.683



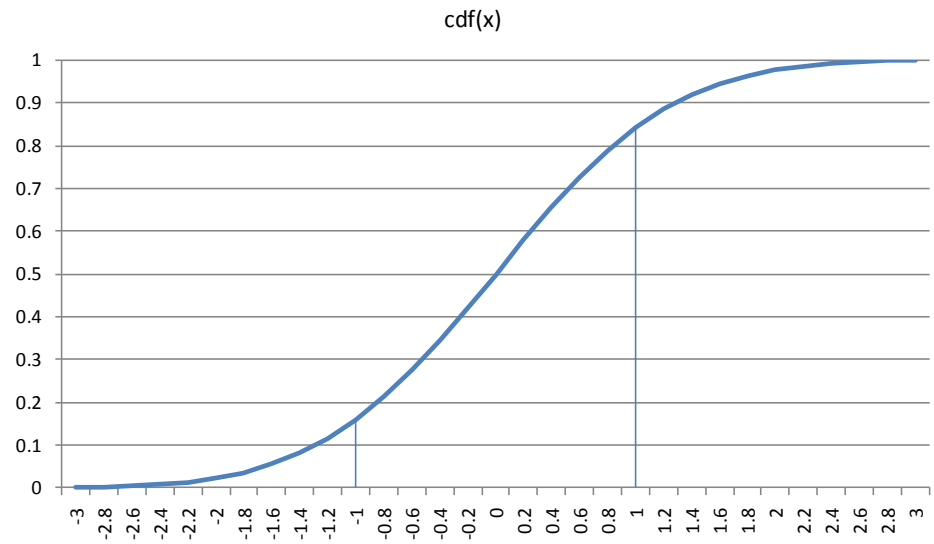


Figure 4.14

Cumulative distribution function. This function makes it easier to compute probabilities across a range.

Some problems ask for the probability of X falling between two values P(a ≤ X ≤ b). This value can be computed by finding the cumulative probability up to value b and subtracting the cumulative probability up to value a. Be careful with discrete distributions. In the sample binomial distribution, what is the probability that the number of heads falls between 3 and 7 (inclusive)? Because N is only 10, this value could be computed by using the simple probability distribution function (non-cumulative) and adding the results. If N is large, this total can be tedious, even with a spreadsheet. With the cumulative approach, the answer is found using: P(X ≤ 7) – P(X ≤ 2) or 0.9453-0.0547 = 0.891.

Continuous Data

With these basic concepts, it is now time to examine distributions for continuous data. The first, most important, concept to understanding continuous data is that the point probabilities are all equal to zero. Recall the binomial example for 10 trials. The probability of obtaining exactly seven heads was almost 12 percent. Check Figure 4.12 to see what happens to that point probability as the number of trials increases. With continuous data, the probability of any specific point is always zero. Think about it in terms of the number of possible values. It has to be impossible to hit exactly one specific number out of infinity.

With continuous data, probability is defined only in terms of the cumulative distribution function. Probability can be found only for ranges of values. Technically, a **probability density function (pdf)** or probability mass function exists, but it is defined only in terms of the cumulative values:

$$f(x) \geq 0 \text{ for all } x$$

$$F(x) = \int_{-\infty}^{+\infty} f(x)\,dx = 1$$

Integration can be interpreted as the area under the pdf curve. It is also the value of the cumulative distribution function. Figure 4.13 shows a sample pdf for a continuous distribution. The shaded area shows the method of computing the probability P(-1 ≤ X ≤ 1). Given the specific form of the pdf, this probability could be computed directly. However, the pdf functions tend to be difficult to integrate, so the values are found from cumulative tables or using computer numerical formulas. In the example, the probability can be found using the cdf: F(1) – F(-1) = 0.683. The probability of X falling between -1 and 1 is about 68 percent.

To obtain those specific values, you need to know that the pdf plotted is for the standard normal distribution. Note that because the probability at any point is zero, you do not need to worry about the end points and inequalities, which are a concern with discrete distributions. Figure 4.14 shows the cumulative distribution function for the example. The approximate cumulative probabilities can be read from the chart, but more precise values can be found in tables or with computer functions.

Joint and Conditional Probabilities

Note: This section shows theory and could be skipped.

Most probability density functions are discussed and displayed in terms of a single variable. These functions are easy to plot and to tabulate values for both the pdf and cdf. However, most business problems involve multiple attributes or variables. Fortunately, the computational tools are built to handle multiple variables. Even better, the behavior of the functions is the same for multiple variables. So, once you understand the functions in terms of single variables, that knowledge can be applied to multiple variables. However, multiple variables add a few complications.

First, as shown with contingency tables and tree diagrams, multiple variables require a joint density function. In discrete terms, the joint probability function could be expressed as P(X=x and Y=y) = p(x,y). Continuous variables require defining probability from a joint density function that is integrated to obtain the cumulative distribution:

$$f(x, y) \geq 0 \qquad \text{for all } x, y.$$

$$\iint f(x,y)dxdy = 1$$

For more variables, another integral (summation) is added for each variable. Recall that the contingency table included margin probabilities that were computed as the sum of probabilities for a given value of x or y (row and column totals). The same principle leads to marginal density functions for continuous variables, replacing summation with integrals:

$$marginal\ pdf\ for\ x: g(x) = \int f(x,y)\,dy$$

$$marginal\ pdf\ for\ y: h(y) = \int f(x,y)\,dx$$

Consequently, conditional probabilities are defined similarly to the probability case:

$$conditional\ pdf\ g(x|y) = \frac{f(x,y)}{h(y)},\quad for\ h(y) > 0$$

Net Return	Probability	X*p
1000	0.05	50
700	0.7	490
-2000	0.25	-500
	Total/EV	40

Figure 4.15

Expected Value computation. Three outcomes exist for an investment. The net return and probability are given for each outcome. The expected value is found by multiplying each value by its associated value and adding. The expected value here is positive so the investment is worthwhile.

With these definitions, Bayes' Theorem also applies to continuous data. With continuous random variables, Bayes' Theorem leads to a method to update the entire density function not just a simple probability. Methods that use this approach begin with a neutral pdf (often a uniform distribution), evaluate the data and adjust the density function. This new pdf is plugged in as the new a priori function and the process repeats until the data is used up. The result is a pdf that incorporates all of the information from the dataset and can be used to compute the probabilities of any range of values.

Expected Value (Mean) and Variance

Probability distributions need to be general so they can be applied to many different types of problems. Most of the distributions are defined in terms of **parameters** that fit the distribution to specific problems. For example, in a non-statistical context, a line can be defined as y = mx + b. The variables x and y represent the data, leaving m and b as parameters. The values for m and b are estimated from the data and represent the slope and intercept values. Probability functions have similar parameters that enable them to be fit to a specific problem. Two of the most common parameters are the mean and variance.

To understand probability functions and parameters, it is important to understand the concept of expected value. Given a random variable X, the **expected value** of X is defined to be

Expected Value

$$E[X] = \sum_{i=1}^{\infty} x_i p(x_i) \qquad \text{X is discrete}$$

$$E[X] = \int_{-\infty}^{+\infty} x f(x)\, dx \qquad \text{X is continuous}$$

Expected value is easiest to see with discrete values because most people find it easier to compute sums than integrals. Expected value is an extremely useful tool for solving basic business problems. Figure 4.15 shows a simple example of an investment that has three possible outcomes. The random variable is the net return, shown for each possible outcome. Probabilities exist for each of the outcomes—they could be subjective probabilities. Compute the expected value by multiplying each outcome value by its associated probability and summing the values.

Outcome	Net Return	probability	X*p	X-E(X)	p(X-E[x])^2
1	1000	0.05	50	1000	50000
2	700	0.70	490	700	343000
3	-2000	0.25	-500	-2000	1000000
			40		1,393,000

Outcome	Net Return	probability	X*p	X-E(X)	p(X-E[x])^2
1	150	0.2	30	100	4500
2	700	0.3	210	700	147000
3	-400	0.5	-200	-400	80000
			40		231,500

Figure 4.16

Variance calculation. A second investment option has the same expected value but the variance is considerably lower.

The result is a 40 gain, so this investment is acceptable. Note that 40 is not the amount you would make or lose from the investment. You will actually obtain one of the three values shown in the net return (1000, 700, or -2000). In fact, with a 70 percent probability on a single investment, the most likely outcome is that you would receive 700. However, the expected value represents the average amount you would receive if the experiment/investment were repeated a huge number of times. Most of the time (70 percent), you would make 700, but 25 percent of the time you would lose 2000. In the end, you would make 40 on average.

Consider one more example using a binomial distribution. The probability of success, say heads on a coin toss, is p. If success is measured as 1 and failure (tails) as zero, the expected value of one toss is 1/2. If the coin is flipped 10 times, the total expected number of heads is: 1/2 + 1/2 +1/2 +…+ 1/2 for ten times which is 10*(1/2) or 5. In general, the expected value of the binomial distribution is n*p. If you are curious, the expected value of the Poisson distribution can be computed to be alpha (α), its single parameter.

The expected value is also known as the mean of the distribution. It is the center point. In physics terms, it is the center of mass of the distribution.

Variance

Expected value is a powerful tool and relatively easy to understand. Unfortunately, too many people focus on just the mean when making decisions. After all, the mean indicates the expected outcome if the experiment is performed many times. But, there might be many ways of getting to that mean. Consider the investment example again and assume it is an investment you can make repeatedly. For the first few times, you are relatively lucky—after all, 75 percent of the time the investment returns positive values. Then, bang, you lose 2000. The investment is characterized by relatively large gains and losses.

Consider a second investment option that has the same expected value. Figure 4.16 shows a second investment option that has been configured to have the same expected value (40) as the first option. Compare the probabilities of the two

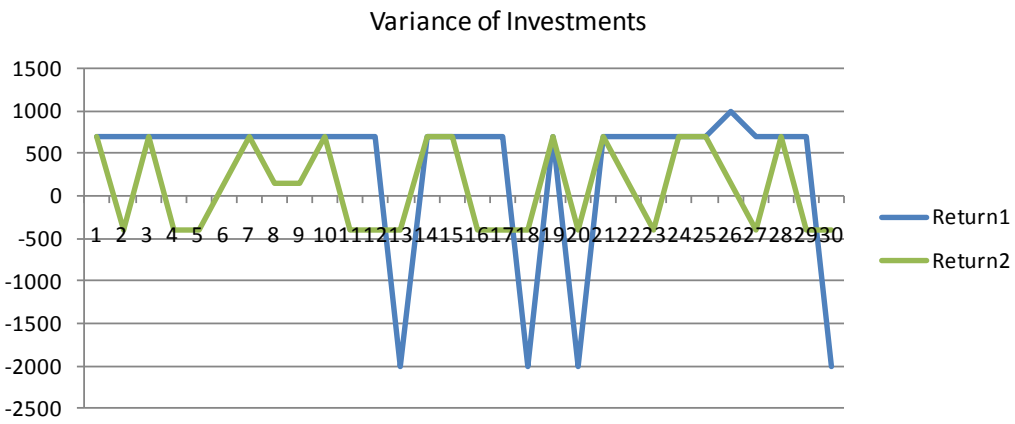


Figure 4.17

Variance effect. The first investment returns bounce wildly around the mean. The second investment is more consistent, even though both have the same mean.

investments. They are roughly the same in the each of the three outcomes for the second investment. They are widely different for the initial investment. Look at the net returns for the three outcomes. Again, in the second case the net returns are closer together in all three outcomes.

This concept is known as **variance** and it has the statistical definition:

$$V(X) = E[X - E(X)]^2$$

That is, the variance is the expected value of X minus the mean, squared. Figure 4.16 shows the direct way to calculate the variance. Subtract the mean from each value, square that difference and multiply by the probability to get the expected value. Add up the values to get the variance. The base formula is useful for understanding that variance measures the deviation from the mean, but an easier calculation method can be found by modifying the formula to:

$$V(X) = E(X^2) - [E(X)]^2$$

With this version, the mean E(X) and the term $E(X^2)$ are computed in one pass through the distribution. Then square the mean and subtract it from the first term.

The point is that when X values have large deviation from the mean or center, the distribution has a higher variance. Figure 4.17 shows the effect using the two investment options. The first investment has large swings or deviations about the mean. The second investment is more consistent. Both distributions have the same mean. Looking only at the mean, it might be tempting to conclude the two investments are equally valuable. Incorporating the variance provides more information. Most investors would choose investment two that has less variance.

Look again at the original definition for variance. It uses (X – mean) squared. The variance is measured in squared units of the random variable. For instance, if the data is dollars, the mean is measured in dollars, but the variance is measured in squared-dollars. Consequently, a more useful definition is to define **standard deviation** as the square root of the variance. The standard deviation is in the same units as the original data, yet still represents variation from the mean.

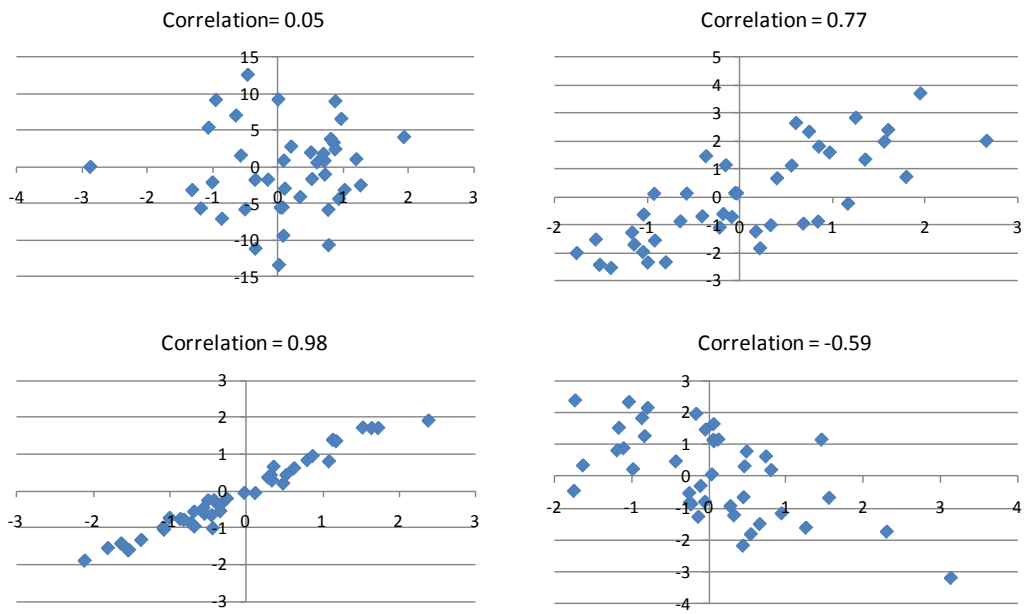


Figure 4.18

Sample correlation coefficients. At zero, the scatter plot is essentially random. As the correlation approaches one, the relationship is closer to a straight line. Negative values exhibit the same relationship, but the slope of the line is negative.

The mean and standard deviation appear in many distributions and many statistical situations. For the distributions, the mean is usually denoted as μ and the standard deviation as σ. Variance is the square of the standard deviation or σ^2.

Correlation Coefficient

When a distribution has more than one random variable (dimension), then the possibility exists for correlation between two variables. These correlations are critically important to data mining. The business analyst wants to find correlation patterns to see which variables move together. If two variables are independent, the correlation will be zero. If positive changes in one variable (say X) are matched by positive changes in a second variable (Y), the two are said to be positively correlated. With negative correlation, the variables move in opposite directions. These concepts are defined statistically with the correlation coefficient:

$$\rho = \frac{E\{(X - E[X])(Y - E[Y])\}}{\sqrt{V(X)V(Y)}}$$

For computational purposes, the correlation coefficient can be rewritten:

$$\rho = \frac{E(XY) - E(X)E(Y)}{\sqrt{V(X)V(Y)}}$$

Distribution	Function	Mean	Variance
Binomial	$\binom{n}{k} p^k (1-p)^{n-k}$	np	np(1-p) or npq
Poisson	$\frac{e^{-\alpha}\alpha^k}{k!}$	α	α

Figure 4.19

Summary of mean and variance for common discrete distributions.

The correlation coefficient is similar to the definition of variance, except that it measures the deviation from the mean for X and multiplies it by Y's deviation from the mean. The denominator uses the two variances to normalize these deviations, so the correlation coefficient varies between -1 and 1. Values of -1 or 1 represent perfect correlation. If the X and Y values were plotted on a chart, they would fall on a single line when the correlation is perfect. This concept is explored in more detail in the regression chapter.

Figure 4.18 shows some sample correlation coefficients. The correlation coefficient measures the relationship between pairs of variables. Near zero, the scatter plot is essentially random—the variables are independent. As the coefficient approaches one (or negative one), the scatter plot becomes closer to a straight line. Negative values exhibit the same effect but the slope of the line is negative. Keep in mind that correlation is a statistical measure only. It does not imply causation. Higher correlation numbers simply mean the variables move together, but it does not mean that changes in one will always cause changes in the other variable. Causation has to be determined by a model that provides a theoretical explanation for the joint movement.

Discrete Distributions

The parameters of a probability function often determine the mean and variance of the distribution. The expected value and variance can be found by applying algebra to the underlying distribution function. Sometimes the process is tedious, but the values have already been computed for the standard distributions. They are listed here without proof, because they are useful to know when working with the selected distribution. Also, they provide some insight into the distributions.

Figure 4.19 shows the mean and variance for the common discrete distributions. For the binomial distribution, the variance is often written as npq, where q is the probability of failure or 1-p. In the coin-flipping example where p=1/2, pq is equal to 1/4, so the variance for tossing the coin 12 times is npq = 3. The following sections describing continuous distributions will explain the value of knowing the standard deviation.

Important Continuous Distributions

Discrete distributions are useful for understanding the basic concepts and for solving specific types of problems. However, many random variables involve continuous data. More importantly, the continuous distributions are critical to solving cer-

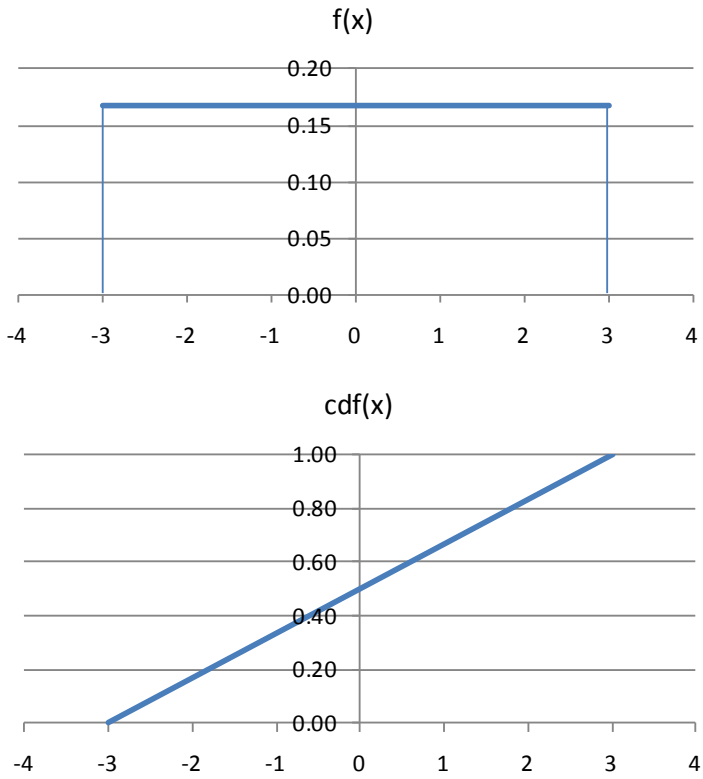


Figure 4.20

Uniform distribution. Data is evenly distributed across a fixed interval. Sometimes used as a neutral starting point for Bayesian analysis.

tain problems. In particular, the normal or Gaussian distribution is the foundation for much of statistics—partly because many of the other distributions, including binomial, converge to that distribution as the number of observations increases. This section presents the most important of the continuous distributions. The section is merely a summary. Details can be found in any statistics textbook.

Remember that probability for continuous variables depends on a specified range. The probability density function defines probability (pdf) as the area under the curve. Consequently, the cumulative distribution function (cdf) is often used to compute actual probability values.

Uniform or Random Distribution

The **uniform distribution** is useful for two reasons. First, it is easy to understand, and second, it represents a basically equally random distribution of data. It is often used as an a priori distribution in Bayesian analysis because it imposes no specific structure on the data. It does require specifying finite minimum (a) and maximum (b) points to limit the range. Figure 4.20 shows the pdf and cdf for the uniform distribution. The distribution has two parameters: a and b, the starting and ending points of the range. Data is evenly distributed, so fixed range has the same probability. Hence, the distribution is sometimes called random because it

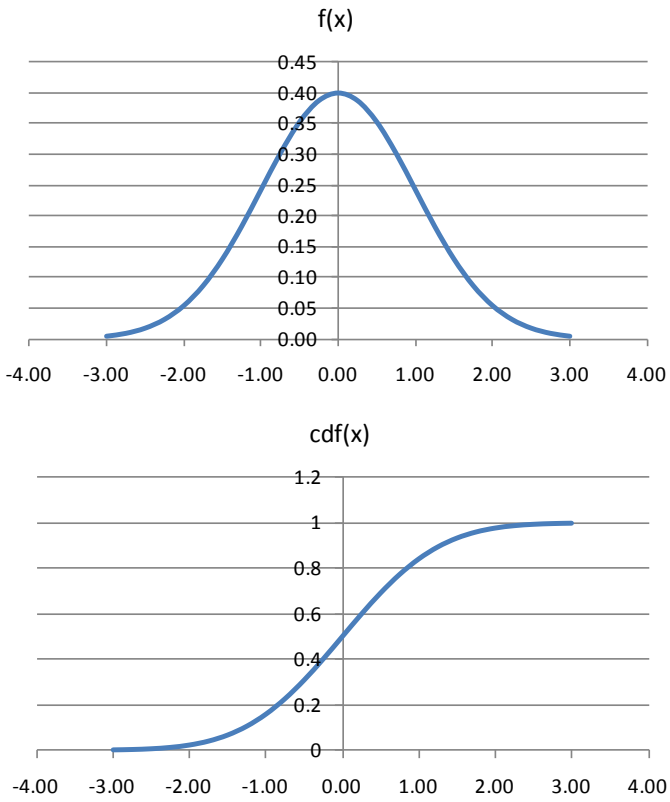


Figure 4.21

Normal or Gaussian distribution. With a large number of observations, most problems will follow a Normal distribution. The distribution is defined by the mean and standard deviation, but all problems can be converted to the standard normal N(0,1) with a mean of zero and standard deviation of one.

imposes no emphasis on any particular values. The pdf and cdf are mathematically easy:

$$f(x) = 1/(b-a)$$
$$cdf(x) = (x-a) / (b-a)$$
$$\text{Mean: } (a+b)/2$$
$$\text{Variance: } 1/12\ (b-a)^2$$

The uniform distribution has limited use except for providing a neutral starting point for some investigations.

Normal or Gaussian Distribution

By far, the most important continuous distribution is the Normal or Gaussian distribution. One of its most important values is that most of the other distributions converge to the Normal distribution as the number of data points increases. So it is the benchmark for most statistical analyses. This section explores some of the fundamental properties of the Normal distribution, but many of the concepts apply

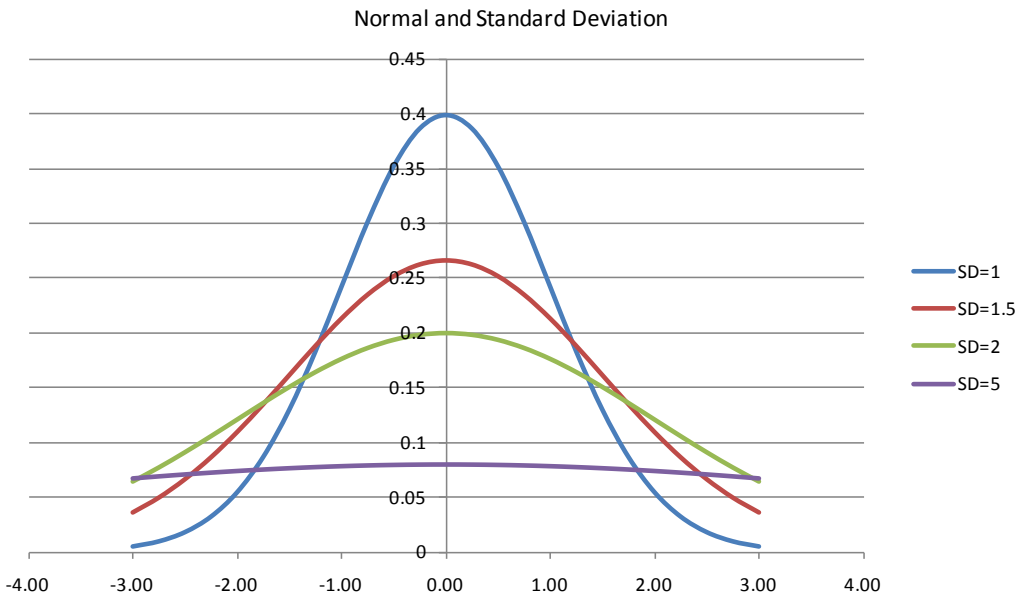


Figure 4.22

Normal distribution and standard deviation. Increases in the standard deviation widen and flatten the pdf, pushing more of the probability into the tails.

to other distributions as well. Figure 4.21 shows the pdf and cdf for the standard normal distribution. The Normal distribution is defined by the mean and standard deviation, but all problems can be converted to the standard normal which has a mean of zero and a standard deviation of one. It is commonly written N(0,1).

The equations for the pdf and cdf are hugely important in statistics and mathematics, but they are mathematically difficult. In particular, there is no simple solution to the cdf. The integration and probability values can be found only through computer numerical analysis.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right)$$

$$cdf(x) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)$$

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

Mean: μ
Variance: σ^2
Conversion to standard normal: $Z = (X - \mu) / \sigma$
Excel: NormDist(X, μ, σ, cumulative)

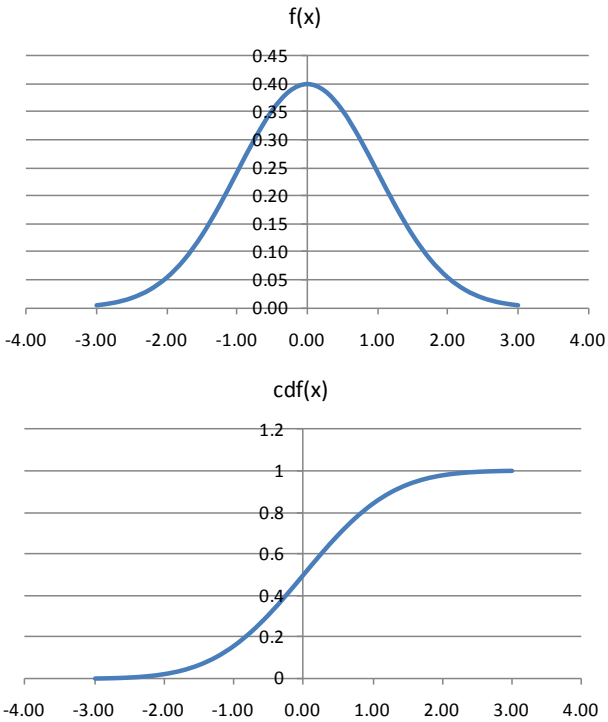


Figure 4.23

Percentage of data falling within one standard deviation of the mean. Area under the pdf between [-1,1], or cdf(1)-cdf(-1) is 0.6827. The percentage for two standard deviations is 0.9545 and for three it is 0.9973.

The Excel function returns the value of the pdf if the entry for cumulative is set to false, otherwise it returns the cdf value.

The mean establishes the center of the distribution—it is symmetric about the mean. As shown in Figure 4.22, increases in the standard deviation widen and flatten the pdf. Consequently, more of the area under the curve or the probability is pushed into the tails away from the center. If the standard deviation is very high, the distribution resembles the uniform distribution, providing little information about the underlying data.

A key aspect to understanding the Normal distribution is to recognize that when a random variable follows the Normal distribution, most of the observations are close to the mean. A smaller standard deviation means that the data are clustered even closer to the mean. To see how close, Figure 4.23 shows how to compute the percentage of data falling within one standard deviation of the mean. The value can be found quickly using the cdf, because it is simply cdf(1) – cdf(-1) = 0.6827. Think about that for a second. In a Normal distribution, over two-thirds of the data fall within one standard deviation of the mean. Check the values for two and three standard deviations: 0.9545 and 0.9973. In a large enough group, with any random variable, almost every value will fall within three standard deviations of the mean. Find a group of people, measure everyone's weight and height. Very few outliers will exist. Give an exam to 100 students and 95 percent of the values will fall

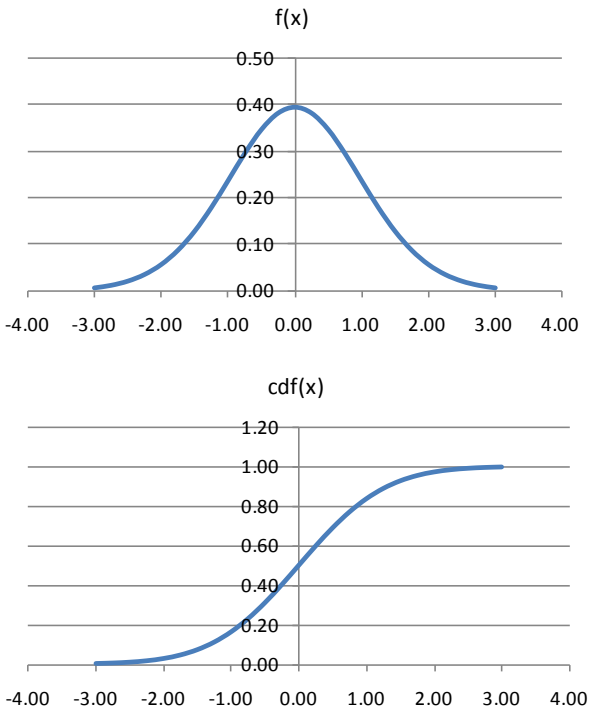


Figure 4.24

T Distribution with 25 degrees of freedom. It is similar to the normal distribution. Standardize the data $T = (X-\mu)/\sigma$. Degrees of freedom are typically n – 1.

within two standard deviations of the mean, and there is only a 0.3 percent chance of someone scoring outside of three standard deviations. These conclusions arise because of the role of standard deviation. If the standard deviation is high, a huge variation in values can still exist.

The mean is important because it shows where the distribution is centered. But the standard deviation describes how tightly the data is clustered around that mean. For example, perhaps daily sales have a mean of 10,000 and a standard deviation of 1000. If the data follow a normal distribution, it is easy to predict that tomorrow the sales total will be between 8,000 and 12,000 with a 95 percent probability. On the other hand, if the standard deviation is 4,000; the 95 percent forecast would be from 2,0000 to 18,000, which is not very useful.

The multivariate normal distribution has also been studied extensively, but it is basically a generalization of the simple normal distribution.

T-Distribution

What happens if there is not enough data to justify using a normal distribution? The Student's T distribution was specifically created to handle problems with small samples. It functions the same way as the Normal distribution, but has one more parameter: degrees of freedom. For most problems, the degrees of freedom are one less than the number of observations, or n – 1. Figure 4.24 shows the pdf and cdf for the T distribution with 25 degrees of freedom. It looks the same as the

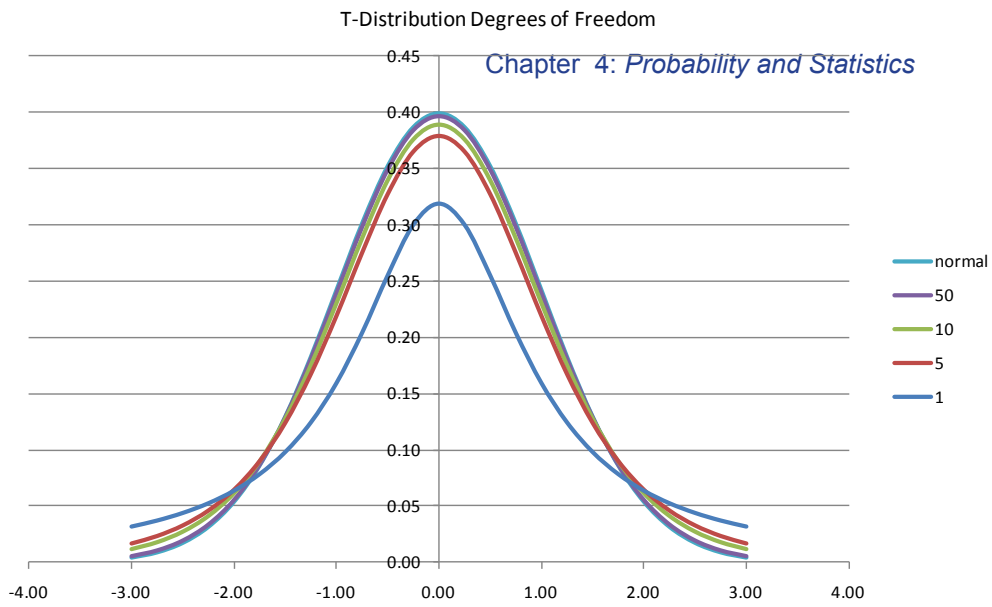


Figure 4.25

T Distribution with different degrees of freedom. Even at df=50, the pdf is almost on top of the normal distribution. At lower degrees of freedom, or smaller samples, more of the probability is moved to the tails.

normal distribution. Just to be complete, the equation for the pdf is given here, but even that is relatively complex because of the gamma function. The cdf is even more complex. Both functions are typically computed using numerical analysis.

$$f(x) = \frac{\tilde{A}\left(\frac{d+1}{2}\right)}{\sqrt{\eth}\tilde{A}\left(\frac{d}{2}\right)}\left(1+\frac{x^2}{d}\right)^{-\left(\frac{d+1}{2}\right)}$$

Mean: 0 (standardized)
Variance: d/(d-2) d>2
d: degrees of freedom, usually n – 1
Excel: TDist(X, df, tails); but it returns the tail, P(T > z)

The Excel formula TDist is different from the normal distribution function. TDist only works for positive values of T and it does not return the pdf or the cdf. Instead, it returns P(T > x), or the probability in the right-hand tail. This value can be converted to the cdf by subtracting it from 1 to get P(T < x) for values of x greater than zero. Because of the way the T distribution is typically used, the Excel formula makes sense, but it is important to realize exactly what is being computed.

To demonstrate the effect of the degrees of freedom, Figure 4.25 plots the T distribution at several different values for the degrees of freedom. With a small number of observations, the pdf is flatter and wider—pushing more of the probability to the tails. Outliers, or extreme effects are more likely to arise. Going the other direction, even at df=50, the plot of the pdf is almost identical to the plot of the normal distribution.

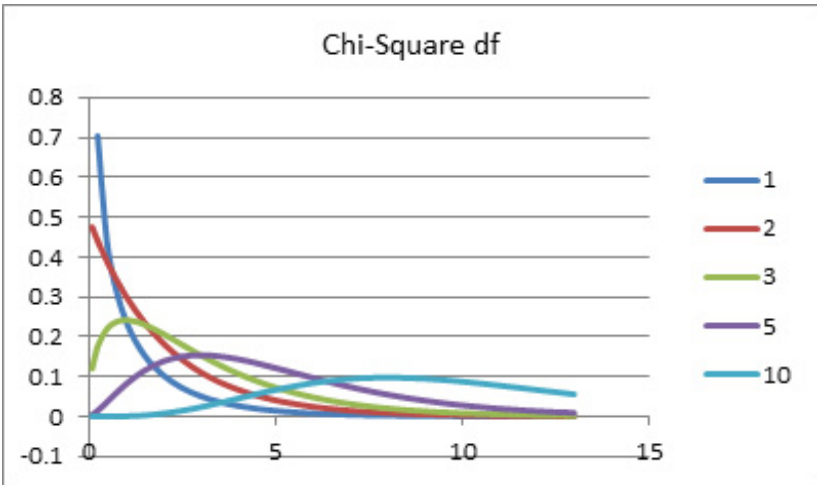


Figure 4.26

Chi-Square distribution at different values of degrees of freedom. Note that it is not symmetric, and more of the probability is clustered at the low end. The different plots are due to different values for the degrees of freedom.

The T distribution, or the normal for large problems, is heavily used in hypothesis testing—which is covered in the statistics sections of this chapter. The T distribution is more accurate for smaller problems.

Chi-Square Distribution

Figure 4.26 shows the Chi-square distribution at varying degrees of freedom. Even a quick glance at the pdf indicates that this distribution is different from the others. It is defined only for positive values of X and positive values of the degrees of freedom. The distribution is not symmetric—more of the probability is concentrated near zero. The distribution is often written with the Greek letter Chi: X_n^2, or Chi-square with n degrees of freedom.

$$f(x) = \frac{\left(\frac{1}{2}\right)^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} x^{\frac{d}{2}-1} e^{-\frac{x}{2}}$$

Mean: d
Variance: 2d
d: degrees of freedom
Excel: ChiDist(X, df)
returns right tail P(X>x)

The Chi-square distribution has some specific uses that are covered in the statistics section. It can also be applied in cases that involve squared random variables. A random variable with an N(0,1) distribution, when squared will have a Chi-square distribution with one degree of freedom. In fact, if the degrees of freedom exceed about 45, and a random variable Y has a Chi-square distribution, a new variable can be defined as sqrt(2Y), and it will have a normal distribution N(sqrt(2n – 1), 1). For now, simply remember that the Chi-square distribution is useful for squared data.

Central Limit Theorem

One of the most important results in the theory of probability and statistics is called the **central limit theorem**. The theorem states that for almost any random variable X, the average of the X values will have a standard deviation of $\sigma/\sqrt{n}$ and the standardized Z value will have a normal distribution for sufficiently large number of observations. In other words, for almost any data the average of that data will follow a normal distribution.

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

This theorem means that almost any data evaluated in the context of data mining can use the Normal distribution. Small samples should use the T-distribution. A few specialized problems benefit from the Chi-Square distribution, but largely, if you understand the Normal distribution, the concepts will apply to most problems.

Statistics

How are statistics used in data mining to find interesting results? The discipline of statistics is used by researchers to determine relationships, find patterns in data, and to test research hypotheses. Many of the tools and concepts developed for use in research have proven useful in data mining. However, some key differences in attitude exist between research statistics and data mining. Data mining is an exploratory process. Its purpose is to provide information to analysts to generate intuition and provide directions for additional research. Consequently, many of the statistical tools are applied differently than they would be for scientific research. In scientific research, certain protocols need to be followed to ensure the data observations are independent and unbiased. Data collection is a critical step in the research. Analyzing data for scientific research also constrains how some procedures are used. This section reviews the key concepts in statistics that often arise in data mining. However, it does not deal with the protocols that are critical to scientific research.

Probability theory in the earlier sections described basic rules of probability and provided the foundations of probability distributions. Probability explains how the population or total set of possibilities is related. Statistics deals with a **random sample** of observations from a population. For example, the population might be the set of all people who are potential customers. A sample might consist of actual customers, or of a set of people randomly chosen and contacted by the marketing department. A **statistic** is a random variable that defines some measure on the random sample. Because a statistic is a random variable, it follows some probability distribution. With enough observations, the central limit theorem states that the normal distribution is a good approximation. The measure can consist of almost anything, such as demographic characteristics of people (age, height, weight, income, and so on), or business outcomes including intention to purchase, amount of money spent each month, or profitability.

Samples

For most research projects, obtaining a good random sample that is representative of the population is critical. A sample is a selection of observations taken from the population of potential data. It costs money and time to obtain research data, so researchers try to obtain the best sample possible at reasonable costs. With data

mining, the concept of a sample is different. Generally, organizations have already collected the data—through traditional transaction processing systems. In many cases thousands, millions, or even billions of rows of data exist. The analyst typically has access to as much data as the computer can process. However, two common situations arise in data mining where samples become useful.

First, sometimes there is simply too much data to analyze it efficiently. If a company has billions of rows of transaction data and needs to perform intensive analyses, it might take too long to obtain results. Some data mining algorithms are designed to handle huge amounts of data efficiently; others that are based on traditional statistics tend to struggle with huge amounts of data. In these cases, it makes sense to take a random sample of observations from the main data set. The analyst can explore the sample data fairly quickly. Any conclusions reached can then be tested against larger samples, or possibly even the entire data set. With preliminary results, it is easier to run one or two tests against the large data set.

Second, data mining faces the risk of over fitting. **Over fitting** is a polite term for stating that the tools and the analyst have pushed too far and tailored the results to the specific set of data. It is one of the key risks that researchers need to avoid, and it is the reason the term data mining was originally a derogatory term. Given a small set of data and enough tools, it is possible to define a model that almost exactly describes every point in the sample. But that model will work only with that specific data set. The model and conclusions could fail completely when applied to other data. One way to reduce the risks of over fitting is to withhold a random sample of the dataset. After the tools are run on the first set of data, they can then be tested against the random sample. If the results are radically different on the second set, then over fitting is a problem. A few data mining tools go even further and split the data into multiple samples. The tools then compute results on each set and combine them to obtain the overall values. By default, most SQL Server tools automatically withhold 30 percent of the data to use as a test sample.

A third possible issue exists with samples, but it is rare in data mining. Sometimes not enough data exists to handle complex analyses. In these cases, a **bootstrap** process can be used. The sample data are analyzed as a distribution and new data is randomly generated that matches that distribution. The initial sample is expanded by randomly adding data that matches the underlying distribution. The process enables the tools to perform more detailed analyses of the data, but there is always a risk that the expanded sample does not completely match the real distribution.

Common Statistics

Some common statistics are computed for almost any measure. Remember that probability distributions have parameters. These parameters provide a mechanism to adjust the distribution to specific cases. In particular, the Normal distribution is completely defined by two parameters: mean and variance or standard deviation. Hence, it is important to estimate values for these two parameters. They are estimated by the **sample mean** and the **sample variance**, which makes them the two most important statistics to be computed for any problem. The sample mean is the simple arithmetic average:

$$sample\,mean\,\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The sample variance is the squared deviation from the mean:

$$sample\,variance\,S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Note that the variance has n – 1 in the divisor instead of simply n. This term is important to keep the estimate unbiased. It also leads to the use of n – 1 as the degrees of freedom when using the T-distribution. Loosely stated, the degrees of freedom are reduced by one (from n) because the same data were already used to estimate the sample mean. Because the sample mean is used in the definition of the sample variance, one degree of freedom is lost.

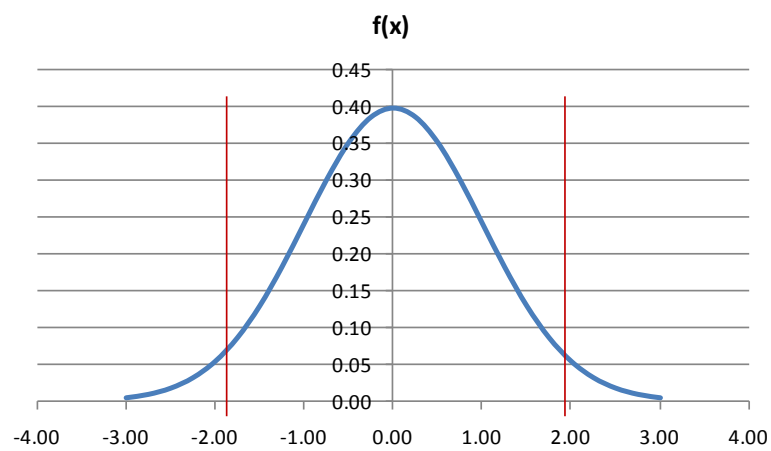
The sample mean and variance commonly appear as parameters in other distributions. They are relatively easy to calculate, and most data mining and statistical packages perform the computations automatically. Even SQL has internal functions to compute these values. Other common statistics include the minimum and maximum values.

With multivariate data, it is also common to compute the sample correlation coefficient to determine the degree of relationship between two variables X and Y. The coefficient is defined as

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Figure 4.27

Find 95-percent confidence interval. Normal distribution, 95-percent in the middle leaves 0.025 probability in each tail. Inverse cumulative normal says that occurs at 1.96, or about 2 standard deviations.



However, it is easier to compute by simplifying the equation:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}}$$

Data mining tools can compute this coefficient automatically. Variations of it are commonly used in regression analysis. It is not necessary to memorize the formula, but it is important to remember that a value of zero means the two variables are uncorrelated or independent. Values of 1 and -1 indicate perfect correlation, but perfect correlation is rare.

Confidence Intervals

Much of the probability and statistics discussion ultimately reduces to one basic concept that is used for two key tools: Confidence intervals and hypothesis testing. The two concepts are two ways of looking at the same question. Remember that a sample is just that—a sample of observations that might or might not be accurate. A key question is to evaluate the sample data and determine its accuracy. Another way to examine the question is to treat the average as a forecast, and then ask about the accuracy of the forecast. Will actual values be close to the forecast, or does a wide range of possible outcomes exist?

A confidence interval is typically created by assuming the statistic follows a normal or T-distribution. Define a Z (or T) value as the difference between the average and true mean divided by the standard deviation of the average:

$$Z = \frac{\overline{X} - \mu}{SD(\overline{X})}$$

The true mean exists but cannot be observed directly, so rewrite the equation:

$$\overline{X} - (Z)SD(\overline{X}) \leq \mu \leq \overline{X} + (Z)SD(\overline{X})$$

Figure 4.28

Find 95-percent confidence interval for average bill in Diners database. A simple totals query retrieves the count, average, and standard deviation of the bill total. Compute the standard deviation of the average and multiply by the Z value of 1.96.

```
Avg(X) = 83.26
StDev(X) = 73,958
N = 73,233
StDev(Avg) = 73.958/sqrt(73,233) = 0.272
Z = 1.96
95-percent CI:  83.26 – 1.96(0.272) , 83.26 + 1.96(0.272)
          = (82.73, 83.79)
```

As shown in Figure 4.27, choose the value of Z so that this interval has a specified probability of containing the true mean. For instance, it is common to pick a 95-percent CI. To get 95-percent of the probability contained in the interval leaves 0.025 percent in each of the two tails. In the old days, you would read through probability tables to find the Z value that leaves 0.025 in each tail. Today, you can use the computerized inverse cumulative Normal function (NormInv in Excel) to learn that point occurs for Z=1.96, which most people round off to 2. An interval that is plus-or-minus two standard deviations wide has a 95-percent chance of containing the mean. In terms of forecasting, future values would have a 95-percent chance of falling within that interval. Some people also like to examine the 99-percent confidence interval. The Z value for a two-tailed 99-percent CI is 2.58.

Finding a confidence interval for the mean requires computing the mean and the standard deviation of the mean. Be careful when reading that sentence. It says "standard deviation of **the mean**," not standard deviation of X. The typical variance or standard deviation functions compute the standard deviation of the original data (X). But the average is the sum of the X values divided by the number of observations n. Based on some probability rules that are beyond this book, the variance of the mean is V(X)/n. Take the square root to get the standard deviation of the mean:

$$SD\left(\overline{X}\right) = SD(X) / \sqrt{n}$$

The process is straightforward, and all of the components have been defined. Consider a simple example using the small Diners database. The owner/chef wants a 95-percent confidence interval for average purchases on a single sale. The data are already stored by individual table or sale and includes the BillTotal. Create a simple totals query:

```
SELECT Avg(Diners.BillTotal) AS AvgOfBillTotal,
StDev(Diners.BillTotal) AS StDevOfBillTotal,
Count(Diners.DinerID) AS CountOfDinerID
FROM Diners;
```

Figure 4.28 shows the results of the query with the average and standard deviation of the Bill Total. Divide by the square root of the number of observations to obtain the standard deviation of the average. Plug these values and the Z value of 1.96 into the CI formula to learn that the confidence interval is (82.73, 83.79). This interval is tight because the standard deviation is low from the large number of observations. Remember that it is the interval for the mean, not for a specific bill. It means that over this time period, the average is an accurate indicator of the overall mean. This number (83.26) could then be used to compute reliable estimates of the revenue for the restaurant. For example, assuming the number of customers can be forecast, multiply by the average bill total to obtain the revenue estimate.

Use the same data to compute the 99-percent confidence interval. The only number that changes is the Z value, which becomes 2.57 instead of 1.96. The resulting interval is (82.56, 83.96). It is wider than the 95-percent CI, which will always happen. To be more confident, the interval has to be wider. But, because the standard deviation of the mean is so low, the effect is negligible—about 25 cents or less than one percent of the mean.

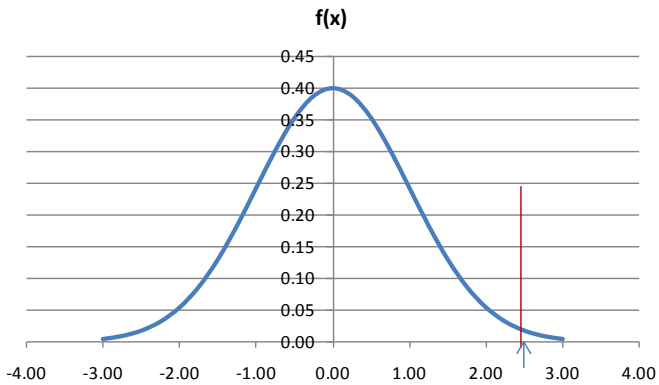


Figure 4.29

Hypothesis test. With large n, compute Z = (Avg-hypothesis)/SD(Avg). Say it is 2.5 in the example. Under the null hypothesis that the mean is zero, what is the probability 2.5 can be observed? N(2.5, 0, 1) = 0.99379 (area to left of 2.5), so tail is 1-0.99379 = 0.00621. Double it for two tails. Still less than a 1.5 percent chance of error if reject null hypothesis.

The process for computing confidence intervals for the mean is similar for any data. Find the average, standard deviation, and count. Compute the standard deviation of the mean and multiply by the desired Z-value. Add and subtract this value from the mean.

The process of creating confidence intervals is applied to many types of problems. The main challenge lies in finding the correct value for the standard deviation. The formula shown in this section applies to the mean. If you wanted to create a confidence interval for a forecast of the original data, the standard deviation is much higher—it includes the standard deviation of the mean and the original data. Many data mining packages create confidence intervals for forecasts and they automatically compute the standard deviations. In most cases, you simply need to understand how to interpret the confidence interval.

Hypothesis Testing

Because it uses the same concepts, mean, standard deviation, and Z-value, **hypothesis testing** is similar to creating confidence intervals. The focus is slightly different, so a few additional concepts and terms are necessary. The goal of hypothesis testing is to decide if a proposed statement (hypothesis) can be rejected. Note the wording—rejected, as opposed to accepted. Technically, probability can only reject statements; which can lead to some interesting configuration of problems. However, most hypotheses in data mining are straightforward. The typical statement is that there is no relationship between variables, or that the effect of a particular dimension is zero. If this type of statement is rejected, then the relationship does exist or the dimension does have a significant impact.

Figure 4.29 shows one way to look at the hypothesis test. Assume enough observations exist so that the mean follows a normal distribution. Compute the test statistic for the problem as Z = (average – hypothesis)/SD(average), where the hypothesized value is zero. Say the result is Z=2.5. Under the assumption that the

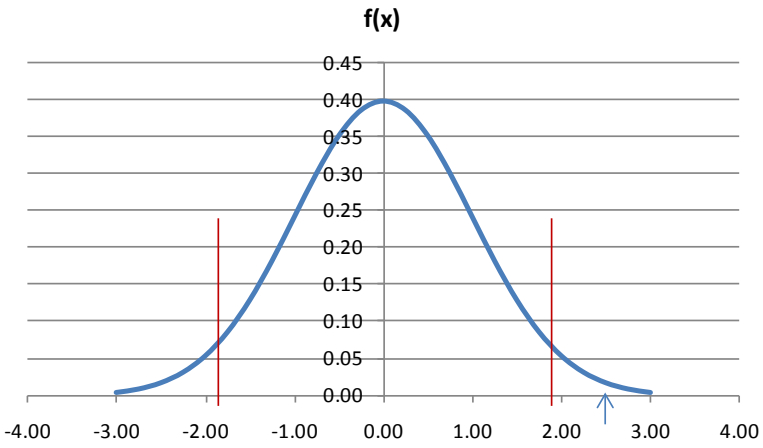


Figure 4.30

Hypothesis test is similar to a confidence interval. Define the level of Type I error to obtain the critical values (Normal with 5 percent in two tails is 1.96). Compute the Z statistic (Avg. – hypothesis)/SD(Avg). If the statistic exceeds the critical value (in absolute value) and falls outside the confidence interval, reject the null hypothesis.

hypothesis is true (zero mean), what is the probability that this result or higher could arise by chance. From the cumulative standard normal formula, N(2.5, 0, 1) = 0.99379 which is the area to the left of the value 2.5. The area to the right in the tail is 1 – 0.99379 = 0.00621; or less than one percent probability. A two-tailed test is more common—where the computed average could be higher or lower than the true mean. In that case, double the tail probability (one for each side), and the result is less than 1.3 percent. So, rejecting the mean could result in being wrong no more than 1.3 percent of the time.

This probability of being wrong is a key element in hypothesis testing. Technically, two types of errors can arise in a hypothesis test:

Type I error: Reject the null hypothesis when it is true.

Type II error: Accept the null hypothesis when it is false.

Of these, the Type I error is usually considered the most important. In dealing with probabilities, no absolute answers exist—random chance could always crop up with atypical results. Most scientists want to be conservative and only reject a hypothesis if strong evidence exists. So, most hypothesis testing begins with a statement of the amount of Type I error you are willing to accept. Common values include 5 percent or 1 percent. A 1 percent error might seem "better," but it increases the amount of Type II error—there is a tradeoff. The only ways to reduce both types of error is to increase the number of observations and decrease the standard deviation.

Defining a level of Type I error simplifies hypothesis testing. The process becomes:

1. Define the null hypothesis.
2. Define a level of Type I error.

Count: 13,497
Average: 153.75
SD(Bill): 93.2479
SD(Average) = SD(Bill)/Sqrt(N) = 0.8026
Z = (153.75– 83.26)/0.8026 = 70.49/0.8026 = 87.82

Figure 4.31

Statistics for Bill Total on Saturdays. Compute the Z-statistic and notice that it is huge! Clearly Z = 87.82 > 1.96, so reject the null hypothesis. Diners on Saturday definitely spend more money on average than diners overall.

3. Look up the critical values as a Z or T statistic, usually as a two-tail test.
4. Compute the Z (or T) statistic: (Avg. – hypothesis)/SD(Avg).
5. If the statistic is larger (absolute value) than the critical value, reject the null hypothesis.

Figure 4.30 shows how the hypothesis test is similar to a confidence interval. When the level of acceptable Type I error is set, the **critical values** can be found (typically 1.96 or 2.58 for 5 percent or 1 percent error). If the test statistic falls outside of this interval, the null hypothesis is rejected because that result can arise with less than the specified probability. That is, the mean is most likely not equal to the hypothesized value (e.g., zero). In the example, the test statistic is 2.5, which is greater than 1.96, so the null hypothesis is rejected.

This approach to hypothesis testing is relatively easy to perform in practice—particularly when the hypothesized value is zero. Most computer programs display the value of the desired item or average, along with the standard deviation. Simply divide the item's value by its standard deviation to obtain the T (or Z) statistic; and many programs display this value automatically. If the absolute value of the ratio is larger than 1.96 (roughly 2), reject the null hypothesis.

Consider a simple example, again using the Dining database. Perhaps in exploring the data, you noticed that people seem to spend more money on some days than on others. Is this difference significantly different from the overall average? To answer this question, formulate a basic null hypothesis: The average bill on Saturdays is equal to the overall average bill of 83.26 found in the previous section. Set a Type I error rate of 5 percent. Remember that the number of observations is in the thousands, so the Normal distribution can be used instead of the T distribution; hence, the critical value is 1.96. Create a database query to compute the average and standard deviation of the bill total on Saturdays.

```
SELECT Count(Diners.DinerID) AS CountOfDinerID, Avg(Diners.
BillTotal) AS AvgOfBillTotal, StDev(Diners.BillTotal) AS
StDevOfBillTotal
FROM Diners
WHERE (Diners.DOW='Sat');
```

Figure 4.31 shows the results of the query, with a count of 13,497 leading to an average Bill Total of 153.75 and a standard deviation of 93.25. Note that the

average bill for all days is about $83 and the average for Saturdays is about $154. Is this difference large enough to be significant? The hypothesis tests answers this question and it does it by including the standard deviation. The figure shows that the Z statistic computes to 87.82 which is a huge number. It is clearly greater than the critical value of 1.96, so the null hypothesis should be rejected, and you decide that people do indeed spend more money on Saturdays than other days.

A statistician would observe that this hypothesis test could be improved. For example, perhaps a better statement would be to compare Saturday averages to averages on other days—leave the Saturday values out of the overall total. It would not change these results, because the average without Saturday is going to be even lower, leading to a larger Z value. But, it could make a difference for other days. Similarly, managers might want to make comparisons between two specific days of the week. Comparing two items to each other leads to problems with defining the standard deviation—because both days might have different values. Statisticians have found several ways to define the standard deviations for comparisons. These distinctions do not arise very often in data mining, but you should ask for help from a statistician if you encounter problems that require comparisons of two sets of numbers. Be particularly cautious if the values are paired—such as observations taken from the same person at different points in times. These cases are not considered here, but the formulas for the variances can be found by searching for paired T test or pooled variance for cases when the variables are independent.

Chi-Square Hypothesis Tests

The T-distribution and Normal distribution are useful for most problems—particularly tests of means. However, some problems are best solved with other distributions. In particular, the Chi-Square distribution is used in two common types of problems. (1) Testing a variance, and (2) goodness of fit. Because variance is defined as the square of the deviation from the mean, the variance follows a Chi-Square distribution instead of a normal distribution—particularly with a relatively

Figure 4.32

Goodness of fit example. A company wants to see if managers are promoting men and women in equal proportions. Assume each manager has the same promotion opportunities. Under the null hypothesis of no discrimination, the total ratio can represent the expected number of promotions.

	Female		Male		Total	
	N	Promoted	N	Promoted	N	Promoted
Manager1	27	15	32	12	59	27
Manager2	21	9	29	11	50	20
Manager3	16	4	21	5	37	9
Manager4	31	7	41	9	72	16
Manager5	9	3	7	2	16	5
Manager6	17	7	11	2	28	9
Manager7	19	5	7	1	26	6
Total	140	50	148	42	288	92
ratio		0.357		0.284		0.319

	Female			Male		
	Obs	**Expected**	**X2**	**Obs**	**Expected**	**X2**
Manager1	15	8.625	4.712	12	10.222	0.309
Manager2	9	6.708	0.783	11	9.264	0.325
Manager3	4	5.111	0.242	5	6.708	0.435
Manager4	7	9.903	0.851	9	13.097	1.282
Manager5	3	2.875	0.005	2	2.236	0.025
Manager6	7	5.431	0.454	2	3.514	0.652
Manager7	5	6.069	0.188	1	2.236	0.683
			7.2347			3.711794

Figure 4.33

Goodness of fit results. To test the chi-square totals, find the chi-square value that has 5 percent in the right-tail with 6 degrees of freedom. The value is 12.59 and can be found using the Excel formula: ChiInv(0.05, 6). Both test statistics are well below this critical value so the null hypothesis cannot be rejected. But, it might be worth looking deeper into some promotions by Manager1.

small number of observations. Hypothesis testing is similar to that for means, but the statistic is:

$$X^2 = (n-1)\, s^2 / \sigma^2$$

The statistic uses the Chi-Square distribution with n – 1 degrees of freedom. It is used to test whether the variance is equal to (or greater than) some value. This type of test sometimes arises in data mining but it is rare.

Goodness of fit is a more interesting application of the Chi-Square distribution. How do you know if the data match a particular distribution? Or, how do you test if data from two sets have the same distributions? The goodness of fit test is one answer. Divide the problem into segments. For each segment, use the probability distribution to compute the expected number of observations in that segment. Then count the observed number of items in that segment. Compute the Chi-Square statistic as:

$$X^2 = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j}$$

This statistic has a Chi-Square distribution with J-1 degrees of freedom. If the number of observed values is consistently different from the expected number, the statistic has a high value and the Chi-Square test will reject the null hypothesis that the data matches the expected distribution. A similar version of this test is used to test for independence between two variables—using a contingency table.

Chi-square goodness of fit is a useful way of comparing data across a company. Figure 4.32 shows a potential use for goodness of fit in a data mining example. A company wants to know if managers are promoting men and women in equal proportions. Retrieving the appropriate data is the first data mining challenge. Under the null hypothesis of no discrimination, and assuming all managers have

Set	H(X)
Uniform random: p=1/25	10.87
Random draw 0 < p < 1	9.5
N(0,1) from -3 to 3	8.1

Figure 4.34

Sample values of Shannon's entropy measure for various sequences of 25 probabilities. Uniform random always has the highest entropy value.

equal opportunities to promote workers, the ratio of the total promoted can serve to compute the expected values for each category and manager.

Figure 4.33 shows the computation of the Chi-square statistics. In the end, the totals seem relatively small. To check, find the Chi-square value with 6 degrees of freedom that has 5 percent probability in the right-hand tail. That value of 12.59 is the critical value for testing within each gender across all managers. Both values are below that critical level, so statistically, the null hypothesis cannot be rejected. It is also possible to test both groups at the same time by adding the two separate sums and checking the Chi-square critical value with 13 degrees of freedom. That critical value is 22.36, which is clearly higher than the total of 10.9. On the other hand, it might be worth checking into promotions by Manager1. Summing the two cells (4.7+0.3) provides a statistic value of about 5. The critical value with 1 degree of freedom is 3.8, so the null hypothesis would be rejected on this data.

Statistical tests of this nature can be powerful tools to automatically scan the data looking for anomalies. But, keep in mind that standard statistics practices are geared towards research. It is quite possible that small patterns exist that will not be revealed with a 5 percent Type I error rate. That is, patterns can exist but not be strong enough to pass a rigorous research test. One solution is to increase the Type I error rate, reducing the critical values and highlighting more potential anomalies. Just be cautious when interpreting the results, and use other information to make critical decisions. Data mining used in this manner is useful for exploration and providing directions for further study. It generally cannot be used for proving a case or making a final decision.

Information Measure

There is a statistic that is commonly used in data mining that is rarely presented in introductory statistics classes. In 1948, Claude Shannon (Shannon 1948) published a paper dealing with communication theory. A key aspect of communication is the ability to deliver information, so Shannon formulated a mathematical definition of information. His main point is that information must contain something new, or perhaps surprising. The classic example is a coin flip. If a coin has two heads and no tail, it will always come up heads. Hence, the coin flips provide zero information—there is no surprise. At the other extreme, if a coin is fair, the probability that a head is revealed constitutes new information because there was no way to predict the value ahead of time.

Shannon's **information measure** or **entropy** is defined for a random variable X as

$$H(X) = E[I(X)]$$

In the definition, E is the standard expectation function from probability and I is an information measure of the X variable. Information is based only on the probability of the event occurring, and events with more surprise have more information (1/p), so the function I is defined as

$$I(X) = \log(1/p) = -\log(p)$$

Because Shannon was dealing with bits transferred, he used the base-2 log and information was measured in bits. However, the definition is neutral and could use any base for the log function. To understand the information function, consider the simple coin toss. With a fair coin, p = 1/2. When a head is tossed, the information revealed is log2(1/0.5) = log2(2) = 1. In a communication sense, it takes one bit to send the information that a head was tossed. Similarly one bit is needed to hold the information that a tail was tossed. The bit could be 1=heads, 0=tails. In a larger problem, rolling a fair die, a specific number (e.g., 5) has 1/6 probability of appearing, so -log2(1/6) = 2.585, or almost 3 bits to carry the information content of the results of a roll of a die.

Combining these definitions, the standard measure of Shannon's entropy is:

$$H(X) = -\sum_{i=1}^{n} p_i \log(p_i)$$

where p_i is the probability of x_i appearing, and the values are summed over all possible values of X.

The formula is easy to compute, but it requires the probability values for each of the outcomes. To understand it a little better, build a simple spreadsheet and compute the entropy measure for a variety of sets. Figure 4.34 shows the results for three different random variables. A uniform random distribution would have the same probability (1/25) at each point. A random draw was created to define 25 probabilities, and the normal pdf was computed at 25 points from 3 to 3. Uniform random variables will always have the highest entropy value, because the probabilities are equal and the results cannot be predicted, so the surprise value is the highest. Consequently, Shannon's entropy is a measure of randomness. Consequently, the measure is often used to evaluate different models to see which one reduces entropy (randomness) the most. The entropy measure is sometimes used to help determine if a dimension variable has enough information to be useful in further analysis. For instance, if gender were a dimension but all of the data was based on women, the Shannon information value would be minimal and there would be little point in including that dimension. Conversely, the information gain on a target variable is a useful way to evaluate the impact of adding new dimensions.

Summary

Probability and statistics are useful and powerful tools that form the foundation of data mining. The rules of probability provide the mathematical foundations that lead to probability distribution functions. These distribution functions are used to evaluate statistics and eventually to make hypothesis tests about the data. Random variables can contain discrete or continuous data and the distributions are different. The binomial and Poisson distributions are commonly used for discrete data. The Normal distribution for continuous data is the most important distribution because the central limit theorem says that all distributions approach the Normal

when the number of observations is large. The T-distribution corrects for degrees of freedom biases in small samples. The Chi-square distribution is used to perform hypothesis tests on the variance, but it is more commonly used to test for equality of distributions and independence.

The Normal and T-distributions are commonly used for testing hypotheses on the mean. The process is the same for both, but the T-distribution requires the number of degrees of freedom—which is usually n – 1. The basic process is to compute the test statistic as (Avg. – hypothesis)/S.D.(Avg). The null hypothesis is typically that the mean is zero, so the computation simply becomes the estimated average divided by its standard deviation. This ratio is generally reported by data mining tools. If the ratio is larger than the critical value (1.96 for Normal at 5 percent error) then the null hypothesis is rejected. Values that exceed the critical level tend to be important in data mining problems.

Bayes' Theorem is another useful result from probability theory that is used in various aspects of data mining. It defines how to compute a conditional probability. Derived from basic probability rules, the theorem is best understood in terms of subjective probabilities. Beginning with a prior believe about probability (or its distribution), Bayes' Theorem shows how to use data observations to update the probability to a new posterior probability.

Probability and Statistics have detailed high-level mathematical definitions. Some basic definitions are presented in this chapter simply to illustrate these foundations. Fortunately, statistical data mining tools incorporate this knowledge and perform the computations automatically, so it is not necessary for managers to know all of the details. Instead, it is critical that managers using data mining tools understand the fundamentals and be able to interpret the results and conclusions. Two of the most difficult problems faced by analysts are (1) choosing the appropriate tools, and (2) understanding the results and limitations of the tools. A basic knowledge of probability and statistics is required to perform these tasks.

Key Words

addition rule
arrangements
Bayesian
bootstrap
central limit theorem
combination
conditional probability
contingency table
continuous
critical values
cumulative distribution function (cdf)
discrete
discretizing
expected value
experiment
general multiplication rule
goodness of fit
hypothesis testing
independent
information measure
joint events
joint probability
margin totals
multiplication rule
mutually exclusive
over fitting
parameters
permutation
Poisson distribution
probability
probability density function (pdf)
probability distribution
probability function
probability mass function
random
random sample
random variable
relative frequency
sample mean
sample variance
Shannon entropy
standard deviation
statistic
subjective probability
uniform distribution
variance

Review Questions

1. What is the probability of two events both occurring if they are independent?

2. When counting outcomes, what is the difference between arrangements, permutations, and combinations?

3. What is the general probability rule for computing the probability that event A or event B can occur?

4. What is Bayes' Theorem and why is it important in data mining?

5. How can a tree diagram represent Bayes' Theorem? Hint: Draw a simple tree and explain it.

6. What is the difference between discrete and continuous data?

7. What is expected value?

8. When evaluating data, why is it so important to look at the variance as well as the mean?

9. What is the purpose of the correlation coefficient?

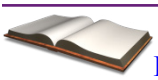
10. When should a T-distribution be used instead of a Normal distribution?

11. How is a confidence interval similar to a hypothesis test?

12. What is the test statistic for comparing an observed distribution of data to another specified distribution?

13. When does Shannon's entropy measure reach its peak values?

Exercises



Book

1. A company has just hired 15 interns and wants to assign them to teams of 5 people each. The team members will be rotated every month. One manager has suggested it would be nice to set up all possible teams. How many possible teams can be made from these 15 people?

2. A company produces a complex machine that has 10 key components. The probability of any individual part failing is given in the table. Fortunately, the system is not dependent on all parts equally. Some parts are essentially backups for others. The system fails if: (a) C1, C5, and C7 all fail, (b) C8, C9, and C10 fail, (c) C2 and C4 fail, or (d) C3 and C6 fail. What is the probability any individual machine will fail? If the company sells 10,000 machines, what is the expected number of failures?

   Individual failure rates:

1	2	3	4	5	6	7	8	9	10
1/100	1/300	1/500	1/50	1/800	1/700	1/100	1/200	1/50	1/600

3. A factory produces identical components from two machines. Machine A is older and is known to produce defects at the rate of 1/1,000. Machine B produces defects at the rate of 1/5,000. After a daily run of 5,000 parts produced from A and 10,000 parts from B, an inspector discovered that a part in a combined box is defective. Draw the decision tree and compute the probability that the defective part came from Machine B.

4. Use Excel to create a table and a chart for a binomial distribution where the probability of success is 3/4 and there are 20 trials.

5. A bank is thinking about adding a second ATM machine. On Fridays at the end of the month during lunch hour, an average of 20 people an hour use the existing ATM. The one machine can barely handle those 20 people because each takes an average of 3 minutes per person to complete their transactions. Compute the probabilities of from 21 – 30 people arriving to use the ATM. The manager has decided that if 30 people arrive too often, a second ATM will be needed.

6. Given the values in the following table, find the probability that the mean is greater than zero using both a Normal and T-distribution for each of the row entries.

Avg	SD(X)	N
5	5	100
5	5	10
5	20	100
2	10	50
2	10	100

7. Given the data in the following table, create the 95-percent confidence interval for the mean for all the data points.

9.12	11.45	5.07	3.86	6.32	2.45	4.40	4.57	7.30	5.81
3.17	10.87	-0.33	2.90	-1.30	5.41	3.47	2.31	5.48	4.66
8.57	5.12	4.07	3.68	5.34	9.33	-0.45	2.51	7.30	4.52

8. Given the following values reported by the data mining software, determine whether the mean is significantly different from zero in each of the row cases.

Avg	SD(X)	N
2.5	70	100
12.3	35	20
-5.4	125	50
18.4	65	100
21.7	25	50

9. Your ten coworkers have accused your boss of making decisions by flipping a coin. In an attempt to disprove this hypothesis and argue that your colleagues do not understand your boss, you have asked each of them to count the number of "yes" responses out of the next ten questions asked of the boss. The following table records the number of times the boss said "yes" out of ten requests. Using the data, is your boss really flipping a coin to make decisions?

# Yes	0	1	2	3	4	5	6	7	8	9	10
Count	0	0	0	0	1	1	2	3	2	1	0

10. Create an Excel spreadsheet to compute the following values:

    a. Normal probability for X > 32 when the mean=25 and std. dev. = 5.

    b. Binomial probability finding three or fewer errors in 10 trials where the probability of any given error is 1/10.

    c. A salesperson is paid $100 for every new customer who places an order in the next month. If the average salesperson gets 3 out of 50 people to place an order, how much money can a salesperson expect to make by calling 1000 people?

    d. Someone has estimated the following probability distribution for events. What is the Shannon Entropy measure?

Event	Total destruction	Irreparable damage	Major damage fixable in a month	Minor damage	Complete success
Probability	0.10	0.15	0.15	0.25	0.35

Rolling Thunder Database

11. Compute sales by state for 2010 and decide if sales to CA are different from the average.

12. Compute sales (count) by gender for 2012 and determine if men buy more bicycles than women.

13. Are sales of bicycles by model type for 2011 evenly distributed based on count?

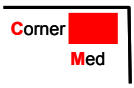
14. For 2012, use the relative frequencies to estimate the probability of bicycle model type. For each model type, estimate the probability of the bicycle having a carbon frame given the model type. Then use Bayes' Theorem to compute the probability that a carbon bike just sold is a race bike.

15. For 2012, what is the correlation between bicycle size (FrameSize) and sale price for road and race bikes (combined)?

Diner

16. Is the purchase of desserts equally probable on each day of the week?

17. Build a frequency distribution of total sales (BillTotal) by day of week. Does it resemble any of the distributions covered in this chapter? Hint: If necessary, use the Excel Frequency function.



Corner Med

18. Create a decision tree with estimated probabilities, from patient gender, age (<5, 6-50, >50), and tobacco use.

19. Create a frequency distribution histogram by count of the number of visits covered by each insurance company.

20. Create a frequency distribution by counting the number of visits for each top-level ICD10 procedure code (first letter).

Basketball

21. For each team, compute the number of wins for non-playoff games in 2010-2011 and compute the team's free throw percentage. Across the teams, what is the correlation coefficient between these two variables? Hint: Use the TeamGameTotals view.

22. Compute each team's win percentage for the regular season (82 games). Using this data as p, compute Shannon's entropy. Compare the resulting value to the entropy of all 30 teams winning exactly 50 percent of their games.

23. Use a Chi-Square test to determine if each division has the same potential/record. That is set the null hypothesis that each division wins the same number of games in the 82-game season for 2010-2011.

24. Across all teams, did guards score more points on average than centers did in 2010-2011?

Bakery

25. In terms of average sales value (quantity * sale price) per day, do breads sell more than the average of all other products? Hint: Compute sales per day by category.

26. Using relative frequencies, what is the probability that bread and cake are purchased at the same time. Hint: Count the total number of sales that include at least one bread item and one cake.

Cars

27. Compute the correlation between MPG and acceleration (seconds to 60 mph). Comment on the results.

28. Create a frequency distribution histogram for the MPG values. Comment on the chart. Does the result resemble any of the distributions covered in this chapter?

29. Divide weight into three categories: light, medium, and heavy. Do the same for horsepower and price. Create a decision tree for these three attributes and include the probabilities for each node. Comment on any patterns.

Teamwork

30. Choose one basketball team. Compute the average free throw percentage and standard deviation for the entire team for one season. Assign at least two players to each person on in your group. Compute the player's free throw average and determine if it is significantly different from the average for the team.

31. Split the bakery data into days of the week and assign one day to each team member. Have each person compute the observed probability of a cake and bread item being purchased at the same time on the specified day of the week. Compare the team's results for the different days.

32. Look at the cars by the predefined category. Assign one category to each team member. Compare the average MPG, Price, and Weight to the overall averages. Are they statistically different for the specified group? Combine the team's results and comment on any results.

Additional Reading

Mlodinow, Leonard, 2008, *The Drunkard's Walk: How Randomness Rules Our Lives*, Pantheon Books: New York. [A collection of examples on how difficult it is to apply probability to everyday life. No mathematics needed.]

Shannon, Claude,1948, "A Mathematical Theory of Communication," Bell System Technical Journal, 47, 379-423. [Shannon's original definition of information.]

Trevor Hastie, Robert Tibshirani, and Jerome Friedman, 2001, *The Elements of Statistical Learning*, Springer: New York. [An outstanding book on data mining, with an emphasis on theory. A graduated-level book that requires a strong mathematics background.]

Zellner, Arnold, 1971, *An Introduction to Bayesian Inference in Econometrics*, Wiley: New York. [A classic book on Bayesian theory, particular focus on subjective probabilities and how they can define traditional analyses. Graduate level with mathematics.]