L. A Math: Romance, Crime and Mathematics in the City of Angels

By James D. Stein

Supplementary Mathematical Material for Website

Mathematical Accompaniments

Chapter 1 - Symbolic Logic

Before you dive into this section, let me repeat something I said earlier – you don't have to read this! I think a fair amount of learning takes place by erosion – if you're simply exposed to something often enough, it will sink in. Maybe not deeply, but enough to give you the idea. Hang around with musicians, you get some idea of what goes into music – maybe not enough to make you a musician yourself, but you'll be a lot more knowledgeable about it than if you spent no time on it at all.

However, I hope that you'll try reading some of these sections. They're not deep, and if you get fed up, you can always go on to the next story.

Sherlock Holmes was fond of telling Watson that, when you eliminate the impossible, whatever remains, no matter how improbable, must be the truth. It's certainly a simple, insightful, and elegantly phrased remark. However, it does not come as a stunning surprise, because most of us are already aware of the inherent logic behind Holmes' statement.

It is perhaps fitting that Sherlock Holmes, for whom logic was a *sine qua non*, was an Englishman, for it was another Englishman, George Boole, who was most responsible for the invention of symbolic logic.

Prior to George Boole, mathematicians concerned themselves with mathematical objects such as numbers and geometric figures. A goal of mathematics then, as now, was to prove theorems -- about numbers, geometric figures, and such. To George Boole goes the honor of being the

first extensive investigator of the nature of proof (not to slight the Greek philosophers, who made the initial contributions in this area).

One of the reasons that mathematics is so successful that it does is its relentless focus on concepts whose definitions are unambiguous. Boole focused his attention on statements or sentences which were either unambiguously true or unambiguously false. Such statements are called propositions. From now on, the letters P and Q will be used to denote propositions, and the letters T and F will be used as abbreviations for “true” and “false”, respectively.

To each proposition there is an opposite, the proposition we shall denote by NOT P. A proposition can be negated simply by sticking the phrase "It is false that ... " in front of the proposition. For instance, if P is the proposition "Today is Thursday," the opposite of P (sometimes called the negation of P) is the proposition "It is false that today is Thursday."

From simple propositions more complicated ones can be built up through the use of the logical connectives OR and AND. At this stage, let us introduce what mathematicians call a "convention" (this is probably short for 'conventional agreement'). We will use the words OR and AND to denote the logical connectives that enable us to construct complex propositions from simple ones according to the following rules.

If P and Q are propositions, the proposition P OR Q is true if either P is true or Q is true, or possibly both are true. This is known as the inclusive “or” – if your waiter asks “Will you have coffee or dessert?” and you answer “yes”, you mean that you will either have coffee, or dessert, or maybe both. There is also the exclusive “or”, as in “I will either drive to San Diego tomorrow or take the plane,” obviously, you’re not going to do both. However, mathematicians decided to

use the inclusive “or”, just like they decided to use the symbol + for addition rather than something else. It’s important to have everyone in agreement about what terms mean.

The proposition P AND Q, however, is true only when both P and Q are true.

This information can be quickly summarized in tabular form. The layout that follows is known as a truth table.

P	Q	P OR Q	P AND Q
T	T	T	T
T	F	T	F
F	T	T	F
F	F	F	F

With the four rows, we have covered all the possible true-false combinations for the two propositions P and Q, and the other columns give the truth values of the proposition at the top of the column for the truth values of P and Q in the same row.

Here's a very informative, very short truth table.

P	NOT P	P OR NOT P	P AND NOT P
T	F	T	F
F	T	T	F

In words, for any proposition P, the proposition P OR NOT P is always true, and the proposition P AND NOT P is always false. Propositions which are always true are known as **tautologies**.

Some truth tables are important, but not especially interesting. It is easy to show that P OR Q and Q OR P have the same truth table. This isn't surprising, as if your waiter asks "Will you have dessert or coffee?", it's the same question as if he asked "Will you have coffee or dessert?" Similarly, P AND Q and Q AND P have the same truth table.

It is possible to use parentheses to construct ever more complicated propositions, much as parentheses are used in arithmetic and algebra for exactly the same purpose. If P, Q, and R are propositions, we can construct the compound proposition P OR (Q AND R) by first constructing the proposition Q AND R, and then taking that proposition and OR-ing (as the computer folk are fond of saying) the proposition P with the proposition Q AND R.

We can compute the truth value of a complicated proposition from the truth values of its components simply by stripping away levels of parentheses. Just as we compute the numerical expression (2+3) x (3 x (4+7)) by working from the inside out, we can do the same thing with complex propositions.

Arithmetically, (2+3) x (3 x (4+7)) = 5 x (3 x 11) = 5 x 33 = 165. Suppose now that P is true, Q is false, and R is false. To compute the truth value of the following logical expression, (P OR NOT Q) AND (NOT P OR R), let's replace each proposition by T or F as we compute it.

1) (P OR NOT Q) AND (NOT P OR R)

2) (T OR NOT F) AND (NOT T OR F)

3) (T OR T) AND (F OR F)

4) T AND F

5) F

Now let's return to Sherlock Holmes. How can we analyze his remark that, when you have eliminated the impossible, whatever remains, however improbable, must be true?

Let's suppose that P and Q are propositions such that P OR Q is true. Suppose further that Q is false. How can we conclude that P must be true?

One look at the truth table for P OR Q should make it fairly obvious.

	P	Q	P OR Q
1)	T	T	T
2)	T	F	T
3)	F	T	T
4)	F	F	F

Since P OR Q is true, line (4) is eliminated. Since Q is false, lines (1) and (3) are likewise out. No matter how improbable P may be, it must be true, as line (2) is the only one remaining, and P is true in line (2).

Now things get a little complicated. Boole decided to assess the validity of the argument IF P THEN Q on the basis of the true-false values of the propositions P and Q. What Boole decided

was that the important thing was to make sure that any argument which started with a true premise (the premise is the P in IF P THEN Q) and ended with a false conclusion (the conclusion is the Q in IF P THEN Q) would be labeled as false. After all, if you start with the truth and reach a false conclusion, your argument must be fallacious. In order to single out these fallacious arguments, Boole made all other IF P THEN Q statements true, by fiat.

This resulted in the following truth table for IF P THEN Q.

P	Q	IF P THEN Q
T	T	T
T	F	F
F	T	T
F	F	T

Let's look at the compound proposition IF ((P OR Q) AND NOT Q) THEN P. So that everything will fit on one line, let R denote the proposition (P OR Q) AND NOT Q

P	Q	NOT Q	P OR Q	(P OR Q) AND NOT Q	IF R THEN P
T	T	F	T	F	T
T	F	T	T	T	T
F	T	F	T	F	T
F	F	T	F	F	T

No matter what the truth values of P and Q, the proposition

IF ((P OR Q) AND NOT Q) THEN P

is always true!

Admittedly, when Sherlock Holmes used it, he assumed implicitly that either P or Q is true. Nonetheless, no matter what the truth values of P and Q, IF ((P OR Q) AND NOT Q) THEN P must be true.

Boolean logic, as this branch of mathematics is known, has gone far beyond what Boole could ever have imagined. Not only is your computer constructed on its principles, every time you do an Advanced Search with a search engine, you are using Boolean logic as well.

Chapter 2 - Percentages

As Pete observes in the story, percentages are a source of substantial confusion. Most of this confusion comes from a mistaken belief that percentages work the same way as numbers, with a gain of 20% compensating for a loss of 20%. As we saw in the story, a gain of 20% does not compensate for a loss of 20%, because the 20% gain is not figured using the same amount as the 20% loss.

A knowledge of basic algebra can be quite helpful in eliminating much of the confusion surrounding percentages. There's a little algebra in the material that follows, but hopefully not enough to cause you sleepless nights.

Innumeracy, as Pete points out in the story, is the arithmetic equivalent of illiteracy. Those who have succumbed to illiteracy, however, realize that they cannot read. They know this can profoundly affect their lives, and often take steps to remedy this problem.

Innumeracy is much more insidious than illiteracy. The victims often do not realize they are innumerate. Illiteracy is condemned, but innumeracy is not regarded in the same light. There are those who feel that such attributes as artistic creativity go hand-in-hand with innumeracy -- indeed, there are even some who proudly flaunt their innumeracy.

This is a great pity because, like any disease, the ripple effects of innumeracy spread throughout our society. It is probably not an exaggeration to say that elimination of innumeracy would save our society tens, and perhaps hundreds, of billions of dollars annually.

Any study of innumeracy would undoubtedly find that confusion concerning percentages is a contributing factor. The term 'per cent' is an abbreviation of the Latin phrase 'per centum', which means 'each hundred'. Thus 3% means 3 for each hundred. 3% of 100 would be 3, and 3% of 400 would be 12.

Computing Percentages

To find a percentage of a given number, multiply the number by the percentage, and divide by 100.

Example 1 - To find 7 1/2% of $350, multiply 7 1/2 by $350, obtaining $2625, and then divide by 100 to get $26.25 .

Computing percentages is straightforward, and most people do not have too much difficulty doing so. Finding the cost of an item on which the sales tax is known is a little more difficult, and requires setting up and solving a simple equation.

Suppose we want to find the purchase price of an item on which a 7% sales tax came to $11.20 . Let P denote the purchase price. 7% of P is 7 P / 100 = .07 P . Therefore

$$.07\ P = \$11.20$$

$$P = \$11.20 / .07 = \$160$$

Notice that we can check that $160 is the right price simply by taking 7% of $160 and observing that it equals $11.60 .

Example 2 - Rutabaga Preferred stock rose 15% last year. If the stock went up 13 1/2 points, at what price did it start the year? (Note: a "point" is actually investor-speak for a dollar. A stock selling for 50 points is selling for $50 for one share of stock.)

Solution: Let S denote the starting price. 15% of S is .15 S, so

$$.15\ S = 13.5$$

$$S = 13.5 / .15 = 90$$

The stock started at 90. Checking, 15% of 90 is 13.5 . ■

From here we move on to mark-ups and mark-downs. Suppose that the total bill for a meal at a restaurant in which the sales tax was 6% came to $33.92 . How much was the meal itself?

Obviously, the total bill was the sum of the cost of the meal, which we call M, and the added tax. As we just saw, 6% of M is .06 M. Therefore

cost of meal + sales tax = $ 33.92

M + .06 M = $ 33.92

1 M + .06 M = $ 33.92

We write M = 1 M so we can use the distributive law. Notice that M + .06 M = 1 M + .06 M = (1 +.06) M = 1.06 M .

1.06 M = $ 33.92

M = $ 33.92 / 1.06 = $ 32.00

Example 3 - After an 8.5% sales tax, the price of a car is $9439.50. What was the price of the car, not including tax?

Solution: If C represents the pre-tax price of the car, the sales tax is 8.5% of C, or .085 C. Therefore

C + .085 C = $9439.50

1.085 C = $9439.50

C = $9439.50 / 1.085 = $8700

Checking, 8.5% of $8700 is .085 x $8700 = $739.50, and so the total price would be $8700 + $739.50 = $9439.50 . ■

Discounts seem to cause more trouble than mark-ups. As an example, suppose that after a 15% discount, a television cost $153. What was the original price of the television?

If we let T denote the original cost of the television, the discount of 15% is given by the expression .15 T. As before

original cost - discount = $ 153

T - .15 T = $ 153

.85 T = $ 153

T = $ 153 / .85 = $ 180

It is always easy to check problems like this. Since 15% of $ 180 is $ 27, the original price less the discount is $ 180 - $ 27 = $ 153 .

CAUTION!!!

A commonly-made mistake is to compute the price from which a discount is taken by adding that percentage to the discounted price.

Example 4 - After a 20% discount, a TV sells for $120. What was the original price?

Solution: The mistake is to take 20% of the discounted price of $120, which is $24, and add that to the discounted price of $120, arriving at an erroneous original price of $120 + $24 = $144.

You, of course, now know better. Let S denote the original price of the TV. 20% of S is .20 S, and so

S - .20 S = $120

.8 S = $120

S = $120 / .8 = $150

Checking, 20% of $150 is .20 x $150 = $30, and when $30 is subtracted from $150, the result is $120. ■

Now let's take a look at the trap that the city of Linda Vista fell into. Why can't we reduce an individual's taxes by 20% if the tax base increases by 20%? If there are originally T taxpayers, and each taxpayer is assessed D dollars, then the total revenue is clearly TD dollars. A 20% increase in the number of taxpayers will add .20 T taxpayers to the original T taxpayers, so there will now be T + .20 T = 1.2 T taxpayers. A 20% reduction in the taxes assessed each taxpayer will be .20 D dollars, so each member will now pay D - .20 D = .80 D dollars. Therefore, the total revenue will be the number of members, multiplied by the tax per member. This comes to 1.2 T x .8 D = .96 TD dollars. This is only 96% of the original revenue.

Incidentally, did you catch onto the fact that Pete was able to compute the number of taxpayers simply from the information that taxes had been reduced from 100 dollars to 80 dollars and that the city of Linda Vista was $396,000 short? Once he found out that everybody assessed had paid up, Pete hypothesized that the City Council had fallen into the classic innumeracy trap of thinking that a 20% gain in the number of taxpayers compensated for a 20% loss in revenue per taxpayer. In that case, as we discovered above, the shortfall would have been 4%, since the revenue was only 96% of the original revenue. If the total revenue is denoted by R, then 4% of R would be $396,000. Therefore

$$.04\ R = \$396{,}000$$

$$R = \$396{,}000 / .04 = \$9{,}900{,}000$$

At $100 per taxpayer, the number of taxpayers in the previous census would have been $9,900,000 / 100 = 99,000 taxpayers.

Innumeracy in this respect can have potentially catastrophic consequences. A doctor may tell a nurse to reduce the dosage of a drug by 50%. When the patient relapses, the doctor tells the nurse to

raise the dosage by 50%. Disaster! The doctor may think the patient is receiving the same amount of medication as he or she did originally, but the patient is only receiving three-quarters of the original amount. One shudders at the thought of a similar error being made with an airplane whose fuel has been depleted by 50% .

Many of the misunderstandings concerning percentages occur because of a failure to realize that the computation of a percentage requires a base number upon which one computes the percentage. Suppose that a stock is selling for 100, and the price rises 30% and then declines 30%. The base number for figuring the price rise percentage is 100. 30% of 100 is 30, so the price after the rise is 130. This number, 130, is the base number for figuring the price decline of 30%. 30% of 130 is 39, so the stock price falls to 130 - 39 = 91. Notice that, even if the stock had fallen 30% first and then risen 30%, the final price would again be 91. That's because we are performing two successive multiplications, and it doesn't matter in which order we perform them.

Failure to understand percentages has repercussions in other areas of mathematics. Percentages are often used to convey probabilistic notions. For instance, a weatherman might say that there is a 50% chance of rain on Saturday and a 50% chance of rain on Sunday. A recent poll showed that many people were under the impression that the above forecast was equivalent to saying that it was certain that it would rain during the weekend! Well, it's not, and we shall have more to say about this in Chapter 9.

Chapter 3 - Averages and Rates

Introduction

If mathematicians were to vote on the most useful notion in mathematics, the concept of averages would be up close to the top. In fact, many would probably award it the title. And that's the reason that there's so much accompanying material to the story for this chapter.

Once again, you don't have to read any of it. But if you're planning on taking any math courses and you're not a stellar math student, it's probably a good idea to read it.

Averages occur throughout all of mathematics. They represent one of the best ways of summarizing past information, and in the absence of more pertinent data, the best way to predict the future. Averages play critical roles in such widely diverse topics as percentages, probability and statistics, algebra, and calculus.

An average is a quotient, and one model for division is sharing a quantity of items in a fair fashion. If four people are to share twelve slices of pizza fairly, how many slices should each person receive?

Trivial as this example may seem, it can be used to provide an easy introduction to the concept of an average. If twelve slices of pizza are shared among four people, the average number of slices each person receives is three. Of course, this does not mean that each person actually receives three slices. An average in this instance represents a way of summarizing data by looking at what would have happened if fair sharing had actually taken place.

When we say that the average is three slices, there is a very important, but often unspoken phrase: 'per person.' An average is a quotient, and a quotient consists of a numerator and a denominator. When we are dealing with real-world quantities, the numerators and denominators are measured in units. The numerator units in the above example are slices of pizza, the denominator units are persons. In order to fully understand an average, one must know what is being shared (the

numerator units -- pizza slices) and among what the shared quantity is being shared (the denominator units -- people). The units of measurement for averages are 'numerator units per denominator unit' -- in this instance, slices per person. All the concepts in this chapter involve quotients.

The Importance of Units

When computing an average, a number 'by itself' is meaningless -- both the numerator and denominator units must be specified. To see how important this is, ask yourself if you would take a job if the salary was 5.

Assuming that the job isn't distasteful or dangerous, you almost certainly would take the job if the salary was $5 per second. You almost certainly wouldn't take the job if the salary was 5 cents per year.

Section 1 - **Averages -- Summarizing the Past, Predicting the Future**

An average is a quotient. Baseball, the national pastime, provides an excellent source for the computation of averages. A player's batting average is the quotient of the number of hits the player has achieved divided by the total number of 'official at-bats' (an official at-bat occurs anytime a player is not automatically awarded first base via a base on balls or being hit by a pitch). If a player has 500 official at-bats, and gets hits in 150 of these, his batting average is $150/500 = .3 = .300$. (A player whose batting average rounded to the nearest thousandth is .273 is said to be hitting two-seventy-three.) A player's batting average is the average number of hits per official at-bat.

Example 1 - After graduating from college, Anne's salary was $30,000 for her first year, $35,000 for each of the next two years, and $42,000 her fourth year. What was her average salary?

Solution: Her total salary for the four years was $30,000 + 2 x $35,000 + $42,000 = $142,000. Her average salary was $142,000/4 = $35,500 per year. ■

Computing averages is not difficult, but it is surprisingly easy to be misled by the form in which the information is presented.

Example 2 - Evan buys $6.00 worth of hamburger at $1.50 a pound, and then goes to another store where he buys another $6.00 worth of hamburger at $1.00 a pound. What is the average price of the hamburger?

Solution: If you computed the answer by saying that, since the same amount was purchased, the average price of the hamburger must be $1.25, the average of the prices $1.50 and $1.00, you have fallen into exactly the same trap as Freddy did during the story!

Many problems involving averages simply require you to keep focused on the numerator and denominator of the quotient which you use to compute the average. In this case, the numerator is $12.00, the amount of money spent, and the denominator is 10 pounds, the amount of hamburger purchased. The average price is therefore $12.00/10 = $1.20 per pound. ▉

Many of the mistakes made in computing averages are variations of the error Freddy made during the story. Suppose that one is given two different data sets and computes an average for each data set, and then computes the average of all the data together. In general, it is *not* true that the average for all the data is the average of the averages for each data set. In the story, Freddy computes the average of two averages rates. In Example 2, the trap is to compute the average of two average prices. The way to avoid this trap is to compute the numerator and denominator for the entire data set -- computing the average of two averages is very likely to lead to a wrong answer.

Example 3 - A student's grade-point average (GPA) is computed as an average of grade-points earned (A=4, B=3, C=2, etc.) per course. Each course has the same number of credit-hours. Maria has a GPA of 3.2, and has taken 5 courses. If she wishes to have an overall GPA of at least 3.3, and

is taking four courses this semester, what GPA must she receive for those four courses? What combination of grades will enable her to reach her goal?

Solution: Let G be the GPA she will receive for those four courses. Then she will receive a total of 4G grade points in those four courses. She has received 5 x 3.2 = 16 grade points in the five courses she has already taken. The total grade points she will receive is therefore 4G + 16 for nine courses. If she wishes to have a GPA of at least 3.3 for those nine courses, she must obtain 9 x 3.3 = 29.7 grade points. Therefore

$$4G + 16 > 29.7$$

$$4G > 13.7$$

$$G > 3.425$$

If she is taking four courses, the only combinations of grades with a GPA > 3.425 are 4 As (GPA = 4.0) , or 3As and a B (GPA = 3.75), or 3As and a C (GPA = 3.5), or 2As and 2Bs (GPA = 3.5). ■

Example 4 - Linda's bowling league is awarding a prize to anyone who averages 200 in a 10-game series. After 6 games, Linda's average is 196. What must she average for the next four games in order to win the prize?

Solution: There are several ways to attack this problem. If Linda averages a score of S in the last four games, then her average for all ten games will be (6 x 196 + 4S)/10. Setting this equal to 200, we get

$$(6 \times 196 + 4S)/10 = 200$$

$$1176 + 4S = 2000$$

$$S = (2000 - 1176)/4 = 206$$

Another way to handle this problem (without algebra) is to realize that a total score of 200 x 10 = 2000. Linda has already scored 6 x 196 = 1176 points, so she needs 2000 - 1176 = 824 points in 4 games, for an average of 206.

Yet a third way to work the problem (again without algebra) is to realize that a score of 196 is 4 below the desired average of 200, so Linda is 6 x 4 = 24 points below the needed pace. She needs to make up those 24 points in 4 games, so she must average 6 points above the desired average of 200 for those 4 games, or 206. ■

Returning to baseball, notice that a batting average can be interpreted either as an average or a percentage. The 'three hundred hitter', who amassed 150 hits in 500 official at-bats, can be said to have gotten hits in 30% of his official at-bats.

Averages are capable of being abused, as the following example shows.

Example 5 - The four executives at Mirage Financial are paid annual salaries of $100,000, and the six assistants get annual salaries of $40,000. The executives got raises of $10,000, and the staff got raises of $2,000. The company told its stockholders that the average raise was 7%, and it told the employees that the average raise was 8.125%. What's going on here?

Solution: Welcome to the wonderful world of creative accounting! There were four raises of 10% and six raises of 5%; the average of these ten percentages is 7%. On the other hand, the initial payroll was $640,000, and after the raises the payroll was $692,000, an increase of 8.125%.

There is some truth in both of these numbers -- they are both correctly figured, but they are differently defined. This example indicates why it is often difficult to find out what is really happening with the finances of a company which has a complicated balance sheet.

Averages are often used as a basis for estimates of future performance. This reliance on averages to estimate future performance exists even where the reasons that resulted in that average are not well understood. If 100 years of weather data shows that the average rainfall in an area is 30 inches per year, planning for water use and distribution is made on the assumption that the future will continue to average 30 inches per year.

Using past averages as future estimates is common practice. A company estimates future sales by looking at past averages. A .300 batting average is often used not only as a summary of the past, but as a predictor of the future (at least, by the hitter's agent when the time comes to negotiate the contract).

Example 6 - An average of 600 million tickets have been sold annually by the motion picture industry during the past five years. Project how many tickets will be sold during the next decade. What factor is most likely to cause this to change?

Solution: The obvious estimate is 10 x 600 million = 6 billion tickets during the next decade. It is likely that, in estimating ticket sales, the industry would adjust this figure by the growth rate of the population. Notice that the growth rate is also likely to be computed as an average! ■

It is important to realize that averages can be used to estimate the future, and plans based on those estimates may often be extremely useful, but averages based on past data cannot predict the future. Even though past rainfall may have averaged 30 inches per year, the coming year may see either a drought or a flood.

Two important branches of mathematics that will be discussed later in this book are probability and statistics. Probabilities are basically future long-term averages. Much of statistics is devoted to a quantitative analysis of how best to use averages as predictors. Mathematical modeling uses probability and statistics, among other disciplines, to enable us to analyze what happens – or will happen – in the real world.

Section 2 - Rates

Rates are quite similar to averages, as both averages and rates are expressed as quotients.

From a mathematical standpoint, there is no difference between an average and a rate. However, there is a sort of unspoken agreement that the word 'average' is generally used to summarize data, whereas the word 'rate' is used to describe an ongoing process. If twelve slices of pizza are eaten by four people, each person eats an average of three slices. We could also say that the rate of pizza consumption is three slices per person. The numerical measures (three) of both average and rate are the same, the units of measurement (slices per person) are the same, so from the mathematical standpoint they are indistinguishable.

One of the most well-known formulas in mathematics is

$$D = R\ T$$

or

distance = rate x time

When the above formula is used in the precise form D = R T, it enables us to calculate how far an object has moved in an amount of time T, provided that it was moving at a *constant* rate R. For instance, if a car is traveling at a constant rate of 40 miles per hour, in three hours it will travel 40 x 3 = 120 miles. When a car travels at a constant rate, the speedometer never changes -- it always reads 40.

However, when we do a little division to write the above equation in the form R = D / T, a wider range of interpretation is possible. Clearly, we can use the above formula to solve the following problem: if a car that travels at a constant rate goes 120 miles in 3 hours, at what rate is it moving? Of course, the answer is 120 miles/ 3 hours = 40 miles per hour.

We can also use the formula R = D / T to construct the idea of an average rate. Let us suppose a car goes 120 miles in 3 hours. Then its average rate is defined to be the total distance travelled, 120 miles, divided by the total time taken, 3 hours. Its average rate is 40 miles per hour. However, we cannot use the average rate thus computed to draw specific conclusions about the distance travelled during any period of the trip. One cannot use the information that a car moves at an average speed of 40 miles per hour for three hours to determine how far the car went during the second hour of the trip. It might have gone forty miles, or sixty miles, or zero miles, or even backward -- we simply do not know.

The same type of situation occurs with averages. When four people share twelve slices of pizza, each person gets an average of three slices, but there is no way to tell exactly how many slices each person received.

The formula D = R T is one of the most used (and abused) formulas in mathematics. It is only to be used if the rate R is either constant or the average rate for the entire time interval to which the problem refers.

The formula R = D/T can either be used to compute an unknown constant rate, or as the definition of an average rate for a trip that covered D distance units (e.g. miles) and took T time units (e.g. hours).

In the story, Freddy made the common mistake of assuming that the average rate is the average of the rates. It certainly sounds convincing, but it is only true when the denominators of both rates are equal. The actual average rate of Freddy's round trip to San Diego can be computed by looking at the total distance traveled (240 miles), and dividing by the total time taken (3 hours to go the 120 miles to San Diego at 40 miles per hour; 6 hours to return at 20 miles per hour). Although the average of 40 and 20 is 30, the average speed on the trip is 240/9 = 26 2/3 miles per hour. As Pete observed, the actual average speed is lower than Freddy's estimate of 30 miles per hour because time, the denominators of the two rates, differed for the two legs of the round trip.

Example 1 - A plane travels 200 miles at 300 miles per hour, and then runs into a headwind and travels 300 miles at 200 miles per hour. What is its average rate?

Solution: This is a problem to which we can apply the formula R = D/T, as long as we are careful to compute D and T accurately. Clearly, D = 200 + 300 = 500 miles. The first part of the trip takes 200/300 = 2/3 of an hour, and the second part of the trip takes 300/200 = 3/2 hours, so the total time taken is 2/3 + 3/2 = 4/6 + 9/6 = 13/6 . The average rate R = 500/(13/6) = 3000/13 = 230 10/13 miles per hour. ▉

The Rate Principle: Exchanging Quantities

Rates enable us to exchange one quantity for another. When we say that apples cost 40 cents per pound, we are specifying a way to exchange one quantity (money) for another (apples). If we are interested in purchasing N items, each of which has a price P, then the total cost C is given by the formula

$$C = N P$$

Example 2 - Hamburgers cost $1.50 each. How much will 6 hamburgers cost?

Solution: This question is of a type that confronts us every day. A simple use of the C = N P formula shows that the total cost is 6 hamburgers x $1.50 per hamburger = $9.00. ▉

In Example 2, the restaurant exchanges 6 hamburgers for $9.00, the rate of exchange being $1.50 per hamburger. Notice that the D = R T formula also involves an exchange -- in this case, an exchange of time for distance. If we drive 3 hours at a constant rate of 40 miles per hour, we are exchanging 3 hours of our time for 3 x 40 = 120 miles of distance.

The rate concept is extremely flexible, and has many different applications. They all have an underlying similarity -- the use of multiplication to exchange one type of quantity for another at a constant rate of exchange.

Example 3 - A survey indicates that commercials are shown at a rate of 15 minutes per hour during sitcoms, and 12 minutes per hour during sports events. If a typical viewer in a certain demographic bracket spends 40% of his time watching sitcoms and the remainder watching sports, at what hourly rate does he watch commercials?

Solution: There are several ways to work this problem. In ten hours of watching, he will see four hours of sitcoms, with 4 x 15 = 60 minutes of commercials, and 6 hours of sports, with 6 x 12 = 72 minutes of commercials. He will therefore watch 60 + 72 = 132 minutes of commercials in ten hours of viewing, a rate of 13.2 minutes per hour.

An alternate way is to realize that a typical hour for this viewer is made up of 40% sitcoms and 60% sports. Therefore, the number of minutes of commercials seen during this hour would be .4 x 15 + .6 x 12 = 6 + 7.2 = 13.2 . ■

Rates occur frequently in the financial world. We shall devote an entire chapter to a study of interest rates and their effects. Currency exchange rates are another important example of the rate concept in daily life.

Example 4 - On a trip to London, Sally exchanged $200 for pounds at a rate of $2.00 per pound. When she woke up the next day, the pound had been devalued. She exchanged $350 more, and found that she now had a total of 300 pounds. What is the new rate of exchange? By what percentage had the pound been devalued?

Solution: When Sally exchanged $200 at the $2.00 rate, she obtained 100 pounds. Therefore, she must have received 300 - 100 = 200 pounds for her $350. The new rate of exchange is therefore 350/200 = $1.75 dollars per pound. The pound has lost $.25 from its $2.00 initial value, which is 100 x ($.25/$2.00) = 12 1/2%. ■

Converting Different Rates

We are often interested in converting rates. If we are staying in England and want to buy a sweater which is selling for twenty pounds, we want to find out what that sweater costs in dollars to see whether it is a good buy. In order to do this, we must know the currency exchange rate of dollars per pound. If the exchange rate is $1.75 per pound, we know that the sweater will cost 20 x $1.75 = $35. What we have actually done is to convert one rate (pounds per sweater) to another rate (dollars per sweater) by knowing a third rate (dollars per pound). If we write the units in which the rates are being measured, such as dollars per pound, as fractions (dollars/pound), we could write the problem as

20 pounds/sweater x 1.75 dollars/pound = 35 dollars/sweater

Notice that we have 'multiplied the units' by canceling a common factor (pounds) that appears in the numerator and denominator of the two fractions.

pounds/sweater x dollars/pound = dollars/sweater

This multiplication process is always followed when one converts rates.

Example 5 - If there are 8 Danish kroner to the dollar and $1.75 to a pound, what is the conversion rate of kroner per pound?

Solution: We are guided by kroner/pound = kroner/dollar x dollars/pound. (Notice that we inserted canceling dollars in both numerator and denominator.) Therefore

8 kroner/dollar x 1.75 dollar/pound = 14 kroner/pound ■

Example 6 - A hamburger recipe calls for 6 ounces of ground beef per hamburger. What is the rate of ground beef required for the hamburgers expressed in pounds per dozen?

Solution: Recall that there are 16 ounces in a pound, so the rate of pounds per ounce is 1/16. Since there are 12 hamburgers in a dozen, the rate of hamburgers per dozen is 12. Therefore

6 ounces/hamburger x 12 hamburgers / dozen x 1/16 pounds/ounce = 4 1/2 pounds/dozen

Notice that 'ounces' and 'hamburgers' appear in both numerator and denominator of fractions, and therefore cancel. ■

Reciprocal Rates

Let's take one last look at the example in which four people eat 12 slices of pizza. If the people were distributed evenly among the pizza slices, each slice would have been eaten by 4/12 = 1/3 of a person. Just as we measure the average number of slices per person by division of the total number of slices by the total number of people, we measure the average number of people per slice by division of the total number of people by the total number of slices. Notice that both these rates (three slices per person, 1/3 of a person per slice) convey the same information, and that the number three and 1/3 are reciprocals of one another. This reciprocity will always occur, because $a/b = 1 / (b/a)$.

Many difficult story problems which involve rates are difficult only because of the failure to perceive the problem in terms of reciprocal rates.

Consider the following problem. It takes a large pipe four hours to fill a swimming pool, and it takes a small pipe twelve hours to fill the same pool. How long does it take both pipes, working together, to fill the same pool? Familiarity with the reciprocal rate concept makes short work of this problem. In a single hour, the large pipe fills 1/4 of the pool, and the small pipe 1/12 of the pool, so together they fill 1/4 + 1/12 = 3/12 +1/12 = 4/12 = 1/3 of the pool. Therefore, it will take three hours for both pipes to fill the pool.

Example 7 - It takes Sue eight hours to paint a room. If she and Jack paint the same room, it only takes five hours. How long would it take Jack to paint the room?

Solution: Problems such as this have given algebra the reputation of being more terrifying than a visit to the dentist! However, it is not a difficult problem when we look at it in terms of reciprocal rates. If Jack takes H hours to paint the room, he would paint 1/H of the room in an hour. Sue paints 1/8 of the room in an hour, and together they paint 1/5 of the room in an hour. Therefore

$$1/H + 1/8 = 1/5$$

$$1/H = 1/5 - 1/8 = 3/40$$

It would take Jack 40/3 = 13 1/3 hours to paint the room. ∎

If we look at a well-known rate formula such as D = R T we can deduce two other formulas: R = D / T and T = D / R. The first of these formulas, R = D / T, is essentially the definition of what an average rate is. If we drive 60 miles in 2 hours, our average rate is 60/2 = 30 miles per hour.

The second formula, T = D / R, is not as easily recognized; it is a little easier if we write it as T = D x 1 / R . 1/R is the reciprocal rate: if R is the speed of the moving object in miles per hour, 1/R represents the number of hours it takes to travel one mile.

All rate formulas basically convey the same type of information -- only the quantities being measured change. Consider, for instance, the parallels between moving objects and eating slices of pizza.

Moving Objects	Eating Slices of Pizza
D = distance travelled	S = number of slices eaten
T = time taken	N = number of people
R = average speed	R = average # slices eaten
Totals D = RT	Totals S = RN
Rates R = D/T	Rates R = S/N
Reciprocal Rates T = 1/R x D	Reciprocal Rates N = 1/R x S

We conclude this section by looking at the swimming-pool problem as if it were a problem in eating slices of pizza.

Example 8 - If John takes 4 days to eat a large pizza, and Susan takes 12 days to eat a large pizza, how long will it take them both to eat a large pizza?

Solution: John eats 1/4 of the pizza a day and Susan eats 1/12 of the pizza a day, so together they eat 1/4 + 1/12 = 1/3 of the pizza a day. Therefore, it takes them 3 days to eat the pizza. ■

It's the same problem, with the same numbers as the swimming-pool problem that we saw a few paragraphs back. Not only does it use the same numbers, it uses the same ideas.

Section 3 - **Ratio and Proportion**

An average or rate is computed in the real world by measuring two quantities. If two people eat six slices of pizza, we clearly counted the numbers of slices of pizza and the number of people. A ratio is a fraction expressed in whole-number form. The ratio of people to slices of pizza, 2 to 6, is written 2:6. We interpret this as saying that to every 6 slices of pizza, there are 2 people. Since the ratio 2:6 corresponds to the fraction 2/6 (the number of people per slice of pizza), and since 2/6 = 1/3, the ratio 2:6 is the same as the ratio 1:3.

Unlike fractions, which are limited to a comparison of two quantities, the idea of ratios can be extended to any number of quantities.

Example 1 - A recipe for cheese omelettes requires 6 eggs, 2 spoonsful of cream, and 4 ounces of cheddar cheese. It serves 3 people. The ratio of people to eggs to spoonsful of cream to ounces of cheddar cheese is written 3:6:2:4.

A **proportion** is simply an equality between ratios. Proportions can be used to find missing information by equating fractions.

Example 2 - The ratio of Democrats to Republicans in Lincoln County is 11 to 9. If there are 12,000 registered voters in Lincoln County, how many are Democrats?

Solution: Let D be the number of Democrats. Then 12,000 - D is the number of Republicans. Therefore, the two ratios 11:9 and D:12,000 - D must be the same. We can therefore equate the fraction form of these ratios by setting

$$11/9 = D/(12,000 - D)$$

$$11 \times (12,000 - D) = 9 D$$

$$132,000 = 20 D$$

So there are 6,600 Democrats and 5,400 Republicans (note that 6,600/5,400 = 11/9). ■

A common use of proportions occurs when recipes in cookbooks are modified to accommodate a different number of servings than the one specified in the recipe.

Example 3 - Suppose that 12 people wish to enjoy the cheese omelettes whose recipe is given in Example 1. How much of each item is necessary?

Solution: The straightforward way to use proportions is to set up a specific equation for each quantity. If E is the number of eggs needed by the 12 people, then 3:12 must be the same ratio as 6:E. Therefore 3/12 = 6/E, and it is easy to see that E = 24.

The shortcut is to realize that, if R represents the quantity of a particular item specified in the recipe, and N the amount needed for 12 people, then 3:12 must be the same ratio as R:N. So 3/12 = R/N. Cross-multiplying, 3 N = 12 R, and so N = 4 R. In other words, just multiply all the recipe quantities by 4. So 24 eggs, 8 spoonsful of cream, and 16 ounces of cheddar cheese will be needed. ■

In Example 3, the quantity 4 by which all ingredients in the recipe must be multiplied is called a **scale factor**. A common use of scale factors can be seen in maps.

The scale factor in a map is usually indicated in two different ways. A notation such as

1 inch:10 miles

means that each inch on the map represents 10 miles in the real world. The above ratio may also be expressed as 1:633,600, and indicates that map distances must be multiplied by 633,600 to obtain real-world distances. Notice that there are 12 inches in a foot, and 5,280 feet in a mile, so there are 10 x 5,280 x 12 = 633,600 inches in 10 miles.

Proportions can be used to solve somewhat more complicated problems. Suppose, for instance, that three painters take two days to paint eight rooms. If all the painters work at the same rate, how many days will it take seven painters to paint twenty-eight rooms?

There is a fairly straightforward way to attack this problem. Three painters took two days to paint eight rooms, so they would have taken one day to paint four rooms, or seven days to paint twenty-eight rooms. It would therefore take one painter twenty-one days to paint twenty-eight rooms, and so seven painters could do it in one-seventh of that time, or 3 days.

Another way to do this is to introduce the concept of a painter-day, and use proportions. If three painters take two days, 3 painters x 2 days = 6 painter-days are required. If we let X denote the number of painter-days required to paint twenty-eight rooms, the ratios of painter-days to rooms must be the same, and so 6:8 must be the same ratio as X:28. Therefore 6/8 = X/28, which has X = 21 as the solution. Since we have seven painters, each must paint for 21/7 = 3 days.

A painter-day is an example of what might be called a **product unit**, a unit which arises through the multiplication of two different units. Although the term 'product unit' is not in common use, it is a worthwhile concept to keep in mind, as some problems can naturally be formulated in terms of product units. Product units can often be found in everyday life. For example, when an airplane carries 40 passengers on a 500 mile trip, the airplane is said to have flown 40 passengers x 500 miles = 20,000 passenger-miles. We sometimes write 40 x 500 = 20,000 passenger-miles if the units that the numbers 40 and 500 describe (passengers and miles respectively) are well-understood within the context of the problem. Mathematicians love to use abbreviations; it saves both time and space, but is one of the reasons that mathematics is sometimes difficult to read.

Example 4 - If a hen and a half lays an egg and a half in a day and a half, how many eggs will six hens lay in nine days?

Solution: This oldie-but-goodie is at least two centuries old, and in the intervening two centuries, many people have fallen into the trap of concluding that a single hen lays an egg a day (you, of course, know better!).

One approach is to use multiplication. Multiplying hens and eggs by four shows us that six hens lay six eggs in a day and a half, so, since 9 = 6 x 1 1/2, the six hens will lay 6 x 6 = 36 eggs in nine days.

To use a proportion, we must realize that it requires 1 1/2 x 1 1/2 = 2.25 hen-days to produce

1 1/2 eggs, so if E eggs are produced by 6 x 9 = 54 hen-days, we must have 2.25/1.5 = 54/E, and so E = 36. ■

Averages, rates, and ratios involve the same information. Every statement about one is a statement about the others as well, as they are all concepts involving the interpretation of quotients. It helps also to note what the units are, as it helps to keep the fractions rightside-up!

Chapter 4 - Sequences and Arithmetic Progressions

The ability to recognize and use patterns is one of the most important aspects of intelligence. It is only in the last ten thousand or so years that man has created societies and the advances that go with them. Undoubtedly, these advances started when man first started to use the pattern of recurring seasons to develop agriculture.

Predictions based on patterns form the foundation of science and technology. Psychologists work to decipher patterns of behavior on which to predict how we will react in certain situations, while medical researchers study the pattern of the beating heart in the hopes of isolating cues which will alert them to incipient heart attacks. Without knowledge based on predictions, life would be reduced to approximately the level of animal behavior.

Numbers provide an environment which makes it possible to recognize patterns, and these number patterns reflect phenomena that occur in the real world. In this chapter, we will study some of the simpler number patterns, where they occur in the real world, and how they can be used.

Section 1 - **Sequences**

A **sequence** is an unending string of numbers. Some well-known examples of sequences are:

The counting numbers -- 1, 2, 3, 4, ... (the dots indicate that the numbers continue)

The odd numbers -- 1, 3, 5, 7, ...

The prime numbers -- 2, 3, 5, 7, ... (a prime number has only two whole-number divisors, itself and 1. 13 is prime because its only whole-number divisors are 13 and 1, 12 is not prime because it has 2, 3, 4, and 6 as whole-number divisors)

The numbers that make up a sequence are called its **terms**. In the sequence 1, 3, 5, 7, ... of odd numbers, 1 is called the first term of the sequence, 3 the second term, 5 the third term, etc.

Letters can be used in algebra when we wish to talk about numbers and their properties without specifying a particular number. For example, we use the letter x in the equation $2 x + 6 = 2 (x + 3)$ to denote any number. When we wish to talk about sequences or properties of sequences without specifying a particular sequence, we use the notation a_1, a_2, a_3, ... to denote a sequence. The numbers 1, 2, 3, ... in the above notation are called **subscripts**. For instance, a_7, which is read "a sub 7," is the seventh term of the sequence. When we wish to talk about a term of the sequence without specifying a particular term, we use the notation a_n.

There are several different ways to describe sequences. The sequence 1, 3, 5, 7, ... of odd numbers can be described

(1) by using words: the n^{th} term of the sequence is the n^{th} odd number.

(2) by using a formula: $a_n = 2 n - 1$. Notice that using a formula enables us to compute the exact value of <u>any</u> term in the sequence. For instance, $a_{100} = 2 \times 100 - 1 = 199$.

(3) by using a **recursive definition**, which describes a sequence in roughly the same way that one would give instructions on how to use a ladder: put your foot on the first step, and whenever you are standing on a rung, put one foot on the next higher step and bring your other foot up to join it. In this instance, the recursive definition would be $a_1 = 1$, $a_n = a_{n-1} + 2$.

We start with the definition that $a_1 = 1$. When we let n=2, the recursive formula becomes $a_2 = a_{2-1} + 2 = a_1 + 2 = 1 + 2 = 3$, so $a_2 = 3$. Now we can let n=3, and the recursive formula becomes $a_3 = a_{3-1} + 2 = a_2 + 2 = 3 + 2 = 5$. We can now use a_3 to help us compute a_4, etc.

Of the three ways of describing a sequence, describing by means of a formula is the most useful, because it enables us to compute directly any term in the sequence. The formula $a_n = 2 n - 1$ can be used to compute the one hundredth term in the sequence as we did above. If we wanted to find the

one hundredth term using words as the description, we would have to write out the first one hundred odd numbers, and if we wanted to find the one hundredth term by using the recursive definition, we would have to use it ninety-nine times (the first use gave us a_2, the second use gave us a_3, etc.)

Unfortunately, sometimes a formula is not available. In the case of the sequence of prime numbers, there is no known formula or recursive definition which enables us to compute the n^{th} prime number. As a matter of fact, mathematicians have actually been able to prove that it is impossible to find such a formula, or recursive definition!

Other examples of sequences that must be defined by words are the daily balances in your checking account, or the rolls of a random die.

Some Well-Known Sequences

Sequence 1: The even numbers -- 2, 4, 6, 8, ...

Formula -- $a_n = 2n$

Recursive definition -- $a_1 = 2$, $a_n = a_{n-1} + 2$

Sequence 2: The squares -- 1, 4, 9, 16, ...

Formula -- $a_n = n^2$

Recursive definition -- $a_1 = 1$, $a_n = a_{n-1} + 2n - 1$

Sequence 3: The triangular numbers -- 1, 3, 6, 10, ...

Formula -- $a_n = n(n+1)/2$

Recursive definition -- $a_1 = 1$, $a_n = a_{n-1} + n$

One of the parts of a standard intelligence test is to predict the next number in the sequence. This is a case where too much knowledge is a dangerous thing, at least from the standpoint of scoring well on the intelligence test! For instance, suppose you were asked to predict the next number in a sequence whose first three terms were 1, 2, and 4. If you decide that the sequence is recursively defined by $a_1 = 1$, $a_n = 2\ a_{n-1}$, then the next term of the sequence would be 8. If you decide instead that the sequence is recursively defined by $a_1 = 1$, $a_n = a_{n-1} + n - 1$, then the next term of the sequence is 7!

Sequences can consist of terms other than numbers. This book, for instance, is a sequence of letters, numbers, and symbols (with all the terms after a million or so being blank).

Arithmetic Sequences

Consider the sequence 2, 5, 8, 11, ... , which we define recursively by $a_1 = 2$, $a_n = a_{n-1} + 3$. Notice that the difference between any two consecutive terms is 3, which we can see simply by taking the recursion formula $a_n = a_{n-1} + 3$, and subtracting a_{n-1} from each side to obtain $a_n - a_{n-1} = 3$. This is an example of an **arithmetic sequence**, more commonly called an **arithmetic progression**, which is a sequence in which the difference between any two consecutive terms is a constant. This difference is called the **common difference**. In an arithmetic sequence, one can find the common difference by subtracting any term from the term immediately following it. In the sequence 2, 5, 8, 11, ... , $3 = 5 - 2 = 8 - 5 = 11 - 8$, etc.

If we are given the first term 2 and the common difference 3 of an arithmetic progression, we can immediately write down the recursive definition: $a_1 = 2$, $a_n = a_{n-1} + 3$. We can also use the pattern to write down the first few terms of the sequence

$a_1 = 2 \qquad\qquad = 2 + 0 \times 3$

$a_2 = 2 + 3 = 5 \qquad = 2 + 1 \times 3$

$a_3 = (2 + 3) + 3 = 8 \qquad = 2 + 2 \times 3$

$a_4 = (2 + 3 + 3) + 3 = 11 = 2 + 3 \times 3$

From this pattern, we conclude that

$a_n = 2 + (n - 1) \times 3$

Since we could have done exactly the same thing if we had an arithmetic progression with first term a_1 and common difference d, we have the following two formulas for arithmetic progressions.

Formulas for Arithmetic Progressions

An arithmetic progression with first term f and common difference d can be defined recursively by

$a_1 = f, a_n = a_{n-1} + d$

or by means of the formula

$a_n = f + (n - 1) d$

Example 1 - Find the common difference and the 250th term of the arithmetic progression whose first three terms are -6, 1, 8.

Solution: Since 8 - 1 = 1 - -6 = 7, the common difference is 7. The nth term is given by the formula

$a_n = -6 + 7(n-1)$

So the 250th term is $a_{250} = -6 + 7 \times 249 = 1737$. ■

Arithmetic progressions frequently occur in daily life in the total amount paid by a consumer making installment payments.

Example 2 - Susan makes a down payment on a car of $2000, and monthly payments of $180. Describe her equity (the total of the payments she has made) as an arithmetic progression. If she must make 48 monthly payments, how much will she have paid to buy the car?

Solution: The total amount Susan paid is an arithmetic progression, with a_1 = $2000 and common difference $180. Therefore a_n = $2000 + $180(n-1). After she has made 48 payments, the total paid will be a_{49} = $10,640. ■

Nature, too, has many examples of arithmetic progressions at her disposal. Many natural phenomena, such as the intervals between eclipses, or closest approaches of planets to the sun (which can be used to mark the changing seasons), represent examples of arithmetic progressions.

Example 3 - The great American author Mark Twain (the same one who found it easy to give up smoking!) died in 1910, a year which marked the 4th official sighting of Halley's Comet, which appears every 76 years. When was Halley's Comet first officially sighted? When did it last appear? When will it next appear?

Solution: We know that $a_4 = 1910 = f + 3 \times 76 = f + 228$, so $f = a_1 = 1682$. (The comet had reappeared many times before then, but this was the year that Halley first recognized it and predicted when it would next reappear.) The last sighting was $a_5 = 1682 + 4 \times 76 = 1986$, and its next appearance will be in $a_6 = 1682 + 5 \times 76 = 2062$.

Incidentally, not only did Mark Twain die during a_4, but he was born during a_3! ■

Section 2 - Sums of Arithmetic Progressions

The basic use of multiplication is for repeated addition of the same number -- 3 x 4 is a shorthand for 4 + 4 + 4. However, there are certain attractive situations where there are formulas in which multiplication can be used to find the sum of different numbers. An example of such a situation was discovered by one of the greatest mathematicians of all time, Carl Friedrich Gauss.

One day while Gauss was attending school, his teacher excused himself from the classroom for a short period, and asked the students to add the numbers from 1 through 100 in his absence. Most people, when confronted with this problem, proceed in the obvious fashion. Faced with finding the sum 1 + 2 + 3 + ... + 100, they compute the sum of 1 + 2, getting 3. To this result they add 3, getting 6. To 6, they then add 4, getting 10. And so on.

Gauss, however, took a different approach. Letting S denote the sum 1 + 2 + 3 + ... + 100, he noticed that S could also be written 100 + 99 + 98 + ... + 1. Writing these two expressions under each other, we have

(1) S = 1 + 2 + ... + 99 + 100

(2) S = 100 + 99 + ... + 2 + 1

Adding both these equations gives

S + S = (1 + 100) + (2 + 99) + ... + (99 + 2) + (100 + 1)

The first number in each parenthesis comes from (1), and the second number from (2). But each sum in parenthesis adds up to 101, and there are obviously 100 such pairs. So

2S = S+S = 100 x 101 = 10,100

and therefore S = 5,050.

Gauss was about eight years old when he discovered this, which mathematicians in his honor refer to as The Gauss Trick. It would be the highlight of many a mathematician's career to come up with as cute a trick.

There is obviously nothing special about the number 100. It could have been 1000, or eight million, or anything at all. Let's suppose, therefore, that we want to add all the numbers from 1

through N. Letting S denote the sum 1 + 2 + ... + N, we write down the sum forwards and backwards.

(3) $S = 1 + 2 + ... + N-1 + N$

(4) $S = N + N-1 + ... + 2 + 1$

Adding (C) and (D), we get

$2S = S + S = (1 + N) + (2 + (N-1)) + ... + ((N-1)+2) + (N + 1)$

As before, the first term in each parenthesis comes from (3), and the second term from (4). Each parenthetical sum is N+1, and there are obviously N pairs of parentheses. Therefore,

$$2S = S + S = N \times (N+1)$$

and so $S = [N \times (N+1)]/2$.

Suppose now that we wish to add the terms in an arithmetic sequence in which the first term is a, the common difference is d, and the last term is a + nd. Letting S denote the sum, and doing a little rearranging, we have

$$S = a + (a + d) + (a + 2d) + ... + (a + nd)$$

$$= (a + a + ... + a) + d (1 + 2 + ... + n)$$

$$= (n+1)a + d (n \times (n+1)/2)$$

$$= (n+1) \times (a + nd / 2)$$

Example 1 - What is the sum of the terms in an arithmetic series whose first term is 8, whose common difference is 3, and whose last term is 377?

Solution: We can simply apply the above formula if we only knew the value of n. However, the last term must be 8 + 3n = 377, so 3n = 369 and n=123. So S = 124 x (8 + 3x123/2) = 124 x 192.5 = 23,870. ▉

The numbers $T_n = 1 + 2 + ... + n = n \times (n+1)/2$ are sometimes called **triangular numbers**, as can be seen from Fig. 4-1.

Triangular Numbers

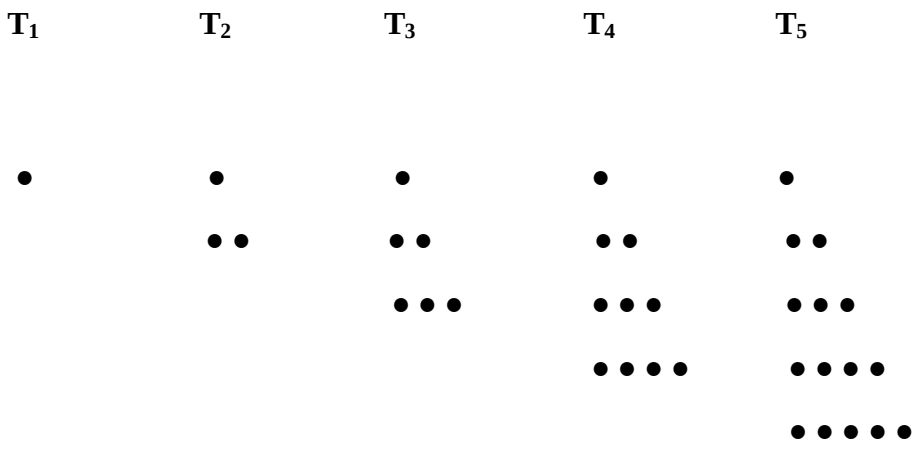


Fig. 4-1

Even in an age of ultra-high speed computers, it is still extremely important to discover short formulas for sums. After all, why should one make even a high-speed computer add a thousand numbers, when the above formula only requires a very few additions and multiplications? Sophisticated problems often require billions or even trillions of computations, and short cuts such as the above formula can sometimes eliminate more than 99% of the calculations.

Example 2 - In the story, Freddy wished to cut down his cigarettes by 1 a day. How many cigarettes would he have had to buy if he had wanted to cut down by 2 a day?

Solution: We have to find the sum of an arithmetic progression whose first term is 40 and whose common difference is -2. Since the last term is 2, $2 = 40 + n \times (-2)$, $n = 19$. Substituting in the formula, we see that

$$S = 20 \times (40 + 19 \times (-2) / 2) = 20 \times 21 = 420$$

So Freddy would have to buy 2 cartons and 1 pack (a carton contains 10 packs, a pack contains 20 cigarettes). We could also have computed this result by realizing that $40 + 38 + ... + 2 =$

$2 (20 + 19 + ... + 1) = 2 \times ((20 \times 21)/2) = 420$. ■

Chapter 5 – Algebra; The Language of Quantitative Relationships

Introduction

This material focuses primarily on the portion of algebra that causes the most trouble for students – setting up and solving standard story problems. If you know an algebra student who is floundering in this area, they would benefit from reading it. Or, if you are a parent whose child is floundering in this area, read it yourself and impress your child with your skills in this area. If your son or daughter is taking algebra, he or she is almost certainly a teen-ager, and this will be your last chance to impress him or her for another decade or so.

Algebra is as much a language as it is a subject. Unfortunately, the way algebra is taught in high school makes it very difficult to see this. In high school algebra one learns how to simplify algebraic expressions, how to factor polynomials, how to solve quadratic equations, and so on and so on ad infinitum (and, unfortunately, often ad nauseam). What is learned is an assortment of techniques.

Most people who use algebra frequently regard it as a language which describes relationships between quantities. One of the more important aspects of a language is that it enables questions to be raised and answered. In this sense, arithmetic is a language too -- the questions that it raises concern the relationships among numbers. An example of an arithmetic question is "What is two plus two?" It has only one correct answer, and is a question about numbers. As we have mentioned, questions about numbers are the concern of arithmetic.

A different type of question is "What time is it?" This question has many different answers – 2 P.M., midnight, 9:35:14, etc. Although only one answer is correct at any given moment, there are a large number of potentially correct answers.

Time is an example of a **variable**, a quantity which may assume different values. Variables are not restricted to numerical values -- a driver's license provides an example of a large number of different variables, such as birth date, hair color, and address.

Algebra is a language which provides a framework for raising and answering questions which have many different potentially correct answers. The techniques of algebra mentioned in the first paragraph (factoring polynomials, etc.) are important, but hopefully this chapter will enable you to see why algebra is a language of quantitative description.

Section 1 - Variables and Functions

As we discussed in the Introduction, a variable is a quantity which may assume different values. The different values which the variable may assume comprise the **domain** of the variable. It is customary to use letters to denote variables. Although there is a tradition of letting x denote a variable, it is often more helpful to use a mnemonically significant letter as the variable. Thus, if one were talking about time as a variable, it would be preferable to use the letter 't'.

Example 1 - Let S denote a state of the United States. S is a variable, and its domain consists of the fifty possible values Alabama, Alaska, ... , Wyoming.

While a variable often will have a natural domain, such as the state variable S in Example 1, one can specify the domain of a variable more exactly if desired. For instance, one could talk about the variable S, whose domain is all states of the United States east of the Mississippi River.

There are branches of mathematics which deal with variables whose domains are quite general, such as the state variable in Example 1. Algebra, however, deals with variables whose domains are sets of numbers. From this point on, we shall always assume that the domain of any variable is a set of numbers. As a result, expressions such as x + 3 ('add 3 to x') or 2x ('multiply 2 times x') make sense, because we are automatically assuming that x is a number.

Functions

Probably the most famous formula in the world is Einstein's *tour de force* from the theory of relativity.

$$E = m c^2$$

The number of people who have heard of this formula, and can even repeat it ("E equals em cee squared") doubtless greatly exceeds the number of people who know what it means. It is a relationship between two variables: m, the mass of the object, and E, the energy obtained when an object of mass m is completely converted to energy. (The letter c is actually a specific constant, like π, and is equal to the speed of light in a vacuum.)

Einstein's formula is an example of a **function**, which is a specific type of a relationship between two variables. What makes Einstein's formula so memorable is that, in addition to succinctness, it is always true -- when we convert a 1 gram mass to energy, we always get the same amount of energy, whether we do it on Tuesday or Saturday, whether we do it on Earth or on the moon, or whether we do it or Nature does it. As long as the converted mass has a mass of 1 gram, the energy produced is always the same.

In Einstein's formula, the variable m is known as the **independent** variable, and E the **dependent** variable. We choose a value for the independent variable, and then use the expression for the dependent variable to compute the corresponding value for it. Functions are characterized by the fact that using the same value for the independent variable always results in the same value for the dependent variable.

Example 2 - In the formula $y = 2 x + 3$, x is the independent variable, and y the dependent variable. If we are given the value 5 for the independent variable x, the corresponding value for the dependent variable y is $2 \times 5 + 3 = 13$.

A function can be thought of intuitively as a process involving an input (the independent variable) and an output (the dependent variable). The process is reliable (the same input always results in the same output) because of the requirement on a function that using the same value of the independent variable always produces the same value for the dependent variable.

In Example 2, the dependent variable y was defined in terms of the independent variable x by the formula $y = 2x + 3$. An alternate notation which is quite useful is

$$f(x) = 2x + 3$$

which is read "f of x equals 2 x plus 3." This does not mean that we multiply f by x; if we wanted to do this we could write f*x or fx. In this notation, the symbol f denotes the function – in this case, the act of multiplying the independent variable by 2 and adding 3. The letter x indicates the independent variable. In Example 2, we saw that letting the value of the independent variable be 5 resulted in a value of 13 for the dependent variable. Using our new notation, this would be denoted by the expression

$$f(5) = 13$$

Suppose we were to define $f(u) = 2u + 3$. Even though the label for the independent variable has changed (from x to u), what the function does has not changed -- it still takes the value of the independent variable, multiplies it by 2, and then adds 3. Two functions with the same domain are equal if they produce the same results when the values of the independent variables are the same. Thus, if

$$f(x) = 2x + 3$$

and

$$g(u) = u + u + 3$$

it is clear that, if the same value is used for both x and u, the resulting values of f and g are the same. When two functions are equal, we write f = g.

Functions are one of the most important concepts in mathematics, because they involve predictability -- one can predict (calculate) the value of the function, knowing the value of the independent variable. From predictability it is only a short step to utility: using the function to calculate the value of the independent variable needed to produce a desired value of the function.

Example 3 - The function $A(s) = s^2$ gives the area of a square whose side is s. Use this formula to determine what must be the side of a square whose area is to be 9 square feet.

Solution: We wish to find a value s of the independent variable such that A(s) = 9. Therefore, we must solve the equation

$$s^2 = 9$$

The solution to this equation is s = 3, so the side of the square must be 3 feet. Yes, s = -3 is also a solution to the equation, but -3 feet is a meaningless expression. ■

Example 3 may appear quite simple, but it saves us an awful lot of work. We don't have to construct squares with different lengths of sides, hoping to stumble upon one whose area is 9 square feet.

It is certainly possible to have functions of more than one variable. One such example is the function describing the area of a rectangle A in terms of its base b and height h. This function is

$$A(b,h) = bh$$

To denote the area of a rectangle whose base is 6 feet and whose height is 3 feet, we use the notation A(6,3) = 18.

Example 4 - Write a function which gives the total cost C of buying H hamburgers and F orders of french fries, if hamburgers cost $2 and an order of fries costs $0.75. Use function notation to describe the cost of 3 hamburgers and 4 orders of french fries.

Solution: Since the cost of H hamburgers will be 2 dollars per hamburger x H hamburgers = 2H dollars, and the cost of F orders of french fries will similarly be 0.75F dollars, the required function is C(H,F) = 2H + 0.75F, where C is measured in dollars. The cost of 3 hamburgers and 4 orders of fries is C(3,4) = 3 hamburgers x 2 dollars per hamburger + 4 orders of French fries x $0.75 per order of French fries = $9. Again, when the units are well understood from context, it saves time and space leave them out until the very end, and write C(3,4) = 3 x 2 + 4 x 0.75 = $9. ▉

It is often convenient to use functions to describe real-world problems and situations.

Example 5 - Suppose that the cost of hamburgers and french fries are as in Example 4. Use the cost function C(H,F) to describe

(a) the various hamburger-french fry combinations that can be bought for exactly $5

(b) the change remaining from $10 when H hamburgers and F orders of french fries are purchased

Solution: (a) This is just the set of all pairs of whole numbers H and F such that C(H,F) = 5 (we have to use whole numbers because you can't buy a fraction of a hamburger)

(b) 10 - C(H,F) ▉

Finding functions which describe the behavior of fundamental quantities is an important goal of science. As a simple but significant example, currently a great deal of effort is being expended to find a function which describes the beating of the heart, in the hopes of being able to use the function to predict when a heart attack will occur.

Section 2 - Linear Functions

The simplest type of function is a constant function, such as

$$f(x) = 3$$

No matter what we use as a value for the independent variable, the corresponding value of the function is always 3. Constant functions, as their name implies, do not depend on the variable -- they are constant, and do not vary.

A linear function is one such as

$$f(x) = 3x - 5 = 3x + (-5)$$

All linear functions have the form $f(x) = a x + b$, where a and b are constants. The constants a and b remain the same no matter how x changes. It is called a linear function because a picture of this function (called a graph) is a straight line.

Let's look again at the function Pete used in the story to compute Freddy's car-rental bill.

$$f(x) = 60 + 0.15x$$

This function is a linear function as described above, with $a = 60$ and $b = 0.15$. In the story, the number 0.15 referred to the 15 cents a mile that Freddy was charged for driving the car. Obviously, the 15 cents is a rate; the rate per mile for driving the car. If we were not aware of the origin of Freddy's function $f(x) = 60 + 0.15x$, and merely regarded it as a mathematical object (rather than a total cost), the number 0.15 still has significance -- it is the average rate of change of the function f(x) in "function units" per "variable unit". For instance, if we compute the values f(50) and f(100), we see that $f(50) = 27.5$ and $f(100) = 35$. The variable changes by $100 - 50 = 50$ "variable units" (miles, in the story) and the function changes by $35 - 27.5 = 7.5$ "function units" (dollars, in the story). The average rate of change would be $7.5/50 = 0.15$ "function units" per "variable unit" (dollars per mile, in the story). Here we used 50 and 100 to compute our average rate of change, but if we had used 127 and 319 instead, the average rate of change would still have been 0.15 .

Mathematical Modeling

A mathematical model is a description of a real-world process using the language and techniques of mathematics. The predictive power of mathematics makes mathematical models highly desirable. Once a model has been constructed, one can make predictions from the model, and then test them against what actually happens in the real world. If the model provides accurate predictions of the process, it can be used not only to describe what is happening and what will happen, but also to improve the understanding of the process.

The Italian mathematician and astronomer Galileo showed, through careful experiments, that the distance that a falling object would drop in t seconds could be described fairly well by the function

$$s = 16\,t^2$$

where s is measured in feet. At approximately the same time, the German astronomer Kepler showed that the orbits of the planets around the sun were ellipses. These models led Isaac Newton to construct his Theory of Universal Gravitation, a mathematical model which provided a rationale for the phenomena that Galileo and Kepler had observed.

Although the physical sciences provide the most striking examples of the power of mathematical models, the value of these models has been proven in many other areas of human activity, from economics to psychology. In fact, as we shall see in later chapters, the desire to model a real-world process mathematically has often motivated the development of new branches of mathematics.

Functions in general, and linear functions in particular, are an important modeling tool. A linear function is almost always presumed as a "first approximation" to the behavior of a function whose actual behavior is not known; it can be shown in calculus that linear functions represent a type of "best guess" to the behavior of a function.

There are many other interesting and important functions besides linear functions. Such functions are called non-linear. One well-known example of a non-linear function is the growth of the

population of the world as a function of time. Life not only exhibits non-linear growth, it exists because of non-linear phenomena.

In particular, plants are green because chlorophyll absorbs some of the colors in sunlight (such as blue and yellow), but reflects the green. When chlorophyll absorbs light, it uses the energy of the light to convert carbon dioxide and water to carbohydrates, some of which we eat. If the amount of energy converted by chlorophyll were a linear function of light color, because green lies between yellow and blue, the amount of energy converted from green light would lie somewhere between the amount of energy converted from blue light and the amount of energy converted from yellow light. However, it doesn't – green light just bounces off chlorophyll molecules and into our eyes – which is why plants are green.

Section 3 - Solution of Systems of Linear Equations

After attending the movies on Wednesday, everyone adjourns to the local pizza palace for pizzas and pitchers. Last Wednesday they had five pizzas and three pitchers, and the bill came to $34. This Wednesday they had six pizzas and four pitchers, for a total of $42. What is the cost of a pizza? What is the cost of a pitcher?

This problem is different from the ones we have encountered earlier, because we are seeking not one, but two items of information. Suppose we let P denote the cost of a pizza, and let B denote the cost of a pitcher (the reader is invited to guess as to why we have chosen this letter). Last Wednesday the five pizzas cost 5P and the three pitchers cost 3B, so the total of $34 results in the equation

(1) $5P + 3B = 34$

Similarly, the six pizzas cost 6P, the four pitchers cost 4B, and since the bill for this came to $42, we obtain the equation

(2) $6P + 4B = 42$

In order to solve this **system of two equations**, there is a standard step-by-step procedure. We pause the solution of this problem to outline it.

Procedure for Solving Two Equations in Two Unknowns

Step 1 - Solve one of the equations for one of the unknowns in terms of the other.

Step 2 - Substitute the result of Step 1 into the other equation.

Step 3 - Solve the resulting equation, which only contains one unknown.

Step 4 - Substitute the result of Step 3 into the result of Step 1 to obtain the value of the other unknown.

Example 1 $2x + 3y = 12$ Equation 1

$4x - 5y = 2$ Equation 2

Solution:

Step 1 : $x = 6 - 1.5y$ from Equation 1

Step 2 : $4(6 - 1.5y) - 5y = 2$

Step 3 : $24 - 6y - 5y = 2$

$24 - 11y = 2$

$22 = 11y$

$2 = y$

Step 4 : $x = 6 - 1.5 \times 2 = 3$ ■

Resuming the solution of the previous problem, and following this procedure, we obtain from (2)

(3) $P = 6.8 - .6\ B$

from equation (1). Substituting this into equation (2), we get

$$6 (6.8 - .6\ B) + 4B = 42$$

$$40.8 - 3.6\ B + 4B = 42$$

$$.4\ B = 1.2$$

So $B = 3$. Substituting this value into (3) yields

$$P = 6.8 - .6\ (3) = 6.8 - 1.8 = 5$$

Although checking is not a formal part of the solution procedure, it should always be performed in order to be sure that an arithmetic or algebraic foul-up has not occurred. Assuming that a pizza costs $5 and a pitcher $3, last Wednesday the five pizzas cost $25 and the three pitchers $9, for a total of $34. This Wednesday the six pizzas cost $30 and the four pitchers $12, for a total of $42.

Notice that we did not check the problem by seeing whether the values $B = 3$ and $P = 5$ satisfied equations (1) and (2); we checked the problem by seeing whether the answers we obtained satisfied the original conditions of the problem. Although it may seem like these two procedures are the same, they are not -- it is possible to derive incorrect equations from the problem, but solve the equations correctly without solving the problem correctly!

The preceding technique for solving two equations in two unknowns, known as Gaussian Elimination, always works, but the results can be divided into three separate sub-cases. Recall from the preceding section that the graph of a linear function is a straight line.

Case 1 - There exists a unique solution to the two equations. In this case, Gaussian Elimination will result in a single equation in a single unknown.

An example of Case 1: $x + 2y = 5$

$$3x - y = 1$$

From the first equation, $x = 5 - 2y$. Substituting this into the second equation yields $3(5 - 2y) - y = 1$. Simplifying, $15 - 7y = 1$, which is a single equation in a single unknown.

Case 2 - There does not exist a solution to the two equations. In this case, Gaussian Elimination will result in a statement, such as $0 = 1$, which can never be true.

An example of Case 2: $2x + y = 5$

$$4x + 2y = 7$$

Solving the first equation for y yields $y = 5 - 2x$. Substituting this into the second equation gives $4x + 2(5 - 2x) = 7$. This simplifies to $10 = 7$, an obvious impossibility.

Case 3 - There exist infinitely many solutions to the two equations. In this case, Gaussian Elimination will result in a statement, such as $2 = 2$, which is always true.

An example of Case 3: $x + 3y = 5$

$$2x + 6y = 10$$

From the first equation, $x = 5 - 3y$. Substituting into the second equation yields $2(5 - 3y) + 6y = 10$, which simplifies to $10 = 10$.

While all three cases can occur, at least in theory, in the real world the great majority of story problems will, if correctly analyzed, yield equations with unique solutions.

There are two major aspects to solving story problems of this type -- setting up the equations, and solving them. Most students do not have much difficulty solving the equations once they are set up, but translating real-world conditions to equations often presents a problem.

There are two main principles involved in setting up equations from real-world conditions.

The Totals Principle and the Rate Principle

Totals Principle - Totals are sums of subtotals. In the problem at the start of this section, the total amount of money spent is the sum of the pizza subtotal and the pitcher subtotal. The Totals Principle is based on addition.

Rate Principle - The cost of a number of objects, each of which sells for the same price, is the product of the price times the number of objects. In the problem at the start of this chapter, the price of five pizzas is five times the cost of a single pizza. The Rate Principle is based on multiplication, and is applicable in any situation involving a rate (not just when the rate is a monetary one).

Here are a few story problems which employ both the Totals and Rate Principles. While story problems can be constructed using other principles, these two are fundamental, and underlie a great number of problems in different areas of mathematics.

Example 2 - A party of five adults and two children buy tickets to a movie for a total cost of $34. A party of three adults and five children pay $28 to see the same movie. What is the cost of an adult ticket? What is the cost of a child's ticket?

Solution: Let A denote the cost (in dollars) of an adult ticket, C the cost of a child's ticket. For the first party, the five adults pay 5A, the two children 2C, which results in the equation $5A + 2C = 34$. Similarly, three adults and five children will pay $3A + 5C = 28$. Solving these two equations yields $A = \$6$, $C = \$2$. ■

The following example uses the Totals and Rate Principles in a different environment.

Example 3 - A triathlete trains by running for two hours and biking for three hours on Tuesday, covering a total of 110 miles. The next day he runs one hour and bicycles four hours, covering 130 miles. How fast does he run? How fast does he bike?

Solution: Notice that we must assume he always runs at the same rate, which we shall denote R, and always bikes at the same rate, which we shall denote B. Using the Totals Principle, we see that the total distance travelled is the sum of the distance run and the distance biked. Using the distance = rate x time formula (the Rate Principle), on Tuesday he ran 2R miles and biked 3B miles, so 2R + 3B = 110. On Wednesday, he ran 1R = R miles and biked 4B miles, so R + 4B = 130. The solution to these two equations is R = 10 miles per hour and B = 30 miles per hour. ■

The technique of Gaussian Elimination can be extended to more than two linear equations in more than two unknowns. Many reasonably-priced pocket calculators can handle the solution of up to thirty equations in thirty unknowns, and computers can deal with thousands of equations in thousands of unknowns. However, as yet no computer, other than the human brain, can handle the much more important and difficult task of translating a real-world problem into the language of algebra. New techniques in algebra are still being discovered in response to the translation of real-world problems into the language of algebra.

Section 4 - Quadratic Functions; Optimization Problems

The utility of a specific description of a process by a function goes beyond problems that reduce to solution of equations. For example, consider the following problem.

A gardener has 100 feet of fencing with which to enclose a rectangular garden. What is the largest area of this garden?

In order to get a start on this problem, we need to recall two formulas involving rectangles. If a rectangle has length L and width W, its area is given by the formula

$$A = L\,W$$

Its perimeter (the length of a walk around the edge of the rectangle) is given by the formula

$$P = 2\,L + 2\,W.$$

Returning to the problem of the rectangular garden, we know that the perimeter of the rectangle must be 100 feet. Therefore, $2 L + 2 W = 100$. We can use this formula to solve for one of the variables in terms of the other, obtaining $2 L = 100 - 2 W$, and so $L = 50 - W$. The area A is therefore given by the function

$$A = L W = L (50 - L) = - L^2 + 50 L$$

We use this function to compute the area for a few sample values of L.

L	10	15	20	25	30	35	40
A	400	525	600	625	600	525	400

The values in this table, and the rising-and-falling pattern of the areas, suggest that we should look a little more closely at the numbers around L = 25.

L	22	23	24	25	26	27	28
A	616	621	624	625	624	621	616

These tables give a strong indication that the maximum area occurs at L = 25 (notice that $W = 50 - L = 25$, so the rectangle would be a square!). We could continue this process, looking ever more closely at the values of L near 25 (we might look at 24.8, 24.9, 25, 25.1, and 25.2; then the next time at 24.98, 24.99, 25, 25.01, 25.02). We would amass more and more evidence that is right to use L = 25 to get the largest possible area of the garden, but this process will never provide us with absolute, incontrovertible proof.

There is a way to obtain such absolute, incontrovertible proof by performing a little algebraic manipulation on the functional description $A = - L^2 + 50 L$, as follows. Recall that $(x + a)^2 = (x + a) (x + a) = x^2 + 2 a x + a^2$.

$$A = - L^2 + 50 L = - (L^2 - 50 L) = - (L^2 - 50 L + 625) + 625$$

$= - (L - 25)^2 + 625$

This last expression is an extremely useful form for A, because we know that the squares of numbers can never be negative. Therefore, if L - 25 is non-zero, $(L - 25)^2$ must be positive, and we will be subtracting a positive number from 625, which obviously must result in a number less than 625. On the other hand, if L - 25 = 0 (which occurs when L = 25), then $A = 625 - (25 - 25)^2 = 625$, and this is the largest possible value for A.

The technique used in the algebraic manipulation, where we added and subtracted the same number in order to write the function in a more useful form, is known as **completing the square**. The problem that was solved is of a type known as **quadratic optimization**.

The big question, of course, is 'how do we know what number to add and subtract'? The following rule makes the adding and subtracting part of the problem unnecessary.

A Rule for Optimizing a Quadratic Function

Suppose that we wish to complete the square on the quadratic function $f(x) = a x^2 + b x + c$, where a, b, and c are constants. This function will have a smallest value (called a minimum) if $a > 0$, and a largest value (called a maximum) if $a < 0$. Moreover, this smallest or largest value always occurs when $x = -b/2a$, and the smallest or largest value is simply obtained by evaluating the function at $x = -b/2a$.

Example 1 - Suppose $f(x) = 3 x^2 - 24 x + 8$. Since $3 > 0$, this function has a minimum at $x = -(-24)/(2 \times 3) = 4$, and the minimum is $f(4) = 3 \times 42 - 24 \times 4 + 8 = -40$.

Of course, the above rule comes from adding and subtracting the correct number. Note that

$f(x) = a x^2 + b x + c = a (x^2 + b/a x + c/a)$

$= a (x^2 + 2 (b/2a) x + b^2/4a^2 + c/a - b^2/4a^2)$

$= a (x + b/2a)^2 + a (c/a - b^2/4a^2)$

In this form, it can easily be seen that the crucial expression is $(x + b/2a)^2$, which is positive unless x = -b/2a, when the expression is 0. If a is positive, unless x = -b/2a, the expression $a (x + b/2a)^2$ is positive, so the function must have its smallest value when x = -b/2a. If a is negative, unless x = -b/2a, the expression $a (x + b/2a)^2$ is negative, so the function must have its largest value when s = -b/2a. In either case, the smallest or largest value of the function occurs precisely when x = -b/2a.

The above rule is an example of a 'good' rule, because it replaces a lot of messy calculation (completing the square) with a simple evaluation. It is far better to remember one 'good' rule than many 'bad' ones (this book has tried to avoid stating any 'bad' rules). However, quadratic optimizations can always be accomplished by completing the square, rather than remembering the rule.

The next problem is a simple variation on the one which started this section.

Example 2 - Suppose that the gardener still has 100 feet with which to fence in the rectangular garden, but that one of the edges of the garden is to be up against the side of the gardener's house, which is 100 feet long, so the gardener does not have to fence in this side. What is the largest area of such a garden?

Solution: If we denote the side of the garden flush with the house by L and the other sides by W, the area is still given by A = L W, but the amount of fencing needed is now only L + 2 W = 100. So L = 100 - 2 W, and so $A = L W = (100 - 2 W) W = - 2 W^2 + 100 W$. Using the Rule, we see that the maximum (since - 2 < 0) value occurs at W = -(100/2 x -2) = 25, and in this instance A = -2 x 252 + 100 x 25 = 1250 square feet. Notice that in this case L = 100 - 2 W = 100 - 2 x 25 = 50 feet, and so the side of the garden flush with the house is twice as long as the other sides. ■

Here's a final example to finish off this chapter.

Example 3 - The Rose Club wishes to sell tickets to an upcoming floral exhibition. They have found that they could sell 400 tickets at $10, and that every additional dollar they raise prices decreases the number of customers by 20. At what price should they sell tickets to maximize the revenue, and how much will they make?

Solution: If they sell tickets for 10 + x dollars, they will be adding x dollars to the original price of the ticket, so they will lose 20x customers. Therefore, the total revenue R from ticket sales at a price of 10 + x dollars will be the number of tickets multiplied by the price, or

$$R = (10 + x) (400 - 20 x) = - 20 x^2 + 200 x + 4000$$

The maximum revenue occurs when the x = -200/(2 x -20) = 5. At a ticket price of $15 (= $10 + $5), they will attract 400 - 20 x 5 = 300 customers, and so will receive $4500. ■

Notice that a quadratic optimization problem allows us to achieve a 'best' result simply by knowing the type of function with which we are dealing. Admittedly, all the problems that we have encountered so far involve optimizing quadratic functions, so the reader may well be wondering whether it is possible to optimize other types of functions. Optimizing other types of functions is the province of **calculus**, which is a branch of mathematics simultaneously discovered by Isaac Newton and Gottfried Leibnitz (although the Greeks had a glimmering of some of the basic ideas of calculus).

Optimization problems have important economic consequences. The next time you go to a supermarket, you might notice that cylindrical juice and soup cans have essentially the same shape, which can be measured by computing the ratio of the height of the can to its diameter. Cylinders that resemble a pizza have a very small height-to-diameter ratio, and cylinders that resemble a strand of uncooked spaghetti have a very large height-to-diameter ratio. If a can of soup is to contain 16 fluid ounces of soup, it is very important to construct the can as cheaply as possible.

Major soup manufacturers make billions of cans of soup a year, so a cost of an extra penny in manufacturing the can amounts to tens of millions of dollars!

Chapter 6 - Mathematics of Finance

Introduction

If you are a typical American, during your lifetime you will be involved with a substantial number of financial transactions. As a result, your ability to understand the mathematics of borrowing and lending will probably be worth a minimum of tens of thousands of dollars to you, and possibly even more. Along with the mathematics of borrowing and lending, you will find some financial advice. Whenever you are sent a prospectus for an investment you will usually see a disclaimer of the form "past performance is no guarantee of future success". The same is true for financial advice; other than a general statement such as "buy low and sell high" (nice work if you can consistently get it), the validity of financial advice varies with both time and place. Hopefully it will help you – but simply being aware of advice that has proved helpful to some should be of help to you.

It is astounding how people who will carefully evaluate which of two types of TV to buy, where the cost of the two TVs may differ by at most a couple of hundred dollars, will immediately accept the first loan shoved under their nose by anyone who agrees to finance their purchase of a car or a house. Just as there are 'best buys' in TVs, there are 'best buys' in credit, for when you shop for a loan, or decide what to do with your investment capital, you are making a purchase of money.

This chapter is unlikely to make you rich. However, it can save you substantial amounts of money during your lifetime -- far more than you might imagine. Since this chapter deals extensively with calculations, you would do well to own a calculator. A simple one, costing only a few dollars, that just has the arithmetic functions (+, -, x, and ÷) will suffice, but life will be easier if you have one with exponentiation (this key is usually denoted y^x). It will only cost a few dollars more.

Section 1 - Simple and Compound Interest

Many jobs pay wages which are computed by multiplying the length of time worked by the pay rate. If you are paid $8 per hour and work for 10 hours, you are paid $80. **Simple interest** can be thought of as the wages earned by money, and is computed exactly the same way, by multiplying the length of time the money is working by the **interest rate**.

Example 1 - Suppose that Sue loans John $500 for 3 years at an interest rate of 6% per year. How much does the money that Sue loans earn? How much does John have to pay back? Assume that simple interest is being charged.

Solution: 6% of $500 is .06 x $500 = $30, so each year John has the money costs $30. Since he borrows it for 3 years, the money 'earns' 3 x $30 = $90 during that period. John will have to repay the $500 he borrowed, plus the $90 interest, for a total of $590. ▉

In Example 1, the amount John borrows ($500) is called the **principal**. The basic time unit used for the computation of the interest rate (1 year) is called the **interest period**. The interest rate (6% per year) is always given as a percentage per interest period.

The general formulas governing computation of simple interest are straightforward.

Rules for Simple Interest

If one borrows a principal P at a rate r per interest period for a total of t interest periods, then the total amount of simple interest I is given by

$$I = Prt$$

The total amount A that must be repaid is

$$A = P + I = P + Prt = P(1 + rt)$$

Although the rate r is always expressed as a percentage per interest period, when computing with it remember to convert the percentage to a decimal (by dividing by 100).

Most problems involving simple interest are, appropriately enough, fairly simple.

Example 2 - Kim borrowed money from Juan at 8% per year simple interest. After two and a half years, she paid Juan $720. How much did Kim borrow?

Solution: This is basically a 'plug-in' problem, with A = $720, r = .08, and t = 2.5. So

$$\$720 = P(1 + .08 \times 2.5) = 1.2\ P$$

Therefore, P = $720/1.2 = $600 . ■

If one is borrowing money, the principal P is sometimes called the **present value** of the loan, and the total amount A to be repaid is called the **future value**.

Notice that in Examples 1 and 2, the interest is added on to the principal, and the entire amount is paid back when the **loan period** (3 years in Example 1, 2 1/2 years in Example 2) expires. **Add-on interest** is the amount of simple interest that is added to the principal.

Compound Interest

Look again at Example 1. Sue loans John $500 for 3 years, and John has the use of the entire $500 for all 3 years. An important feature of simple interest is that the recipient has the use of all the money loaned for the full period of the loan.

Now let's take a look at the money in the Alma Steadman Trust in the story. If she deposits the original $2,000,000 at 6% compounded annually, it is the same as loaning the money to the bank. At the end of every year, the money that was deposited at the start of the year, plus the interest that money has earned, is returned by the bank to Alma, and she immediately redeposits the new amount for another year.

Year	Balance on Jan. 1	Interest	Balance on Dec. 31
2004	$2,000,000.00	$120,000.00	$2,120,000.00
2005	$2,120,000.00	$127,200.00	$2,247,200.00
2006	$2,247,200.00	$134,832.00	$2,382,032.00
2007	$2,382,032.00	$142,921.92	$2,524,953.92
2008	$2,524,953.92	$151,497.24	$2,676,451.16
2009	$2,676,451.16	$160,587.07	$2,837,038.23
2010	$2,837,038.23	$170,222.29	$3,007,260.52
2011	$3,007,260.52	$180,435.63	$3,187,696.15
2012	$3,187,696.15	$191,261.77	$3,378,957.92
2013	$3,378,957.92	$202,737.48	$3,581,695.40

Notice that, after 1 year, the balance on 12/31/04 is $2,000,000 x 1.06. After 2 years, the balance on 12/31/05 is $2,000,000 x 1.06^2 . After 3 years, the balance on 12/31/06 is $2,000,000 x 1.06^3 . Finally, after 10 years, the balance on 12/31/13 is $2,000,000 x 1.06^{10}. $2,000,000 is the original principal, 1.06 is 1 plus the annual interest rate, and 10 is the number of years that the deposit has been earning interest. This shows that the future value A of a principal P deposit at an annual interest rate r for N years is given by the formula

(1) $A = P (1 + r)^N$

This equation is extremely important. It involves four quantities: the principal P, the future value A, the annual interest rate r, and the number of years N that the money has been earning interest. If

any 3 of these quantities are known, it is possible to solve for the fourth quantity in terms of the three missing quantities. For example, the present value P as a function of A, r, and N is given by

$$P = A / (1 + r)^N$$

Example 3 - Ellen deposits $8,000 in a bank at 5% compounded annually. How much will be in the account after 4 years?

Solution: This is simply a matter of using formula (1) with P = $8,000, r = .05, and N=4.

$$A = \$8,000 \times 1.05^4 = \$9,724.05$$ ■

With a calculator which has an exponential key, one computes 1.05^4, and then multiplies it by $8,000. If you have a very basic calculator, you can obtain the same result by multiplying $8,000 x 1.05 x 1.05 x 1.05 x 1.05.

It is often important to compute the present value of an amount that will be needed in the future.

Example 4 - Jose's parents decide to give him a car when he graduates from college in 4 years. If they estimate the cost of the car as $10,000, and a bank is paying 6% compounded annually, how much must they deposit now to be able to buy the car when Jose graduates?

Solution: We must find the present value P of an amount whose future value A = $10,000, when money is compounded at an annual rate r = .06 for 4 years. So

$$P = \$10,000 / 1.06^4 = \$7,920.94$$

We can check that this is correct by seeing that $7,920.94 deposited for 4 years at 6% compounded annually yields a future value of $10,000. ■

Other Compounding Periods

When money is compounded annually, the interest is computed and added on at the end of the year, and the new total used as the principal for the next year. Other frequently-used compounding periods are semi-annual compounding (twice a year), quarterly compounding (four times a year), monthly compounding (12 times a year), and daily compounding (360 times a year). The fact that a banking year is only 360 days is a reminder of how difficult computation was B.C. (Before Calculators), because it is much easier to work with semi-annual, quarterly, and monthly compounding when the year is 360 days rather than 365.

We could simply use formula (1) to compute the future value of any principal if we are given the interest rate r per compounding period. Then N would be the number of compounding periods.

Example 5 - Meredith deposits $3,000 in a bank which pays a quarterly compounding rate of 1.5%. What is the amount in her account at the end of 3 years?

Solution: Since there are 4 quarters in a year, there will be 12 quarters in 3 years. Using formula (1)

$$A = \$3{,}000 \times 1.015^{12} = \$3{,}586.85 \quad \blacksquare$$

Most of the time, however, a loan does not specify the interest rate r per compounding period, but the annual compounding rate and the number of times per year that the loan is compounded. In Example 5 above, the bank would say that it paid 6% compounded quarterly. The quarterly compounding rate is determined by dividing the annual compounding rate of 6% by 4, the number of compounding periods in a year.

This leads to the following modification of formula (1). If a principal P is borrowed for N years at an *annual* rate r which is compounded t times a year, then the future value A is given by

$$(2) \qquad A = P\left(1 + r/t\right)^{Nt}$$

Under the same conditions, the present value P can be computed from the future value by means of the formula

$$P = A / (1 + r/t)^{Nt}$$

Example 6 - Suppose that $8,000 is deposited in an account for 5 years at 8% annually. Compute the amount in the account if compounding is done (a) annually, (b) semi-annually, (c) quarterly, (d) monthly, and (e) daily.

Solution:

(a) $A = \$8{,}000 \times 1.08^{5} = \$11{,}754.62$

(b) $A = \$8{,}000 \times 1.04^{10} = \$11{,}841.95$

(c) $A = \$8{,}000 \times 1.02^{20} = \$11{,}887.58$

(d) $A = \$8{,}000 \times 1.0066667^{60} = \$11{,}918.77$

(e) $A = \$8{,}000 \times 1.0002222^{1800} = \$11{,}934.07$

While (a), (b), and possibly even (c) can be done without a calculator which can handle exponentials, (d) and (e) are simply too much work unless such a calculator is available. In (d) and (e), the fractions have been rounded off to decimals after seven places, such as .08/12 = .0066667, but the actual calculations were made without rounding off. ■

Example 6 makes two things clear. The first is that the more frequently money is compounded at the same annual rate, the greater the future value. The second is that either financial calculators, or calculators with an exponentiation key, are indispensable for computing when the compounding period is monthly or daily. Prior to the invention of electronic computers, the very word 'computer' did not refer to a machine, but to an individual who was employed to do these sorts of calculations

every day! A 'computer' was a job description, not something that could be bought at an electronic supply store.

Effective Yield

Let's look at Example 6. When we compute the future value of the $8,000 when the rate is 8% compounded semi-annually, we need to evaluate $A = \$8{,}000 \times 1.04^{10}$. Since $10 = 2 \times 5$, we could also write $A = \$8{,}000 \times (1.04^2)^5$. Since $1.04^2 = 1.0816$, we have $A = \$8{,}000 \times 1.0816^5$. In other words, if we invested the $8,000 at 8.16% compounded annually, the future value would be the same as if we invested the $8,000 at 8% compounded semi-annually. In this case, we say that the effective yield of 8% compounded semi-annually is 8.16%.

The effective yield of a rate is the rate at which money must be compounded annually to obtain the same future value. If money is loaned at an annual rate r compounded t times per year, the effective yield E is given by

$$E = (1 + r/t)^t - 1$$

Notice that the effective yield in the above formula, like the annual rate r, is a number. To convert it to a percentage, it must be multiplied by 100.

Example 7 - Compute the effective yields of the following, and express them as percentage rates.

(a) 5% compounded quarterly

(b) 7% compounded monthly

(c) 6% compounded daily

Solution: Applying the formula for E, we obtain

(a) $E = (1 + .05/4)^4 - 1 = .050945$, or 5.0945%

(b) $E = (1 + .07/12)^{12} - 1 = .07229$, or 7.229%

(c) $E = (1 + .06/360)^{360} - 1 = .061831$, or 6.1831% ■

The effective yield can be used as a 'benchmark', to compare which rate is better (will return a greater future value).

Example 8 - Compute the effective yields of 8 1/4% compounded monthly and 8 3/8% compounded semi-annually. Which gives the greater return on investment?

Solution: The effective yield of 8 1/4% compounded monthly is

$$(1 + .0825/12)^{12} - 1 = .085692, \text{ or } 8.5692\%$$

The effective yield of 8 3/8% compounded semi-annually is

$$(1 + .08375/2)^{2} - 1 = .085504, \text{ or } 8.5504\%$$

8 1/4% compounded monthly gives the greater return on investment. If you are only investing a few thousand dollars, this difference is obviously not significant. However, if you are investing the pension fund for the employees of the state of California (this fund is currently about $200 billion), the difference comes to about $37,725,000! ■

Section 2 - Closed-End Credit Plans; Annuities

A closed-end credit plan is a method of paying back a loan such that the date of the last payment is known in advance. The most common type of closed-end credit plan is the installment plan, in which one makes the same payment at the same interval (usually monthly) until the loan (plus interest) is paid off.

Compound Interest as Wages Paid on Money; the APR

Another way of viewing compound interest is by regarding interest as money paid for dollar-years, which is the financial equivalent of man-hours. If you receive 8% compounded semi-annually on $100, the $100 "works" for the first half-year, for a total of $100 x 1/2 year =

50 dollar-years. For 50 dollar-years, you are paid 8% of 50 = .08 x 50 = $4. For the second half of the year, one has $100 + $4 = $104, which now "works" for the second half-year, for a total of $104 x 1/2 year = $52 dollar-years. For 52 dollar-years, one is paid 8% of 52 = .08 x 52 = $4.16. At the end of the year, you now have $104 + $4.16 = $108.16. You have supplied 50 + 52 = 102 dollar-years, and have earned $108.16 - $100 = $8.16 interest. Notice that $8.16/102 = .08, which is 8%. This method of computing interest rates is called the *annual percentage rate*, or *APR*. In 1969, Congress passed the Truth-in-Lending Act, which requires all finance plans to state their interest rate as an APR. From the example above, one can see that whether 8% interest is paid simply or compounded, the APR is still 8%.

Computing the Annual Percentage Rate (APR)

The APR is defined as follows:

APR = 100 x total interest paid / dollar-years used

Example 1 - John borrows $98 from Sue. He pays Sue $60 after 6 months, and $45 at the end of the year. What is the APR?

Solution: The total interest paid is $60 + $45 - $98 = $7. It is customary to assume that the interest is paid off in proportion to the amount of the payment. Since the first payment of $60 is 4/7 of the total debt ($60/$105 = 4/7), 4/7 of the total $7 interest, or $4, went for interest. Therefore $60 - $4 = $56 of the principal has been paid off, leaving $98 - $56 = $42 to be used for the second half-year.

The total number of dollar-years John has used is $98 for 1/2 year plus $42 for 1/2 year, or $98 x 1/2 + $42 x 1/2 = 70 dollar-years. The APR is therefore 100 x 7/70 = 10%. ■

Add-on Interest and Payment Plans

Recall that add-on interest is simple interest added to the amount of the loan. If you borrow $200 for 2 years at 8% add-on interest, the total amount of interest owed is $200 x .08 x 2 = $32. The total that you owe is $200 + $32 = $232.

Suppose you arrange to pay this back in four equal installments of $232/4 = $58 every half-year. To compute the APR, we think of $232 as $200 principal + $32 interest, so each $58 payment is $200/4 principal + $32/4 interest = $50 principal + $8 interest. We therefore have the use of $200 for 1/2 year (100 dollar-years), $150 for 1/2 year (75 dollar-years), $100 for 1/2 year (50 dollar-years), and $50 for 1/2 year (25 dollar-years). We have therefore used 100 + 75 + 50 + 25 = 250 dollar-years, and paid $32 interest. The APR is 100 x 32/250 = 12.8%. It should not surprise us that the APR is higher than the add-on interest rate, because we computed the interest on the entire $200 for the entire two-year period, but we did not get the use of all of the money all of the time.

Computing Payments and APRs from Add-on Rates

To compute the amount of each payment (a.k.a. installment), divide the total amount to be repaid (principal + interest) by the number of payments.

Suppose that we borrow a principal of P dollars for Y years at an add-on rate of r (this is a number, not a percentage), and we intend to pay it back in N equal installments. The total interest paid is I = PrY, and so the amount that must be paid back is P + PrY = P(1 + rY). Since we pay it back in N equal installments

Each installment = P (1 + r Y) / N

The relation between the add-on rate r, which is a number, and the APR, which is a percentage, is

APR = 100 x (2 N r / (N+1))

In the example given at the start of our discussion of add-on interest and payment plans, we computed the APR of a $200 loan for 2 years at 8% add-on interest paid back in 4 installments to be 12.8%. Notice that, if we just plug r = .08 and N = 4 into the above formula for converting add-on rates to APRs, we obtain APR = 100 x (2 x 4 x .08 / 5) = 12.8%.

Example 2 - Maria buys a home-entertainment unit for $1200 at 6% add-on interest. She plans to pay for it in monthly installments over a 3-year period. How much is each payment? What is the APR?

Solution: The total interest Maria must pay is $1200 x .06 x 3 = $216, so Maria must repay $1200 + $216 = $1416. There are 36 monthly payments in 3 years, so each installment will be $1416/36 = $39.34 (payments are always rounded up to the nearest cent).

The APR is 100 x (2 x 36 x .06 / 37) = 11.68% . ■

Sometimes the APR is given as a monthly rate (credit cards are good examples) or a daily rate (one can make 1-day investments at a bank -- these are sometimes called "overnight paper"). To obtain the APR from the monthly rate, multiply the monthly rate by 12. To obtain the APR from the daily rate, multiply the daily rate by 365. There are many peculiarities in the credit business, and one is that the compounding year consists of 360 days, whereas the APR year consists of 365 days.

Ordinary Annuities and Annuities Due

One of the major expenses that parents must plan for is the college education of their children. With expenses at some schools approaching $30,000 per year or even more, it is impossible for many families to make a lump-sum payment at the time the child is born which will suffice to pay for schooling when the child reaches college age. One way to obtain the money is to make periodic deposits in a savings account.

Let's suppose that, at the end of each month, Erin's parents deposit $100 in a savings account which pays 6% compounded monthly. We can ask how much will be in the savings account on Erin's 18th birthday.

One way to do this is to compute the future value of each of the 12 x 18 = 216 deposits that will be made at the end of each month. Each deposit earns interest for a different length of time. The

first deposit earns interest for 17 years, 11 months = 215 months; the second for 17 years, 10 months = 214 months; ... ; the next-to-last deposit earns interest for 1 month, and the last deposit earns no interest at all. Since the interest rate is .06/12 = .005 per compounding period, we can compile the following table.

Which Deposit	Future Value of Deposit
Last	$100
Next-to-last	$100 x 1.005
...	...
2nd	$100 x 1.005^{214}
1st	$100 x 1.005^{215}

Adding up the total of these 216 numbers would be excruciating, even with a calculator. Fortunately, if we write down the total, we can see a pattern!

$$\$100 + \$100 \times 1.005 + ... + \$100 \times 1.005^{214} + \$100 \times 1.005^{215} =$$

$$\$100 \times (1 + 1.005 + ... + 1.005^{214} + 1.005^{215})$$

There is a formula for the sum in parentheses; it is

$$(1.005^{216} - 1) / (1.005 - 1) = 387.35$$

Therefore, the value of the account on Erin's 18th birthday will be approximately $38,735.

The type of payment scheme described above is called an **ordinary annuity**.

Ordinary Annuities

An ordinary annuity is a payment plan such that (1) the frequency of payments is the same as the frequency of compounding, and (2) the first payment is made at the end of the first time period (as are all subsequent payments).

Assume that each payment is m, and the payments are deposited into an account with annual rate r compounded t times a year for N years. Let i = r/t denote the interest rate per compounding period. Then the future value A of the annuity when the last deposit is made is

$$A = m [(1 + i)^{Nt} - 1] / i$$

Ordinary annuities are frequently used when one wishes to plan for a long-term expense, or even a not-so-long-term one, as indicated in Example 3 below.

Example 3 - Carlos plans on making deposits into a Christmas Club account so that he will have $500 to spend on presents next Christmas. His bank offers a Christmas Club plan that compounds at 4 1/2% monthly. Carlos plans to make his first deposit on January 31st, and his last on November 30th. How much should he deposit each month?

Solution: This is an ordinary annuity with future value A = $500, a total of 11 payments, and an interest rate i = .045/12 = .00375. Therefore, if the monthly payment is to be m, the formula shows that

$$\$500 = m (1.00375^{11} - 1) / .00375 = 11.21 m$$

Carlos should deposit $500/11.21 = $44.61 monthly. ■

A Christmas Club of the type described in Example 3 is known as a **sinking fund**. This is something of a misnomer. It is not the fund that is sinking -- its value is actually rising as time goes on. The depositor is sinking money into the fund.

Sinking funds are frequently used when a business wishes to make a major purchase some time in the future.

Annuities Due

The only difference between an ordinary annuity and an annuity due is that in an **annuity due**, each payment is made at the beginning of the interest period. In an ordinary annuity, each payment is made at the end of the interest period.

The effect is that each deposit receives one more interest period's worth of interest. If we return to the example of Erin's parents who earlier set up an ordinary annuity based on monthly deposits for 18 years at 6% compounded monthly, imagine this time that they set up an annuity due, making the first deposit the day Erin is born. This time, the value of the deposits would be

Which Deposit	Future Value of Deposit
Last	$100 x 1.005
Next-to-last	$100 x 1.005^2
...	...
2nd	$100 x 1.005^{215}
1st	$100 x 1.005^{216}

The sum of the future values of the deposits would be

$$\$100 \times (1.005 + 1.005^2 + ... + 1.005^{215} + 1.005^{216})$$

The sum of the quantity in parentheses would be

$$1 + 1.005 + ... + 1.005^{216} - 1 = (1.005^{217}-1)/.005 - 1 = 390.29$$

Therefore, the value of the annuity due is about $39,029.

An annuity due is an ordinary annuity in which the first payment is made at the beginning of the first time period (as are all subsequent payments).

Assume that each payment is m, and the payments are deposited into an account with annual rate r compounded t times a year for N years. Let i = r/t denote the interest rate per compounding period. Then the future value A of the annuity at the end of the last period (notice that this is not the moment when the last payment is made!) is

$$A = m [(1 + i)^{(N+1)t} - 1] / i - m$$

$$= m f \quad \text{where } f = [(1 + i)^{(N+1)t} - 1] / i - 1$$

An annuity due has a higher future value than the corresponding ordinary annuity because each payment in an annuity due is made one payment period before the corresponding payment in the ordinary annuity, and thus has more time to accumulate interest.

Example 4 - Vera deposits $200 on the first of every month into a TSA (tax-sheltered annuity), starting January 1, 2014. The account compounds at the rate of 5% monthly. How much will she have when she makes her last deposit on December 1, 2028?

Solution: Vera will have deposited money for 15 years, which will be 180 deposits. The multiplication factor will be

$$[(1 + 5/1200)^{181} - 1] / 5/1200 - 1 = 269.40$$

Her balance will therefore be $200 x 269.40, or approximately $53,880. ■

As in Example 3, one can also compute the payments required to meet a future obligation with an annuity due.

Example 5 - At the start of the fiscal year, Rutabaga Chemicals plans to build a new plant 10 years in the future. The cost of $35,000,000 must be delivered three months before construction begins.

They will make quarterly deposits starting immediately in an account which pays 6 1/2% compounded quarterly. How much is each quarterly payment?

Solution: There are 40 payments to be made in an annuity due, with i = .065/4 = .01625. The multiplication factor f is therefore

$$(1.01625^{41} - 1)/.01625 - 1 = 57.632$$

Each quarterly deposit must be $35,000,000 / 57.632, or approximately $607,300. ■

Section 3 - Financing Plans

When one makes a major purchase, there are three primary ways of paying for it. You can pay cash (or check), paying off the entire amount. You can put it on your credit card. The third alternative is to finance it, possibly paying part of the amount initially (this is known as the **down payment**), and paying off the rest, usually by making periodic payments of the type discussed in the previous section.

The Importance of Financing

If you purchase a bicycle for $100 or so, the finance charges (if any) will only amount to a few dollars. If you purchase a car for $10,000, financing charges will run to the thousands of dollars. If you buy a house for the national average (approximately $190,000 as of this writing), the total dollar amount of finance charges will likely exceed the initial cost of the house by a substantial amount.

Consequently, you should shop for financing with the same care that you shop for the item you are purchasing. Loaning money is a competitive business, and there are many competitors who want your business. While there are no absolutes in the financing game, there are a few guidelines which are fairly reliable.

1) Every organization that offers you financing must notify you of the APR. It doesn't hurt to calculate the APR on your own, because mistakes have been made in doing so.

2) In general, when making a purchase such as a car, the dealer will try to get you to let him (or her) finance it. Auto dealerships often make more money from the financing than from the car. Don't be a 'one-stop shopper'. Financing through a dealer can be very expensive.

3) Investigate financing possibilities before making a major purchase such as a car, not after.

4) If you or some member of your family belongs to an organization that has a credit union, check out the possibility of obtaining financing through it. This can be one of the greatest benefits of the organization.

Credit Cards

As of the writing of this book, prevailing interest rates are in the 4% - 5% range for home mortgages, but credit card rates are MUCH higher. As a result, it has become easier and easier to get credit cards, and students are often inundated with credit card offers on graduation.

A credit card gives the cardholder the opportunity to defer payment on the items purchased until the next billing period (billing on credit cards is done monthly, and the cardholder usually has approximately two weeks after billing to pay). A credit card is an *open-end payment plan*, because neither the amount of the payment nor the date of the last payment is scheduled in advance. The cardholder is required to pay a minimum percentage (usually about 4%) of the bill; a bill of $300 would therefore require a minimum monthly payment of $12.

If the cardholder does not pay the entire amount of the bill, the amount remaining is called the *unpaid balance*. Interest is charged on this unpaid balance at either a daily (on a 365-day year) or, more usually, a monthly rate.

This balance can be computed in three different ways.

Computing the Unpaid Balance on a Credit Card Account

1) Previous balance: interest is charged on the balance as of the previous month.

2) Adjusted balance: interest is charged on the balance as of the previous month, with payments made to the credit company deducted and credit transactions added.

Example: If the previous month's balance was $500, and the cardholder made a payment of $100 and bought a weed whacker for $75, the adjusted balance is $500 - $100 + $75 = $425.

3) Average daily balance: every day the balance is recomputed based on payments made and credit transactions added. The daily balances are added, and the result divided by the number of days in the billing period. The interest is then computed as follows: suppose that D is the average daily balance, A is the APR expressed as a rate (not a percentage), and M is the number of days in the month. The interest I is given by the formula

$$I = D \times A \times M/365$$

Observe that the interest rate that is used is a monthly rate which is 1/12 of the APR if the interest is computed by the previous balance or adjusted balance method. If the average daily balance method is being used, it is calculated as the APR multiplied by a fraction that is 28/365 for February, 30/365 for April, June, September, and November, and 31/365 for all other months.

The average daily balance is simply a weighted average, with the weighting factor being supplied by the number of days during the billing period that a particular balance remained the same.

The date the payment is due is called the **due date**. If payment is received on or before the due date, no interest is charged, but after that, the 'interest meter' starts ticking.

Example 1 - Kim's credit card company charges 1 1/4% per month on the previous balance. Her last monthly bill showed a balance of $800. During the previous month she sent in a payment of $250 and bought a CD player for $140. What will be the balance on her next bill?

Solution: Unless Kim pays off the entire amount, interest is charged on the previous balance of $800. This amount will be .0125 x $800 = $10. Her next balance will therefore be

$800 - $250 + $140 + $10 = $700. ■

Notice that, in Example 1, interest is charged on the full previous balance because the entire balance was not paid off. Credit is given for payments made on the adjusted balance method, provided that the payment is received on or before the due date.

Example 2 - The previous balance on Alicia's credit card was $540. Her credit card company charges 18% APR. Alicia made a payment of $85 before the due date, and bought a windsurfer for $320. What will be the balance on her next monthly bill?

Solution: Because the payment was received prior to the due date, she will receive credit for it. Her adjusted balance is therefore $540 - $85 + $320 = $775. Since 18% APR is 18%/12 = 1 1/2% per month, the interest charged on the adjusted balance will be .015 x $775 = $11.63. Her next balance will be $775 + $11.63 = $786.63. ■

More computation is involved when a company computes average daily balances, but this is perhaps the fairest method. Computers make it easy for a finance company to determine average daily balances.

A credit card can actually make you money (well, sort of) - although not a whole lot, so don't quit the day job! If you buy an item close to the end of the billing period, there will be a lag of about two weeks between the time of purchase and the moment the 'interest meter' starts ticking. If you buy it close to the beginning of the billing period, this 'grace period' can approach six weeks. That means you can keep the money in a bank earning interest during this grace period. In contrast, if you pay in cash or by check, this earning potential is lost.

Example 3 - The balance as of August 31st in Maria's credit card account was $880. The company bills on the 1st of every month by charging 1 1/2% interest on the average daily balance during the preceding month. All transactions are reflected in the revised balance the next day. Maria made a payment to the card company of $90 which was credited on September 8th, bought a microwave for

$135 on September 14th, and a DVR for $280 on September 23rd. What will her bill read as of September 30th?

Solution: The balance of $880 was applicable for 8 days (1st through 8th), the balance of $880 - $90 = $790 was applicable for 6 days (9th through 14th), the balance of $790 + $135 = $925 was applicable for 9 days (15th through 23rd), and the balance of $925 + $280 = $1,205 applicable for 7 days (24th through 30th). Her average daily balance was
(8 x $880 + 6 x $790 + 9 x $925 + 7 x $1,205)/30 = $951.34

The APR is 12 x .015 = .18 when expressed as a rate, so the interest charged will be $951.34 x .18 x 30/365 = $14.08 (it is actually $14.0746, but that .46 cent will be rounded up to the nearest penny). This will be added to the last balance of $1,205, so Maria's next bill will be for $1,219.08 . ■

Some credit cards have no grace period, but feature a lower interest rate instead.

We are concerned with the mathematics of financing, rather than the ins and outs (of which there are many) of car-buying. Buying a car is not like buying a hamburger at McDonald's -- you cannot haggle over the price of the hamburger, whereas you MUST haggle over the price of the car to avoid being ripped off. If you buy a new car, there will be a sticker on the window which lists what features the car has, and a bottom-line price, known as the **sticker price**.

DO NOT ACCEPT THE STICKER PRICE!!! Instead, offer the dealer 85% of this price in a bad economy, and 90% in a good one. A dealer will never tell you to get lost, although he (or she) may sneer at your offer. He (or she) will come back with a counter-offer, and from now on you're on your own. Do not pay more than 95% of the sticker price under any circumstances.

If you are buying a used car, either get the Blue Book, which will give the average price of the particular model of used car you are interested in, or read the advertisements in your local paper to

get a 'feel' for the asking prices of the car. Once again, this is a haggler's market, as usually the seller states the price wanted, and you make a counter-offer.

Financing Your Car

After all the negotiation has been completed, you will probably have to finance your car, especially if you have purchased a new car from a dealer. If you have not arranged for financing in advance, the dealer will probably offer you several different payment plans, including varying down payments and different lengths for financing. The dealer is obligated to tell you the APR, but many mistakes have been made in computing this number.

To compute the APR given the amount of money P to be financed, first compute the total interest I. If r is the annual add-on rate and y is the number of years the car is financed, then

$$I = Pry$$

So

$$r = I / Py$$

Now compute the APR from the formula

$$APR = 100 \times 2 N r / (N + 1)$$

where N is the total number of payments made.

Example 4 - You decide to buy a car, and after much haggling, the final price is $13,600. You pay 15% down, and finance it with monthly payments of $287.82 over a four-year period. What is the APR?

Solution: First, determine the total interest I. You will make 48 payments of $287.82, or $13,815.36. Your down payment is 15% of $13,600, or $2,040, so the amount P to be financed is $13,600 - $2,040 = $11,560. The interest I is therefore I = $13,815.36 - $11,560 = $2,255.36. So

r = I/Py = $2,255.36/(4 x $11,560) = .0488

Therefore, the APR is

APR = 100 x 2 x 48 x .0488 / 49 = 9.56% ■

It is not the purpose of this book to cast stones at car dealers, but since this calculation is fairly simple to make, do a few practice problems before you go to a car dealer with a calculator, and then go through the APR calculations in the salesperson's office. If you suspect you are being hoodwinked, remember, to err is human, and the mistake could be yours. If a dealer makes you an offer, he or she would like you to sign on the dotted line. You can always claim you want to talk it over with your spouse, even if you don't have one. Go home, and go through the calculations in peace and quiet. If your calculations agree with the dealer's statement of the APR, then you are both very likely to be right, because the odds against your both miscalculating them and agreeing is very high.

Example 5 - After agreeing on a price of $11,400, the dealer offers you the choice of two plans. Either pay 20% down and $430.27 a month for 24 months, or pay 10% down and $337.54 a month for 36 months. Assuming you can afford both, which has the more attractive APR?

Solution: If you pay 20% down, $430.27 a month: you will be financing 80% of $11,400, which is $9,120. Your total payments will be 24 x $430.27 = $10,326.48. The interest is $10,326.48 - $9,120 = $1,206.48. The add-on rate r is

r = I/Py = $1,206.48/(2 x $9,120) = .0661

The APR is

APR = 100 x (2 x 24 x .0661) / 25 = 12.69%

10% down, $337.54 a month: you will be financing 90% of $11,400, which is $10,260. Your total payments will be 36 x $337.54 = $12,151.44. The interest is $12,151.44 - $10,260 = $1,891.44. The add-on rate r is

$$r = I/Py = \$1,891.44/(3 \times \$10,260) = .0615$$

The APR is

$$APR = 100 \times (2 \times 36 \times .0615) / 37 = 11.97\%$$

The 10% down, 36 months to pay gives you the better APR. ▉

When shopping for a car, it is helpful to know the approximate price range you can afford to buy. If you know the down payment and monthly payments you can afford, you can get some idea of the price range at which you should be looking.

Example 6 - You can afford a $2,000 down payment and monthly payments of $250 for 48 months. If the APR is 8%, what would be the price of a car with these parameters?

Since APR = 100 x (2 x N x r) / (N + 1), and APR = 8, N = 48, we have to solve

$$8 = 9600r/49$$

So the add-on rate r = 8 x 49 / 9600 = .0408. We will be making 48 payments of $250, which totals 48 x $250 = $12,000. If P is the amount to be financed, then the interest I = Pry = P x .0408 x 4 = .1632 P. Therefore

$$\$12,000 = P + I = P + .1632\ P = 1.1632\ P$$

The amount to be financed is about $12,000/1.1632, which is approximately $10,300. Since you are willing to pay $2,000 down, your ceiling is in the $12,000 range.

A Final Hint

It is generally best if you have arranged for the possibility of financing in advance, but NEVER tell the dealer you have done so until you have agreed on the price of the car! Sometimes a dealer will allow you to haggle the price of the car down a little further, figuring to make it up on the financing.

Section 4 - Buying a Home

Buying a home is a complex procedure, made more complicated by variations in regulations and terminology from state to state and from year to year. This section makes no claims to being complete, and when the time comes to buy a home you should talk to someone *in your area* who is knowledgeable about both market conditions and the availability of different financing. Many books on the subject are also obtainable.

This section is concerned with brief introductions to three topics: how homes are generally financed, determining what you can afford, and tax implications of home ownership. What you will read here is very general, and will help you familiarize yourself with the major factors of buying a home. It is not 'Everything You Need to Know About Buying a Home in One Easy Lesson'!!!

Financing Your Home

Buying a home is similar to buying a car in that one generally makes a down payment and finances the remainder through monthly payments. Because financial organizations such as banks and S & Ls (Savings & Loans) generally finance homes, banks quote their interest rates on 360-day years and regard the payments as ordinary annuities (cars are financed using add-on interest rates which are converted to APRs based on a 365-day year).

When one buys a house, in addition to the down payment, there are a bewildering number of initial costs which generally add up to thousands of dollars. When financing is obtained, the loan (which is called a *mortgage*) can be financed in several different ways. The most common are *fixed-rate mortgages* and *adjustable-rate mortgages*, known as *ARM*s.

The interest rate on a fixed-rate mortgage is fixed (remains the same) for the duration of the loan. The interest rate on an ARM is initially lower than the rate on a fixed-rate mortgage, but if interest rates increase, the rate on the ARM will go up. In effect, an ARM represents a gamble that interest rates will not go much higher. The mathematics of ARMs depends on the structure of the particular ARM, and can get quite complicated. The interest rate paid on an ARM depends upon the interest rate environment.

Foreclosure: The F Word

If you do not make payments on a mortgage, you are subject to foreclosure, an ominous word which means the lender has the right to take possession of your property.

When you decide to purchase a home, you must take precautions to guard against foreclosure. If you finance with a fixed-rate mortgage, you know in advance what your monthly payments will be, and can plan with some degree of accuracy. If you finance with an ARM, an inflationary spiral can cause your monthly payments to skyrocket, greatly increasing the chances of foreclosure.

Because the monthly payments on a fixed-rate mortgage remain the same throughout the duration of the loan, it contains no unpleasant surprises. It is therefore easier to make plans with a fixed-rate mortgage, and we shall study only this type. However, if you anticipate that your financial situation will improve substantially in the next few years, ARMs become more attractive because you will be more able to survive the risk that the payments will increase in exchange for the reward of initially lower payments.

What Monthly Payments Can You Make?

Despite the movie images of the greedy banker eagerly waiting to foreclose on the old homestead, banks do *not* want to foreclose. They make their plans based on your continuing to make your payments, and generally do not want to assume the risk of owning real estate.

Banks have adopted the following rule of thumb: home payments should be 36% of your net monthly income (take-home pay less expenses).

Example: Suppose your take-home pay is $4,000 a month, and your expenses (exclusive of rent) are $1200 a month. You can afford monthly mortgage payments of 36% of $4,000 - $1,200 = $2,800, which is approximately $1,000.

The question of how much of a loan you can afford is related to the question of how much of a down payment you can afford. Banks will usually require a down payment of from 10% to 25%. If one considers the cost of the house as 100%, the ratio of the loan percentage to the down payment percentage is called the **loan-to-value**, or **debt-to-equity**, ratio.

Example 1 - A house costs $250,000. The bank has an 80% to 20% loan-to-value requirement. What down payment do you need, and how much will the bank loan you?

Solution: Note that even though the usual way to express a ratio of 80 to 20 is 4 to 1, it is expressed as 80% to 20% to make the computations obvious. You will be required to make a down payment of 20% of $250,000, which is .20 x $250,000 = $50,000. The bank will loan you 80% of $250,000, which is .80 x $250,000 = $200,000. ▉

The Mechanics of Home Loans

A fixed-rate mortgage usually involves two different types of payments. The monthly payment is made in the same way as monthly car payments. However, there are payments that must be made *immediately* as well (this is not the case with car loans). These payments consist of a processing fee (usually several hundred dollars, which defrays the cost of the necessary paper shuffling), and **points**. If the loan is for $50,000, and 3 1/2 points are demanded, the 'points' payment is .01 x 3.5 x $50,000 = $1,750.

Example 2 - What are the up-front (immediate) charges for a loan of $70,000 which costs 4 1/2 points and a processing fee of $350?

Solution: The points charge is .01 x 4 1/2 x $70,000 = $3,150, so the up-front charges total $3,150 + $350 = $3,500. ■

The following section presents a method for comparing two loans which have different long- and short-term parameters.

Comparing Two Different Home Loans

To compare home loan rates, an approximate formula has been devised which incorporates four factors: the amount of the loan L, the processing fee F, the interest rate r of the loan, and the number of points P expressed as a rate. The Comparison Loan Rate CLR is given by

$$CLR = r + .125 \times (P + F/L)$$

Example: On a loan of $80,000 a fee of $300 and 4 1/2 points are charged, in addition to the mortgage rate of 6%. The Comparison Loan Rate is

$$CLR = .06 + .125 \times (.045 + \$300/\$80,000) = .0661$$

In this case, the CLR is about 6.61%.

In computing the CLR, the interest rate of the mortgage is a long-term parameter. Points and the processing fee are short-term parameters, so the long-term parameter is weighted 8 times as heavily as the short-term parameters in computing the CLR.

Example 3 - On a loan of $80,000, the Continental Bank offers a mortgage rate of 8 1/2%, and charges a processing fee of $250 and 5 1/2 points. The International Bank offers a mortgage rate of 8 3/4%, with a processing fee of $300 and 4 1/2 points. Which loan is the better value?

Solution: The CLR for the Continental Bank is

$$CLR = .085 + .125 \times (.055 + 250/80,000) = .0923$$

The CLR for the International Bank is

CLR = .0875 + .125 x (.045 + 300/80,000) = .0936

You want the loan which costs less, which is the one offered by the Continental Bank. ■

Determining Your Monthly Payment

While there are formulas which enable one to compute exact monthly payments, as a rough approximation it is easier to use a table, such as the one listed in the table below.

Monthly Cost Per $1,000 Worth of Financing

Interest	15 Year Term		30 Year Term	
Rate %	Monthly Payment	Total Amount	Monthly Payment	Total Amount
4.00%	7.39	1331.43	4.77	1718.69
4.13%	7.45	1342.74	4.84	1744.73
4.25%	7.52	1354.1	4.91	1770.9
4.38%	7.58	1365.51	4.99	1797.42
4.50%	7.64	1376.98	5.06	1824.06
4.63%	7.71	1388.51	5.14	1850.9
4.75%	7.77	1400.09	5.21	1877.93
4.88%	7.84	1411.73	5.29	1905.14
5.00%	7.9	1423.42	5.36	1932.55
5.13%	7.97	1435.17	5.44	1960.15

5.25%	8.03	1446.97	5.52	1987.93
5.38%	8.1	1458.83	5.59	1015.89
5.50%	8.18	1470.75	5.68	2044.04
5.63%	8.24	1482.72	5.76	2074.36
5.75%	8.31	1494.73	5.84	2100.86
5.88%	8.37	1506.81	5.92	2129.54
6.00%	8.44	1518.94	6	2158.38
6.13%	8.51	1531.13	6.08	2187.4
6.25%	8.58	1543.36	6.16	2216.58
6.38%	8.64	1555.65	6.24	2245.93
6.50%	8.72	1567.99	6.33	2275.44
6.63%	8.78	1580.39	6.4	2305.12
6.75%	8.85	1592.83	6.49	2334.95
6.88%	8.92	1605.34	6.57	2364.94
7.00%	8.99	1617.89	6.65	2395.09
7.13%	9.06	1630.49	6.74	2425.39
7.25%	9.13	1643.15	6.82	2455.83
7.38%	9.2	1655.86	6.91	2486.43
7.50%	9.27	1668.62	6.99	2517.17

7.63%	9.34	1681.43	7.08	2548.06
7.75%	9.41	1694.29	7.16	2579.08
7.88%	9.48	1707.2	7.25	2610.25

Each entry in the table is the monthly payment for each $1,000 worth of financing at a given interest rate and loan duration. Thus, the monthly payment per $1,000 worth of financing for a 5 1/2% loan to be paid off in 30 years is $5.68. You will pay $2,044.04 for every $1,000 you borrow.

Example 4 - Your dream house will cost $400,000. The bank wants a loan-to-value ratio of 85%-15%, and will offer you a 5 1/2% mortgage to be repaid in 30 years. What is your down payment? What is your monthly payment?

Solution: Your down payment is 15% of $400,000, which is $60,000. The balance of $340,000 is the amount of the loan. Each $1,000 worth of financing requires a monthly payment of $5.68. Since you have $340,000 to be financed, your monthly payment will be 340 x $5.68 = $1931.20. ■

In Example 4, notice that during the course of the 30 years that you pay off the mortgage, you will make 12 x 30 = 360 monthly payments of $1931.20, which totals $695,232. If you decide to pay off the mortgage in 15 years instead, each $1,000 worth of financing will cost $8.18 a month, and so the monthly payments will be 340 x $8.18 = $2,781.20. During the life of the mortgage, you will make 15 x 12 = 180 payments, for a total of $500,616. Ostensibly, paying off the mortgage fifteen years earlier has saved you $695,232 - $500,616 = $194,616. (A question for you to consider -- in what sense have you saved nearly $195,000?)

Tax Implications

This is an extremely complicated subject, but the basic idea is fairly simple: interest on loans is tax-deductible. This means that, if you are paying a certain monthly rent, you can generally afford

higher monthly payments when buying a home, because the interest that you pay is tax deductible. This is most important at the outset of the loan, because that is when the lion's share of your monthly payments goes to pay off the interest.

An example will illustrate the basic idea. Let's say that you take out a loan for $100,000 at 5 1/2% for 30 years. Each $1,000 worth of financing costs $5.68, and so the monthly payments are 100 x $5.68 = $568. The monthly interest is 1/12 x 5.5% = 0.46% (approximately). Therefore the first month's interest on $100,000 is .0046 x $100,000 = $460. Of the $568 payment, $460 goes to pay off the interest (and is therefore tax deductible), and the remaining $108 = $568 - $460 goes to build equity in the property by reducing the balance of the loan. At the beginning of the second month, the *loan balance* is therefore $100,000 - $108 = $99,892.

The table above shows the loan balance, equity, and interest payments for the first six months of the loan described above.

First Six Months Loan Breakdown

Month	Balance	Interest Payment	Equity Payments	Total Equity
1	$100,000.00	$458.33	$109.67	$109.67
2	$99,890.33	$457.83	$110.17	$219.84
3	$99,780.16	$457.33	$110.67	$330.51
4	$99,669.49	$456.82	$111.18	$441.69
5	$99,558.31	$456.31	$111.69	$553.38
6	$99,446.62	$455.80	$112.20	$665.59

A computer can crunch out the entire 360-line table in seconds, but the basic idea is that initially, interest payments are high and equity build-up is slow. As the loan matures, the loan balance decreases, so the interest payments are smaller and equity build-up is faster.

For most people, buying a home is the most important financial decision they will ever make. This section should have given you enough information to get started, but you should always consult with someone knowledgeable before buying a home.

A Word of Caution

Many factors enter into the purchase of a home. Some of these factors are mathematical, some are psychological, and some are economic. Because home ownership is perceived to be an important component of the American Dream, many people overextend themselves financially in order to buy a home. If two incomes are necessary to make the payments and one person loses his or her job, it may prove necessary to sell the home if the payments cannot be met. If this sale is under forced circumstances and the real estate market has been declining, the resulting loss can be substantial.

Buying a home should probably be done only when there is a reasonable safety margin in all the necessary financial transactions. The down payment should still leave a cash reserve; and the monthly payments should not force a major change of lifestyle.

Chapter 7 - Set Theory

Introduction

Set theory is a branch of mathematics that is very different from arithmetic or algebra. The fundamental questions of arithmetic and algebra deal with computation -- what rules govern it, how best to perform it, and where to use it.

Although there are computational aspects to set theory, the primary use of set theory is to provide a framework in which to phrase and study a variety of important mathematical topics. It is impossible to conceive of mathematics without arithmetic and algebra, but almost all the branches of mathematics that use set theory were in existence long before the invention of set theory, and were surviving quite nicely without it.

Nonetheless, once set theory had been invented, it was immediately put to use in a variety of situations.

Sets and Inclusion

The first order of business in constructing a new branch of mathematics is to define the objects we shall consider. A *set* is defined to be a collection of things. Of course, this raises the obvious question: what is a 'thing'? Well, we can't define it, but we know what 'things' are -- a triangle, Abraham Lincoln, and the number 7 are all examples of things.

A set S can be specified by listing its things. For example, S = { 1,2,3,4 } is the set consisting of the whole numbers 1, 2, 3, and 4. This method of listing the **elements** ('element' is a more-impressive sounding synonym for 'thing') of a set works well if the set has only a few elements, but for sets with many elements it is inefficient, and is replaced by the 'set-builder' method. To describe the set of all whole numbers between 1 and 1000, one uses **set-builder notation**. When set-builder

notation is used, the "entrance requirements" for membership in the set are described by using letters as variables.

An example of describing a set using set-builder notation is to write the set S of all whole numbers from 1 and 1000 as S = { x : x is a whole number greater than or equal to 1 and less than or equal to 1000 }. This is read, "S is the set of all x such that x is a whole number greater than or equal to 1 and less than or equal to 1000." The symbol 'x' is sometimes called a **dummy variable**, because the particular symbol 'x' plays no part in the actual set. We could have used any other symbol to denote the same set, such as

$$S = \{ y : y \text{ is a whole number, } 1 \leq y \leq 1000 \}$$

or

$$S = \{ \$: \$ \text{ is a whole number, } 1 \leq \$ \leq 1000 \}$$

Given a specific thing, which we shall denote by t, and a set S, we can ask whether or not t belongs to the set S. We write $t \in S$ to indicate that thing t belongs to set S, and $t \notin S$ to indicate that thing t does not belong to S. Note that drawing a diagonal line through the symbol $\in$ negates it. This is a standard mathematical convention, which we are already familiar with from arithmetic: $2 + 2 \neq 3$ means that 2 plus 2 does not equal 3. The same convention is followed with public-information signs: a picture of a cigarette with a diagonal line through it means 'no smoking'. It is customary to use capital letters, such as A, B, and C, to denote sets, and lower-case letters, such as a, b, and c, to denote elements of sets.

Example 1 - Formulate the sentence, "George Washington was a President of the United States who did not play baseball," using sets. Use set-builder notation, and $\in$ and $\notin$ symbols.

Solution: Let A = { x : x was a President of the United States }

Let B = { x : x played baseball }

George Washington $\in$ A, George Washington $\notin$ B ■

Once a type of mathematical object has been defined, (sets in our case), there are many questions that can be asked. Can we compare two such objects? How can we combine them? We are of course familiar with these questions for numbers, but these questions are among those typically asked when new mathematical objects are first studied.

Two sets A and B are equal if they contain the same elements. If A = { 1,2,3,4 } and B = { 3,1,4,2 }, these sets are equal (which we write A = B) because they contain the same things, even though they were listed in different orders. In this sense, a set is like a lunchbox, because only the contents of the lunchbox matter -- the order in which they were placed in the lunchbox, or the order in which they were removed from the lunchbox, is unimportant.

Subsets

We say that A is a **subset** of B if every element of A is an element of B. If A = { 1,2,3 } and B = { 1,2,3,4 }, then A is a subset of B, written $A \subseteq B$. If A is a subset of B, we sometimes (but not often) say that B is a **superset** of A, written $B \supseteq A$. Notice that if A is not a subset of B, there must be some element of A which is not a member of B. Informally, if $A \subseteq B$, we say that A is contained in B, and B contains A.

Example 2 - Let A = { p, q, r }. Is p a subset of A? Is {p} a subset of A? Is { r, q, p } a subset of A?

Solution: p is not a subset of A, it is an element belonging to A. {p} is the subset of A consisting of the single element p. Every element of { r, q, p } belongs to A, so it is a subset of A (it is, in fact, A itself, so A is a subset of A). ■

The subset relation between sets is in some ways similar to the ' less than or equal to ($\leq$) ' relation between numbers. Both include equality as a possibility: we know that $8 \leq 8$, and also $A \subseteq A$

(every element of the set A on the left side of the $\subseteq$ symbol is certainly an element of the same set A on the right side of the $\subseteq$ symbol). Note that both symbols, $\leq$ and $\subseteq$, have horizontal bars under them to allow for the possibility that the object on the left of the symbol may be equal to the object on the right. This parallel among symbols can be extended to eliminate the possibility of equality by removing the horizontal bar: the symbol $3 < 5$ means 3 is less than 5, and $A \subset B$ means that A is a subset of B, but is not equal to B (sometimes we say A is a **proper subset** of B).

Similarities Between $\leq$ and $\subseteq$

Reflexivity: $a \leq a$ $A \subseteq A$

Anti-Symmetry: if $a \leq b$ and $b \leq a$, then $a = b$

if $A \subseteq B$ and $B \subseteq A$, then $A = B$

Transitivity: if $a \leq b$ and $b \leq c$, then $a \leq c$

if $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$

The first property, reflexivity, is more an observation than something which is generally useful, but the last two properties are quite useful. The second property, anti-symmetry, is often used to show that two sets are equal, by showing that each is a subset of the other.

Universal Sets and Complements

As we shall see, set theory provides a useful framework for discussing certain types of problems. However, often the items under discussion during these problems are limited: we might be discussing positive numbers, or presidents of the United States, or poker hands. In such situations, it is useful to have a **universal set** available, which may be loosely regarded as the 'smallest' set containing all the items under discussion. If we are discussing positive numbers, for instance, we don't want to worry about presidents or poker hands, and so it would be sensible to define U as a

universal set consisting of all possible positive numbers. It is then understood that all sets mentioned will be subsets of this universal set.

Unlike most of the concepts in mathematics, the concept of universal set is a little nebulous. For instance, if A = { Iowa, Illinois }, and B = { Illinois, Indiana }, possible universal sets are states or places beginning with the letter 'I'.

Once a universal set has been defined, it makes sense to talk about all the objects in the universal set which are not in a given set. The **complement** of the set A, written A', is $\{ x : x \in U, x \notin A \}$.

Example 3 - Let U be the universal set consisting of positive whole numbers. If A = { 1, 2, 3 }, then describe A' in two different ways, using both words and set-builder notation.

Solution: A' is the set of all positive whole numbers greater than 3. Using set-builder notation, this would be A' = { x : x is a positive whole number greater than 3 }. ■

Notice that the complement of a set is a relative concept, in that it needs a universal set in order to make sense. In Example 3, for instance, it is almost certain that we don't want to have to worry about hippopotamuses when considering the set A'!

The Connection Between Logic and Set Theory

Set theory has an extremely useful connection with logic. If P(x) is a proposition about the variable x (such as 'x is a number greater than 7'), then P = { x : P(x) is true } is the set of all things for which P(x) is a true statement. Similarly, we define Q = { x : Q(x) is true }.

If both P(x) and Q(x) are propositions and the sets P and Q are defined as above, then the logical statement $P(x) \Rightarrow Q(x)$ for all x supplies the same information as the set theory statement $P \subseteq Q$.

For example, if P(x) is the proposition that 'x is a number greater than 7', and Q(x) is the proposition that 'x is a number greater than 4', clearly $P(x) \Rightarrow Q(x)$. Notice that, if P = { x : x > 7 } and Q = { x : x > 4 }, then $P \subseteq Q$.

To prove the assertion, suppose first that $P(x) \Rightarrow Q(x)$ for all x. If y is an element for which P(y) is true (and therefore $y \in P$), since $P(y) \Rightarrow Q(y)$, then Q(y) is true, and so $y \in Q$. So $P \subseteq Q$.

Now suppose $P \subseteq Q$. If P(y) is false, then $P(y) \Rightarrow Q(y)$ automatically (remember from Chapter 1 that a false statement automatically implies any other statement). If P(y) is true, then $y \in P$, and so $y \in Q$, since $P \subseteq Q$. Therefore Q(y) is true, so $P(x) \Rightarrow Q(x)$ for all x.

Note: in each of the two preceding paragraphs, when we used the letter 'x' to denote a dummy variable, we then used the letter 'y' to denote a specific element. This was to prevent confusion as to which 'x' we were talking about, just as Mom might call James Sr. 'Jim' and James Jr. 'Jimmy' to avoid confusion.

The Null Set

The **null set** is the set which contains no elements. It is sometimes called the **empty set** -- if sets are lunchboxes, there's nothing in this lunchbox. The null set is written $\varnothing$.

One of the most common mistakes is to confuse the null set, $\varnothing$, with the number 0. As we shall see, there are certain similarities between the two objects, but they belong to different mathematical systems. It is as if one were to confuse the contents of an empty lunchbox (the null set) with the cost of purchasing those non-existent contents (zero).

The null set has an extremely important property: it is a subset of <u>every</u> set! While this appears surprising at first, its proof simply involves the logical negation of the statement: $\varnothing \subseteq A$. If it is false that $\varnothing \subseteq A$, then there must be some element in $\varnothing$ which is not an element of A. But this is a contradiction, because there are <u>no</u> elements in $\varnothing$ -- and once an assumption leads to a contradiction, that assumption must be false.

For example, the empty set is a subset of G, the set of all giraffes, for if it were not, there would be a giraffe in the empty set.

Section 2 - Unions and Intersections

We turn now to the question of how to combine two sets. Any interesting combination of two sets must itself be a set, just as the interesting ways to combine numbers (addition, subtraction, multiplication, and division) result in numbers. The two basic ways to combine sets are intersection and union.

The **intersection** of two sets A and B, written $A \cap B$, is the set of all elements common to both sets. If we define this using set-builder notation, $A \cap B = \{ x : x \in A \text{ and } x \in B \}$.

Example 1 - Let $A = \{ 1, 2, 3 \}$ and $B = \{ 2, 3, 4 \}$. Then $A \cap B = \{ 2, 3 \}$.

If A and B have no elements in common, then A and B are said to be **disjoint**. In this case, $A \cap B = \varnothing$.

Example 2 - If $A = \{ 1, 2, 3 \}$ and $B = \{ 4, 5 \}$, then A and B are disjoint, and $A \cap B = \varnothing$.

The other basic way of combining the two sets A and B, the **union** of A and B, written $A \cup B$, is the set of all elements which belong to either (or both) of the two sets. If we define this using set-builder notation, $A \cup B = \{ x : x \in A \text{ or } x \in B \}$.

Notice that the 'or' employed in defining the union of two sets is the inclusive 'or', which is also the 'or' of choice in logic. In day-to-day conversation, it is usually clear from context whether the inclusive or exclusive 'or' is intended, but all the definitions of mathematics use the inclusive 'or'.

Example 3 - Let $A = \{ 1, 2, 3 \}$ and $B = \{ 3, 4 \}$. Then $A \cup B = \{ 1, 2, 3, 4 \}$ and $A \cap B = \{ 3 \}$.

Example 4 - Let $A = \{$ rat, cat, cow $\}$ and $B = \{$ dog, rat, cat, pig, yak $\}$. Then $A \cap B =$

$\{$ rat, cat $\}$, and $A \cup B = \{$ rat, cat, cow, dog, pig, yak $\}$

In Example 3, although the number 3 appeared in both A and B, it is not listed twice in $A \cup B$. When one is asked the contents of a lunchbox, it is repetitive to say, "a ham sandwich, uh, a ham sandwich, a bag of potato chips, and an apple."

When more than two sets are involved, parentheses are used to indicate which operations should be performed first, just as they are in arithmetic. Suppose that A = { Alan, Betty, Carol }, B ={ Betty, David, Frank }, and C = { Alan, Carol, Edward }. Then $(A \cup B) \cap C$ is determined first by computing $A \cup B$ = { Alan, Betty, Carol, David, Frank }, and then $(A \cup B) \cap C$ = { Alan, Carol } . Notice that, if we were to compute $A \cup (B \cap C)$, we would first determine $B \cap C = \varnothing$, and then $A \cup (B \cap C)$ = { Alan, Betty, Carol } . Therefore, the expression $A \cup B \cap C$ must be parenthesized in order to be unambiguously determined.

Unlike arithmetic, where multiplication takes precedence over addition (so that, in the expression 2 + 3 x 4, the multiplication is performed first), there is no precedence hierarchy for the operations of union and intersection. Union and intersection are both **associative** operations, that is,

(1) $(A \cup B) \cup C = A \cup (B \cup C)$

(2) $(A \cap B) \cap C = A \cap (B \cap C)$

If only unions are involved, or intersections, it is not necessary to parenthesize, but if both operations appear in the same expression, parenthesization is imperative.

There is an analogy between numbers and sets. Observe that

(3) $A \cup \varnothing = A$

and

(4) $A \cap \varnothing = \varnothing$

If one regards the empty set as the set-theory analog of the number 0, then (3) is analogous to the arithmetic property a + 0 = a, and (4) to the arithmetic property a x 0 = 0.

We have encountered both unions and intersections in the preceding story. If we let L denote the set of all members of Lisa's animal-rights group who planned on attending the lobbying effort, and R the set of members who planned on raiding the laboratory, then $L \cup R$ is the set of all the activists, those members who planned to participate in at least one of the two activities. $L \cap R$ consists of the 'hard core' who planned to participate in both.

Many of the most interesting sets have only a finite number of things in them. If A is such a set, then N(A) denotes the number of things in A. If A = { Alice, Betty, Carlos }, then N(A) = 3. Notice that, although there are many sets A for which N(A) = 3, there is only one set for which N(A) = 0, and that is the empty set.

The Fundamental Counting Principle

In the preceding story, Pete makes use of an extremely important principle of counting.

If A and B are finite sets, then $N(A \cup B) = N(A) + N(B) - N(A \cap B)$

This valuable formula plays an important role in the study of probability, and so it is worth spending a little effort to verify it.

Pete's explanation was that, when we add N(A) to N(B), obtaining N(A) + N(B), we are counting everything that appears in $A \cap B$ twice. Therefore, if we wish to compare $N(A \cup B)$ with N(A) + N(B), we must realize that N(A) + N(B) counts everything in $A \cap B$ twice. So $N(A) + N(B) = N(A \cup B) + N(A \cap B)$, since elements that belong to A or B, but not both, are counted once on each side of the equation, and elements that belong to $A \cap B$ are counted twice on each side of the equation. Subtracting $N(A \cap B)$ from both sides now yields the Fundamental Counting Principle.

Example 5 - A survey of the 140 customers at an electronics supply store who owned either a HDTV or a camcorder revealed that 105 of them owned an HDTV and 28 owned both. How many owned a camcorder?

Solution: It is certainly possible to simply 'plug into' the Fundamental Counting Principle. Let H be the set of HDTV owners, C the set of camcorder owners. Then $N(H \cup C) = 140$. Since $N(H) = 105$ and $N(H \cap C) = 28$, we have $N(H \cup C) = N(H) + N(C) - N(N \cap C)$, so $140 = 105 + N(C) - 28 = 77 - N(C)$. Therefore $N(C) = 140 - 77 = 63$. ■

The Difference of Two Sets

Another set which is of interest is the **difference** $A \setminus B$, which is defined as the set of all things in A but not in B. That is, $A \setminus B = A \cap B'$, or, using set-builder notation, $A \setminus B = \{ x : x \in A \text{ and } x \notin B \}$. In Example 5 and 6 (below), we have used the idea of $A \setminus B$ before we had specifically defined it.

Example 6 - If A = { dog, pig, rat, cat, yak } and B = { rat, cat }, what is $A \setminus B$? What is $B \setminus A$?

Solution: $A \setminus B$ is the set of all things belonging to A but not to B, which is { dog, pig, yak }. $B \setminus A$ is the set of all things belonging to B but not to A, which is $\varnothing$. ■

We have mentioned the analogy between arithmetic and set theory. There is a temptation to regard the set-theoretic difference as being analogous to subtraction, because when we construct the set $A \setminus B$ we are 'taking away' the things in B from A. However, this analogy should not be taken too far. For example, if the number b is larger than the number a, the difference a - b is negative. If the set B is 'larger' than the set A (i.e. $B \supseteq A$), the set-theoretic difference $A \setminus B = \varnothing$, the empty set. This situation occurs in the second part of Example 7. There is no concept in set theory analogous to negative numbers.

Section 3 - The Nature of Infinity

We start this section by discussing a couple of examples which indicate some of the counterintuitive properties of infinite sets.

Hilbert's Hotel: Where They Never Say, "Sorry, We're Full Up"

In an ordinary hotel (one with 20, 200, or 2000 rooms), once every room is full, a late-arriving guest cannot obtain an unoccupied room. David Hilbert, one of the greatest mathematicians of the early twentieth century, concocted the following example, which is called Hilbert's Hotel in his honor.

Imagine a hotel with an infinite number of rooms, one room for every positive whole number. Suppose that every room is occupied, and a guest arrives looking for an empty room. Rather than turn the guest (and the accompanying cash) out in the cold, the hotel manager simply moves the occupant of Room 1 to Room 2, the occupant of Room 2 to Room 3, the occupant of Room 3 to Room 4, etc. After all this moving is done, every former occupant still has a room, but Room 1 is now free, and the new guest moves in!

Later in the evening, the hotel is still full, and an infinite number of new guests arrive. Not to worry. The manager gives each new guest a number, just as one would at a store where you have to wait in line. While the guests are in the bar, the manager moves the occupant of Room 1 to Room 2, the occupant of Room 2 to Room 4, the occupant of Room 3 to Room 6, etc. All the previous occupants are now in even-numbered rooms. The odd-numbered rooms are now vacant, so new Guest 1 moves into Room 1, new Guest 2 moves into Room 3, new Guest 3 moves into Room 5, and so on. Everyone is happy (especially the stockholders)!

Money for Nothing

Suppose you have ten friends and no money. You talk five of them into loaning you $1 each, and you loan $1 each to the other five. You have created five debtors and five creditors, but your net worth hasn't changed -- you still have no money.

Now let's suppose you still have no money, but you now have infinitely many friends, whom we will number Friend 1, Friend 2, etc. You talk each of your odd-numbered friends into loaning you $1, and you loan $1 to each of your even-numbered friends. You now decide to square accounts with everyone, but you do it in an unusual way.

Step 1 - Collect the $1 each owed to you by Friends 2 and 4, and pay $1 back to Friend 1. You gain $1, which you deposit in the bank.

Step 2 - Collect the $1 each owed to you by Friends 6 and 8, and pay $1 back to Friend 3. You gain $1, which you deposit in the bank.

Step 3 - Collect the $1 each owed to you by Friends 10 and 12, and pay $1 back to Friend 5. You gain $1, which you deposit in the bank.

You continue this process. After you have finished, you are all square with everyone -- you have paid back every dollar you owe, and no one owes you any money. You also have infinitely many dollars in the bank! Money for nothing – except a lot of work!

Countable Sets

The above two examples illustrate that the rules which govern manipulations with infinite sets must be very different from the ones with which we are familiar. When we first learn the properties of numbers, we learn about "three-ness" by discovering that all sets with three elements in them can be placed in one-to-one correspondence with each other. Moreover, this idea of one-to-one correspondence can be used to define the ideas of 'greater than' or 'less than'. Suppose we have two sets, one consisting of children, and one consisting of cookies. If we can give one cookie to each child and at the end of doing so each child gets a cookie and no cookies are left over, the set of children and the set of cookies have exactly the same number of items in them. Mathematicians say that these two sets have the same **cardinality**. If we have cookies left over, the set of cookies has

greater cardinality than the set of children. Similarly, if some children don't get cookies, then the set of children has greater cardinality than the set of cookies.

The example of Hilbert's Hotel illustrates that, for infinite sets, one-to-one correspondence will not enable us to distinguish 'less than' from 'precisely the same number as.' Indeed, the concepts of 'less than' and 'precisely the same number as' require careful definition in the context of infinite sets. For instance, the set { 2, 3, 4, ... } appears to have 'fewer' elements than the set { 1, 2, 3, ... }. However, they can be placed in one-to-one correspondence by means of the following arrangement.

2 3 4 ....

⇳ ⇳ ⇳

1 2 3 ....

Therefore, the sets { 2, 3, 4, ... } and { 1, 2, 3, ... } have the same cardinality.

All of the infinite sets we have examined are subsets of the positive integers, and they can all be put in one-to-one correspondence with the integers (and also with each other). Such sets are called **countable**. This raises a question: is it possible to describe an infinite set S of people who cannot be housed in Hilbert's Hotel? This question was originally answered by Georg Cantor (who suffered a mental breakdown in the process – you hear about situations such as this with artists but rarely with scientists).

An Uncountable Set

To describe a set S of people who cannot be housed in Hilbert's Hotel, imagine that each individual in the set S has a name which can be written out using an infinite sequence of letters and blanks. The first letter of Elvis Presley's name is E, the second is L, ... , the twelfth is E, the thirteenth is Y, and all the other letters, such as the eighty-first or three-millionth or

nine-quadrillionth, is blank. We will show that Hilbert's Hotel is not large enough to contain the set of people with all possible infinitely-long names.

To see this, imagine that Elvis Presley is in Room 1, Jimmy Hoffa is in Room 2, Amelia Earhart is in Room 3, Judge Crater is in Room 4, and every person in the set has been placed in a room. To show that this cannot have been done, we shall exhibit the name of a person who cannot have been given a room in the hotel. We shall construct that person's name using the initials of the King, the letters E and P.

Let's look at the guest register.

Room 1 - ELVIS PRESLEY (the first letter is underlined)

Room 2 - JIMMY HOFFA (the second letter is underlined)

Room 3 - AMELIA EARHART (the third letter is underlined)

Room 4 - JUDGE CRATER (the fourth letter is underlined)

... (and so on)

We construct our mystery guest's name one letter at a time, according to the following rule: if the N^{th} letter of the inhabitant of Room N *is* an E, then the Nth letter of our mystery guest's name will be a P. If the N^{th} letter of the inhabitant of Room N is *not* an E, then the N^{th} letter of our mystery guest's name will be an E. We have underlined the first letter in the first name in the diagram above, the second letter in the second name, etc.

Since Elvis Presley is in Room 1, and the first letter of his name is E, the first letter of our mystery guest's name is P. Since Jimmy Hoffa is in Room 2, and the second letter of his name is I (not E), the second letter of our mystery guest's name is E. Since Amelia Earhart is in Room 3, and the third letter of her name is E, the third letter of our mystery guest's name is P. Since Judge Crater

is in Room 4, and the fourth letter of his name is G (not E), the fourth letter of our mystery guest's name is E. So the mystery guest's name is PEPE ... .

In what room is our mystery guest? Not in Room 1, since the first letter of his name (P) is different from the first letter (E) of Elvis Presley, the occupant of Room 1. Not in Room 2, either, since the second letter of his name (E) is different from the second letter (I) of Jimmy Hoffa, the occupant of Room 2. Our mystery guest cannot be found in Room 3, since the third letter of his name (P) is different from the third letter (E) of Amelia Earhart, the occupant of Room 3. Nor is he, or she, or it, in Room 4, since the fourth letter of his, or her, or its name (E) is different from the fourth letter (G) of Judge Crater, the occupant of Room 4. In fact, our mystery guest is nowhere to be found in Hilbert's Hotel, since the N^{th} letter of his name differs from the N^{th} letter of the name of the guest in Room N!

This extremely ingenious proof is known as Cantor's Diagonal Argument, and is one of the standard techniques used in proving results about infinite sets.

Any infinite set which cannot be put in one-to-one correspondence with the integers, such as the set of all infinitely-long names, is called **uncountable**. The set of real numbers, which is essentially the set of all infinitely-long names using the alphabet of digits 0,1,2,…,9 rather than the alphabet of Roman letters A,B,…,Z, is uncountable.

Chapter 8 – The Chinese Restaurant Principle; Combinatorics

Introduction

The need for counting techniques arises because it is often necessary to count very large numbers of items. After all, if there are only five or so items in a set, one just counts them. On the other hand, if there are several hundred items in a set, counting them is likely to take some time, and there is the possibility of a mistake. It may even happen that the number of items we wish to count is so large that we would never be able to do it directly.

We have already discussed one of the most important counting procedures, the Fundamental Counting Principle, which is summarized in the formula

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

In this chapter, we shall investigate counting techniques based on successive choices. Many decisions can be viewed as a procedure based on successive choices. When we order dinner in a restaurant, we successively choose an appetizer, an entree, and a dessert. When a basketball coach selects a starting line-up, he chooses the center, the two forwards, and the two guards.

Mathematics started with the problem of counting, and counting problems still underlie many of the most important areas of mathematics, such as probability, statistics, and decision theory. The techniques we establish in this section will reappear frequently throughout the remainder of this book.

The Chinese Restaurant Principle

The Chinese Restaurant Principle Pete referred to in the preceding story derives its name from the following situation. A staple of Chinese restaurants has been the fixed-price meal, consisting of one appetizer chosen from a list of many different appetizers, and one main course from a similar list. It is traditional for the menu to present these choices in the form of two columns.

Menu

Column A	Column B
Won Ton Soup	Sweet and Sour Pork
Spareribs	Lemon Chicken
Rumaki	Oyster Beef
Egg Roll	Moo Goo Gai Pan
Dumplings	Shrimp with Lobster Sauce
	Mixed Fried Noodles
	Pressed Duck

For a fixed price, the diner gets to choose one dish from Column A, and one from Column B. An obvious question now arises: how many different possible fixed-price meals can one have?

We must first agree that two meals are different if they have either different appetizers, or different main courses, or both. There is a straightforward way to write out all the different possible meals. First write out all the different meals with Won Ton Soup as the appetizer, then all the different meals with Spareribs as the appetizer, etc. We do this below, abbreviating with dots (...) in order to simplify the task.

Won Ton Soup - Sweet and Sour Pork

...

Won Ton Soup - Pressed Duck

This would give us seven different meals with Won Ton Soup as the appetizer.

Spareribs - Sweet and Sour Pork

...

Spareribs - Pressed Duck

This would give us seven more meals, all different from the ones with Won Ton Soup as the appetizer. Similarly, there would be seven different meals with Rumaki as the appetizer, seven different meals with Egg Roll as the appetizer, and seven different meals with Dumplings as the appetizer. This makes a total of $7 + 7 + 7 + 7 + 7 = 5 \times 7 = 35$ different meals.

Notice that this number is obtained by multiplying the number of different possible appetizers (5) by the number of different possible main courses (7). Although we counted the different meals by listing them by appetizer, we could certainly have first counted all the different meals by counting the different meals in which Sweet and Sour Pork was the main course (5 different meals), then the different meals in which Lemon Chicken was the main course (5 different meals), etc. Counting this way would have given us $5 + 5 + 5 + 5 + 5 + 5 + 5 = 7 \times 5 = 35$ different meals, just as before.

It is important to realize that the choice of main course is **independent** of the choice of appetizer. That is, once we have chosen an appetizer, we are perfectly free to choose any of the possible main courses (or vice-versa, if we decide to choose the main course first). If the choice of one depends on the choice of the other (for instance, if the restaurant has a rule preventing you from ordering a poultry main course when you have chosen Rumaki as an appetizer, or if you are not allowed to

order two pork dishes, such as Spareribs and Sweet and Sour Pork), the choices are no longer independent, and the counting procedure we have established above is no longer valid.

The Chinese Restaurant Principle (even professional mathematicians use that phrase, possibly because they often congregate in such establishments) states that, if two choices can be made independently, the number of ways of making both choices is the product of the number of ways of making each choice separately. In the above example, "making both choices" corresponds to selecting a meal. The number of ways of making the appetizer choice is 5, the number of ways of making the main course choice is 7, and so the number of ways of making both choices is $7 \times 5 = 5 \times 7 = 35$.

The Chinese Restaurant Principle (2 Choices)

If two choices can be made independently, and there are p ways of making the first choice, and q ways of making the second choice, than the number of ways of making both choices is pq.

Example 1 - A store is selling mix-or-match outfits consisting of slacks and sport shirts. If 4 styles of slacks are available and 9 styles of shirts, how many different outfits are available?

Solution: Since the choices are independent, the number of possible choices is $4 \times 9 = 36$. ■

Since no meal is complete without dessert, let's see what happens to the Chinese Restaurant Principle when we decide to have dessert as well. Suppose that the menu we were using before (in Fig. 8-1) offers a complete dinner, including a choice of one appetizer from Column A (5 choices), one main course from Column B (7 choices), and one dessert from Column C (3 choices). How many different complete dinners are possible?

We do not need to perform a lengthy analysis to solve this one. A complete dinner can be viewed as the result of two independent choices: the first choice being the appetizer-main course combination, which we already know can be made in 35 different ways, and the second choice

being the dessert, which can be made in 3 different ways. The Chinese Restaurant Principle thus enables us to conclude that there are 35 x 3 = 105 different complete dinners. Notice that 35 x 3 = 5 x 7 x 3 .

As a result, we can now extend the Chinese Restaurant Principle to any number of independent choices. If there are several different independent choices to be made, the number of different ways of making all choices is equal to the product of the number of ways of making each choice separately. We can state this formally as follows.

The Chinese Restaurant Principle (p Choices)

Suppose that we have p independent choices to make, and the first choice can be made in N_1 different ways, the second in N_2 different ways, ... , and the p^{th} choice in N_p different ways. Then the number of different ways of making all choices is

$$N_1 \times N_2 \times ... \times N_p$$

It is easy to envision the Chinese Restaurant Principle as simply being the number of different p-course meals, where there are N_1 different first courses, N_2 different second courses, ..., and N_p different last courses.

Example 2 - Mammoth Tours offers fixed vacation packages from Los Angeles to either Honolulu, Acapulco, the Bahamas, or Puerto Rico. One can travel by boat or plane, and can stay at a First Class, Deluxe, or Economy Class hotel. How many different vacation packages are offered?

Solution: Since the choices of destination, method of travel, and accommodations are independent, by the Chinese Restaurant Principle there are 4 x 2 x 3 = 24 different vacation packages. ■

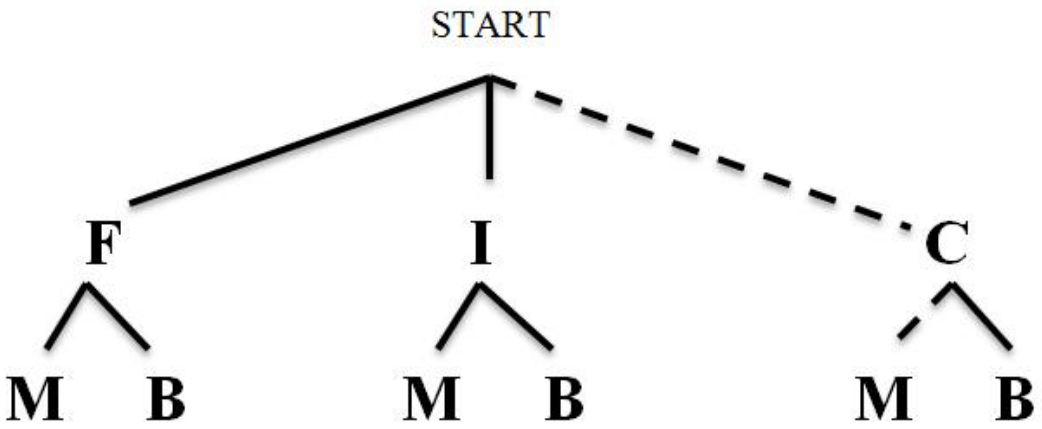
When the choices are not independent, one can often compute the total number of available choices by computing the number of possible choices, and subtracting the number of excluded choices.

Example 3 - In the previous example, assume that if one travels by boat, one must stay in a First Class hotel. Under these restrictions, how many vacation packages are offered?

Solution: The number of excluded choices can be computed by the Chinese Restaurant Principle. There are four excluded destinations, one excluded mode of transportation (plane), and two excluded types of accommodation, so the number of excluded packages is 4 x 1 x 2 = 8. Therefore, the total number of packages offered is 24 - 8 = 16. ■

Trees

Trees offer a convenient way to see sequential choices. Suppose that we are planning an evening out, and can have dinner at either a French (F), Italian (I), or Chinese (C) restaurant, and follow it with either a movie (M) or a basketball game (B). The tree constructed to depict the available choices is



A path from top to bottom which goes from a choice at one level to a choice directly below it is referred to as a **branch** of the tree; each branch of this tree represents a different evening on the town. In the above example, the branch consisting of a Chinese dinner followed by a movie has been shown as a dashed line.

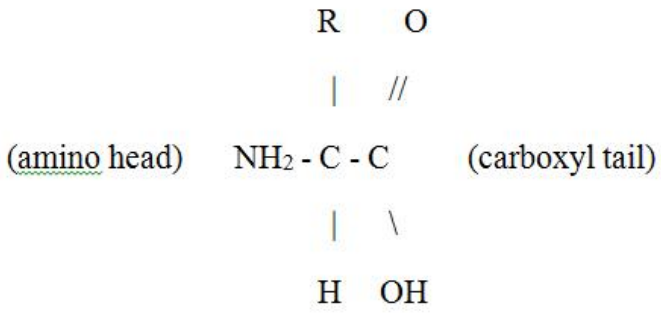
The Chinese Restaurant Principle is the single most important principle in **combinatorics**, which is the branch of mathematics that deals with counting problems. However, its use is not confined simply to mathematics. Knowledge of the Chinese Restaurant Principle helped to decipher the genetic code!

By the late 1950s, biologists had discovered that DNA formed a blueprint for constructing proteins, the important chemicals that are the basis of all life. DNA consists of a very long string of four substances, adenine (A), cytosine (C), guanine (G), and thymine (T). Using these four letters, a section of DNA might look something like ... AACTGAGATTCCAA ... .

Proteins are constructed from smaller sub-units called amino acids. Even though over 200 amino acids are known, only 20 of them are used to form proteins. Each amino acid has a "head" and a "tail"; the "head" of any amino acid fits neatly into the "tail" of any other.

The Structure of an Amino Acid

The chemical structure of an amino acid is



Amino acids form proteins by linking up head-to-tail, with the upper 'jaw' (the O) and the lower 'jaw' (the OH) of the carboxyl tail clamping down on the amino head of the next amino acid in the line. The R is called the 'side chain', and varies from one type of amino acid to another.

It was known that cells somehow "read" the string of letters in DNA to determine which amino acids should be strung together, and in which order, in order to make proteins. The question that

puzzled the biologists in the late 1950s was: what was the correspondence between the letters in the DNA and the amino acids used to make proteins?

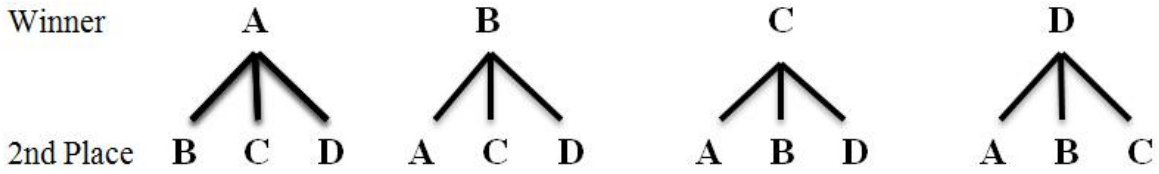
This is where the Chinese Restaurant Principle came into play. It was clear that a single letter of DNA did not correspond to an amino acid; after all, there were only four letters and 20 amino acids. Was it possible that each two-letter combination, such as AA or CT or GA, corresponded to particular amino acids? Since the Chinese Restaurant Principle shows that there are only 4 x 4 = 16 two-letter words that can be made from the DNA letters of A, C, G, and T, there would not have been enough two-letter words to correspond to the 20 amino acids. When one considered three-letter words, such as CAT or GCG, there were now 4 x 4 x 4 = 64 different possible words.

An ingenious series of experiments, taking several years to complete, did indeed determine that each three-letter word, which is called a **codon**, corresponded to a particular amino acid. Since there are 64 such words and 20 amino acids, some amino acids are represented by several different words. For example, cysteine is represented by only one word (TGT) and tyrosine by two words (TAT and TAC). Arginine and leucine have the most words corresponding to them; they are represented by six words each. The genetic code is another supporting argument for the theory of evolution, as the genetic code is the same in all known forms of life, from bacteria to human beings.

Section 2 - Permutations

It is often much easier to describe a concept in mathematics (or anywhere else) by starting out with an example. With that in mind, let us imagine that there are four entrants in a race: Al, Bob, Charlie, and Dave. In how many ways can the first two places be awarded, assuming there are no ties?

Drawing a tree makes this problem fairly simple. We let A stand for Al, B for Bob, etc.



Since there are 4 x 3 = 12 branches of the tree, there are 12 different ways of choosing the first two places.

Even though multiplication is involved, this is not exactly an application of the Chinese Restaurant Principle, which deals with **independent** choices. The choices here are not independent; different winners result in different possible second-place choices, even though the number of possible second-place choices is the same no matter who wins. When we look at the tree for this problem, we can see that there are the same number of branches at each junction on the same level, but the branches are not the same.

We now extend this idea to a slightly more complicated problem. Suppose we have eight entrants in a race, and wanted to find out the number of different possible ways of awarding the first three places, assuming no ties. We can see that, as before, the number of possible ways of choosing the first two finishers is 8 x 7 = 56. For each of these 56 cases, there will be six different choices for third place. Again, the possible third-place finishers will depend on who actually finished first and second, but the number of possible third-place choices, six, will always be the same no matter who finished first and second. Consequently, there are 8 x 7 x 6 = 336 different ways of awarding the first three places.

We are now in a position to generalize. Suppose we have n entrants in a race, and wish to award the first k places. The winner can be chosen in n ways, the second-place finisher in n-1 ways, the third-place finisher in n-2 ways, ... , and the k^{th} place finisher in n - (k-1) = n - k + 1 ways. Therefore, the total number of possible ways of awarding the first k places is given by
n x (n-1) x (n-2) x ... x (n-k+1) ways.

Permutations

A **permutation of n things taken k at a time** is an ordered selection of k objects from an original collection of n objects. The number of such permutations is referred to as P(n,k), and is given by the formula

$$P(n,k) = n \times (n-1) \times (n-2) \times \ldots \times (n - (k-1))$$

$$= n \times (n-1) \times (n-2) \times \ldots \times (n-k+1)$$

Notice that a permutation of n things taken k at a time can be viewed as two successive choices. The first choice is which k things to take, and the second choice is in what order to take them. When we were looking at the problem of figuring out the number of ways to award first and second place in a race among Al, Bob, Charlie, and Dave, the permutation Charlie-Bob consisted of the choice of Bob and Charlie to receive the prizes, and then the choice of Charlie to finish first and Bob to finish second.

In particular, the number of permutations of n things taken n at a time, P(n,n) =

$n \times (n-1) \times \ldots \times (n - n + 1)$, is the product of the integers from n down through 1 or, alternatively, the product of the integers from 1 up to n. It is the number of ways of awarding the first n places in a race with n entrants. Alternatively, it is the number of possible ways that the n entrants can finish the race (assuming no ties). Mathematicians call this number "n factorial", and it is written as $P(n,n) = n!$

Example 1 - A standard deck of cards contains 52 cards. Use factorial notation to express the number of ways the deck can be shuffled.

Solution: Shuffling a deck of cards creates a permutation of 52 cards taken 52 at a time, so there are 52! possible ways to shuffle the deck. ▉

Using factorial notation, we can write

$$P(n,k) = n! / (n-k)!$$

Example 2 - From a collection of 10 songs, a deejay will select 4 and play them. In how many ways can four songs be selected and played, if the same songs played in a different order counts as a different selection?

Solution: This is clearly a permutation problem, and so the answer is $P(10,4) = 10 \times 9 \times 8 \times 7 = 5{,}040$. ■

A key word in dealing with the concept of permutations is 'order'. Another phrase often used to describe a permutation of n things taken k at a time is an **ordered selection of n things taken k at a time**. Even though we will not start a thorough discussion of unordered selections until the next section, this is a good time to point out the difference between the two. As one might expect, in an ordered selection, the order in which the choices are made is extremely important, whereas in an unordered selection, that order is irrelevant. In Example 2, for instance, if the only thing that matters is which songs are selected (and the order in which they are played does not matter), that would constitute an unordered selection.

Returning to the Chinese restaurant, if one orders a meal consisting of won ton soup, oyster beef, and pineapple sherbet, one presumably wants the soup first, the beef second, and the sherbet at the end. This is an ordered selection. However, if one goes to the restaurant with some friends and orders several different entrees to be shared, presumably the order in which those entrees arrive is unimportant. This is an unordered selection.

Another example of the difference between an ordered selection and an unordered one occurs when we look at a race, such as the 100-meter dash in the Olympics. Let's suppose there are 10 entrants. If we only wish to know who the three medalists are, this would constitute an unordered selection of 10 things taken three at a time. If we wish to know who won the gold, who won the

silver, and who won the bronze, this would constitute an ordered selection of 10 things taken three at a time.

We now have two different counting principles at our disposal: the Chinese Restaurant Principle, and the permutation formula. They can obviously be combined, as Example 4 indicates.

Example 3 - The Smith family consists of three men and four women. In how many different ways can they line up for a photograph with the men standing next to each other?

Solution: We can analyze this using a combination of the two principles. By themselves, we can line the four women up in 4! = 24 different ways. Similarly, we can line the three men up in 3! = 6 different ways. If we consider the Smiths as lining up in positions 1 through 7 (left to right, for example), we can place the leftmost man in any one of the first five positions. This determines where all the men are. We then fill the remaining positions with women.

We therefore have three independent choices to make: how to line up the women, which can be done in 24 different ways; how to line up the men, which can be done in 6 different ways, and where to place the leftmost man, which can be done in 5 ways. There are therefore 24 x 6 x 5 = 720 different ways of lining up the Smiths for a photograph. ▉

Factorials get awfully large awfully fast, as can be seen from looking at the following numbers.

1! = 1 2! = 2 3! = 6 4! = 24 5! = 120

6! = 720 7! = 5,040 8! = 40,320 9! = 362,880 10! = 3,628,800

Many newspapers carry a feature called a cryptogram, which consists of a passage or quote printed in a substitution code. A substitution code is obtained from a permutation of all the letters of the alphabet.

A Substitution Code

If the actual letter is A B C ... X Y Z

↓ ↓ ↓ ↓ ↓ ↓

Substitute B C D ... Y Z A

The message "I LOVE YOU" becomes "J MPWF ZPV". It loses something in substitution.

Example 4 - How many different substitution codes are there?

Since each permutation of the alphabet results in a different substitution code, each substitution code is a permutation of the 26 letters in the alphabet taken 26 at a time. So there are 26! substitution codes (we have counted the identity substitution, in which all letters remain the same). 26! is approximately equal to 4 followed by 26 zeroes. A computer that could examine a billion such codes every second would have to have been started approximately at the time of the Big Bang in order to have finished by now. Of course, this "brute force" approach to cryptography is usually avoided in favor of subtler techniques of decoding messages, which rely on characteristics of the language, such as frequencies of letters, and knowledge of which pairs of letters cannot occur in the language, in order to reduce the problem to manageable proportions. This is another use of the Chinese Restaurant Principle, where one eliminates possible options to reduce the total number of choices.

Permutations, while interesting in themselves, also play a vital role in the construction of other counting principles, as we shall see in the next section.

Section 3 - Combinations

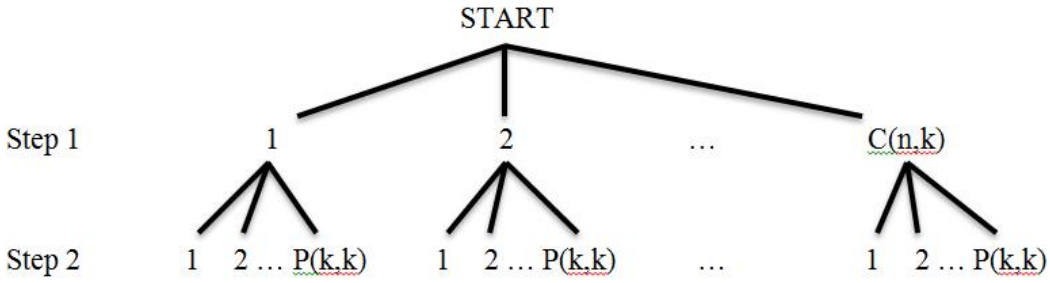
If you have ever gone to a Chinese restaurant with some friends, and shared an assortment of dishes, you have made an **unordered selection**, which mathematicians call a **combination**.

A **combination of n things taken k at a time** is an unordered subset consisting of precisely k items, chosen from an initial collection of n items. To fill in a ticket (which costs $1) in the

California Lottery (before they introduced the concept of a Mega Number), one simply chooses 6 numbers between 1 and 51. Since the order in which one selects the numbers does not matter (all you have to do is match the numbers randomly chosen at the drawing), this is known as an unordered subset. A California Lottery ticket used to be a combination of 51 things taken 6 at a time. Similarly, if you and your friends choose 4 different entrees on a menu from a list of 10 possible items, this is a combination of 10 things taken 4 at a time. C(n,k) is the standard notation for the number of combinations of n things taken k at a time. C(51,6) represents the number of different California Lottery tickets. Another way of looking at C(51,6) is that it is the number of different subsets of 6 items, taken from a collection of 51 items.

The standard way to compute C(n,k) is to use what we know about permutations. As we observed in the previous section, there is a two-step procedure involved in selecting a permutation of n things taken k at a time: first, we choose the k things, and second, we select an order for those k things.

We compute P(n,k) by drawing a tree. The first level of the tree consists of the combination of n things taken k at a time. The second level consists of arranging the things we selected. Therefore, if we wanted to compute P(n,k), we make two choices: Step 1 - choose a combination of n things taken k at a time (which can, by definition, be done in C(n,k) ways), and then Step 2 - arrange the selected k things in order. This last can be done in P(k,k) = k! ways. The tree therefore looks like



The total number of branches in the tree is P(n,k), and so

$$P(n,k) = C(n,k) \times P(k,k) = C(n,k) \times k!$$

Therefore $C(n,k) = P(n,k) / k! = n! / (k! \times (n-k)!)$.

Combinations

A combination of n things taken k at a time is simply a subset of k things taken from a universe consisting of n things. The number of possible combinations of n things taken k at a time is denoted by C(n,k), and is computed by the formula

$$C(n,k) = n! / (k! \times (n-k)!)$$

The combination formula completes our arsenal of counting procedures. The Big Three of counting consists of the Chinese Restaurant Principle, the formula for P(n,k), and the formula for C(n,k). While each is important, there seem to be more applications of C(n,k) than of P(n,k), as in many situations in the real world the order in which the choices are made is unimportant -- what matters is the actual choices that are made.

Example 1 - A basketball team consists of 12 players. At any moment, 5 of these are actually on the floor, playing. How many different squads can actually be out on the floor?

Solution: Since it doesn't matter in what order the players went on the floor, the number of different squads that can be out on the floor is $C(12,5) = 12! / (5! \times 7!) = 792$. ■

Notice that C(n,n), which is the number of subsets of n items from a collection of n items, must be 1 -- after all, the only such subset consists of <u>all</u> of the n items. According to the formula, $C(n,n) = 1 = n! / (n! \times 0!) = 1 / 0!$. Therefore, to make the formula work out correctly, we must define

$$0! = 1$$

One of the "games" mathematicians play is to derive mathematical relationships, and then see if those relationships reveal an underlying truth. Note that

$$C(n,n-k) = n! / ((n-k)! \times (n - (n-k))!)$$

$$= n! / ((n-k)! \times k!)$$

$$= C(n,k)$$

Suppose that we have a team consisting of n members, only k of whom can play at a time, while the rest of the team sits on the bench (for example, the basketball team of Example 1 has 12 men on a squad, only 5 of whom can play at a given moment). The number of different playing contingents, C(n,k), is the same as C(n,n-k), the number of different bench-sitting squads. When we think about it, this is not so surprising, as there is clearly a one-to-one correspondence between playing contingents and bench-sitting squads: given a playing contingent, the corresponding bench-sitting squad is precisely those team members on the bench! This observation can be used in reverse; since the number of playing contingents is precisely the same as the number of bench-sitting squads for the reason given above, this shows that C(n,k) = C(n,n-k) without going through any mathematical manipulations.

Example 2 - A poker hand consists of 5 cards, dealt from a 52 card deck. How many different poker hands are there? If a joker is added to the deck, how many different poker hands are there?

Solution: Since it doesn't matter in what order the cards are dealt, each poker hand is a combination of 52 things taken 5 at a time. There are $C(52,5) = 52! / (5! \times 47!) = 2,598,960$ different hands. With a joker, there are $C(53,5) = 53! / (5! \times 48!) = 2,869,685$ different hands. ■

Example 3 - A full house in poker consists of three cards of a specified rank (such as 5), and two cards of another specified rank (such as king). How many different full houses are there?

Solution: A full house consisting of three 5s and two kings is sometimes said to contain 5s over kings. There are three independent choices to be made to construct a full house. The first choice is what type of full house (such as 5s over kings); this can be done in $P(13,2) = 156$ different ways. The second choice is which specific three of the same rank (such as the 5s of clubs, diamonds, and

spades) are in the hand; this can be done in C(4,3) = 4 ways. Finally, the third choice is which specific two cards of the same rank (such as the kings of diamonds and hearts) are in the hand; this can be done in C(4,2) = 12 ways. Therefore, by the Chinese Restaurant Principle, the total number of full houses is 156 x 4 x 12 = 3,744. ∎

The Binomial Theorem

The combination symbol C(n,k) is ubiquitous in mathematics. It shows up in the Binomial Theorem, which was Isaac Newton's first major step on the road to eventual greatness. The Binomial Theorem is written

$$(x + y)^n = C(n,0)\, x^n + ... + C(n,k)\, x^{n-k}\, y^k + ... + C(n,0)\, y^n$$

It can be proved using algebraic methods and a lot of elbow grease, but there is a more elegant method using combinations. The n^{th} power of x+y is the product of n factors of x+y.

$$(x + y)^n = (x + y) \; (x + y) \; ... \; (x + y)$$

1st factor 2nd factor n^{th} factor

If one were to multiply the expression on the right out, there would be a total of 2^n expressions, each of which consists of a string of n letters, each of which is either an x or a y. One such string would be xxx ... x, consisting of the result of multiplying x in the 1st factor by x in the 2nd factor, ... , by x in the n^{th} factor. This is, of course, abbreviated as x^n . Similarly, another such string would be xyxyyy ... y, which is obtained by multiplying the x's in factors 1 and 3 by y's in all other factors. This string is abbreviated $x^2\, y^{n-2}$. Another way to obtain an $x^2\, y^{n-2}$ is to multiply x's in factors 1 and n, and y's in all other factors.

It can therefore be seen that $(x + y)^n$ is a sum of expressions of the form $x^k\, y^{n-k}$. A typical such expression is obtained by choosing k x's from k of the n factors, and y's from the remaining n-k factors. In how many ways can we construct expressions of the form $x^k\, y^{n-k}$? Obviously, each

choice of k numbers from the integers 1 through n determines such an expression -- we simply write down an x from each factor which was one of the k chosen numbers, and a y from all of the other n-k factors. For instance, if n=6, k=2, and the chosen combination consisted of the numbers 3 and 5, the corresponding expression would be yyxyxy -- note that the x's are in positions 3 and 5, and the y's occupy all other locations.

But how many ways can we make such a choice of k numbers from the integers 1 through n? Precisely C(n,k) ways! You might take an algebra textbook and look up the proof of the Binomial Theorem. It is highly computational. Worse, it takes forever!

Mathematicians, like almost everyone else, have an appreciation for elegance. Hopefully, you may acquire a taste for the elegant in mathematics, because it will certainly help you appreciate some of the pleasures that lurk within it.

Which Counting Principle Should Be Used?

Problems sometimes arise when one tries to decide whether to use the Chinese Restaurant Principle, the permutation formula, or the combination formula. Here are a few guidelines.

1) When in doubt, try drawing a tree, or a portion of it.

2) If choices are "used up" -- in other words, if selection of an item precludes its being selected again, the choices are not independent, and the permutation or combination formula will probably be required.

3) Permutations are involved when the order of the choices made is critical, but combinations are involved when order is irrelevant. To decide whether order is important or not, choose two items. Does flip-flopping the order of the choice of these two items create a different situation?

Chapter 9 - Probability and Expectation

Introduction

Probability is a mathematical tool which has two primary functions. The first, and most straightforward, is to summarize existing information. The second, which is by far the more interesting, is to make predictions about the future. The crystal ball presented by probability is a little clouded, however, for the future that it uncovers is not a future of individual events, but a future of long-term averages. The correct use of these long-term averages allows both individuals and society to plan more sensibly for the future.

Most of us have at least a passing acquaintance with probability through the weather reports. We have all heard the weatherman make a prediction such as: the probability of rain tomorrow is 80%. Intuitively, we know that this means that it's pretty likely that it will rain tomorrow. If we were to analyze this more thoroughly, we would conclude that what the weatherman means is this: if we were to keep a running total of what the weather is actually like on all the days when the weatherman says there is an 80% probability that it will rain, we would expect that it would rain on 80% of those days, but it would not rain on 20% of them.

We use the weatherman's prediction to help us make decisions. When the probability of rain the next day is 80%, we strongly consider taking an umbrella along with us, and we are very unlikely to schedule a picnic. Yes, the weatherman could be wrong -- the day could be nice and sunny, and we would regret looking ridiculous as we carry our umbrella, and we might also regret not having a picnic on such a beautiful day. On the other hand, it would be a lot worse if it rained and we didn't have an umbrella and got soaked to the skin, and of course everyone knows what a disaster rain is when you have a picnic. We have no guarantee of making the right decision, but that 80% number is pretty convincing, so we take the umbrella and cancel the picnic. Of course, if the weatherman

said that the probability of rain was 100%, we'd have no problem deciding, because it is certain to rain. Similarly, if the weatherman said that the probability of rain was 0%, rain would have no chance of occurring, so the umbrella goes back in the closet and the picnic is on.

Probability, as we see it in predictions of rain, is a number which represents the relative frequency with which rain will occur: from a low of 0%, when rain will never occur, to a high of 100%, when rain will definitely occur. The higher the number, the more likely the rain.

Section 1 - Theoretical and Empirical Probabilities; Sample Spaces

Summer days in Los Angeles belong to three basic types: sunny, rainy, and cloudy. We'll assume that these three types of days are defined to be non-overlapping; a day is either sunny, rainy, or cloudy. To make sure that the types of days are non-overlapping, we might define a day to be rainy if any rain falls, cloudy if no rain falls but a cloud appears in the sky, and sunny otherwise. The weather forecast for the next day might be a 20% probability of rain, a 30% probability of sunshine, and a 50% probability that it will be cloudy.

The mathematical terminology that is used in the above example is that observing the next day's weather constitutes an **experiment**. The three possible **outcomes** to the experiment are sunny, which we shall abbreviate as S, cloudy (C), and rainy (R). The three possible outcomes, which must be non-overlapping and cover all possibilities, constitute the **sample space** of the experiment.

Notice that the probabilities assigned to the three outcomes add to 100%: 20% + 30% + 50% = 100%. While it is perfectly possible to describe probability theory using percentages, there are technical reasons, which we shall see in a later section, which make it preferable to use numbers between 0 and 1, which we can obtain by dividing the percentages by 100. The probability of a rainy day becomes .2 after this division; this is usually abbreviated P(R) =.2 . Similarly, we see that P(S) = .3 and P(C) = .5, and we observe that P(R) + P(S) + P(C) = .2 + .3 + .5 = 1.

Here's a summary of this.

Experiments, Sample Spaces, and Probabilities

An **experiment** is the observation of something that actually happens or could happen. The **outcomes** of this experiment are the different observations that might be made. The outcomes must be non-overlapping, and must cover all possible observations.

The **sample space** of the experiment is the set of outcomes. A **probability function P** is an assignment of numbers between 0 and 1 to the different outcomes in such a way that the sum of the probabilities of all the outcomes is 1.

The probability assigned to an outcome represents the relative likelihood that the outcome will actually occur when the experiment takes place. If the probability of an outcome is 0, the outcome cannot occur, and if the probability of an outcome is 1, then the outcome is certain to occur.

Experiments, and sample spaces, belong to one of three basic types, which depend on how the probabilities are assigned to the outcomes. There are sample spaces in which the probabilities are assigned on a subjective basis. The probability of acceptance of a marriage proposal is such an example, although generally the person offering such a proposal usually feels that this probability is close to 1. Exceptions, of course, exist – celebrities continually receive such proposals from hopeful, though probably not optimistic, strangers.

Probabilities can also be assigned on an empirical basis. If a quarterback has thrown 100 passes, completed 60, had 5 interceptions, and thrown 35 incomplete passes, one can assign probabilities by using the empirically-determined relative frequencies: P(complete) = 60/100 = .6, P(interception) = 5/100 = .05, P(incomplete) = 35/100 = .35 . Notice that in this instance the probabilities can be interpreted as averages: for each pass thrown, the quarterback has averaged .6 of a completion, .05 of an interception, and .35 of an incompletion.

The third, and most readily analyzable assignment of probabilities, is the theoretical model. An easy example is the flip of a coin, in which the outcomes are heads and tails, denoted H and T

respectively. We use the symbol P(H) to denote the probability of a head occurring as a result of the flip.

Example 1 - The Fair Coin

One of the simplest experiments is the flip of a fair coin. There are only two outcomes: heads (H), and tails (T). In order to compute P(H) and P(T), we use the assumption that the coin is fair, which means that heads and tails are equally probable, so P(H) = P(T).

Since H and T are the only possible outcomes, we must have P(H) + P(T) = 1. Since P(H) = P(T), we obtain P(H) = P(T) = 1/2.

Expressed in terms of percentages, the probability of heads is 50%, and the probability of tails is also 50%. This has worked its way into the language: equal chances are sometimes called fifty-fifty chances.

The fair coin is a specific example of a **uniform probability space**. In a uniform probability space, all the outcomes are assumed to be equally likely (one of the meanings of 'uniform' is 'the same'). There are many interesting sample spaces in which the outcomes are all equally likely, such as the roll of a fair die or the drawing of a single card from a deck of 52 cards.

Example 2 - Let S be the sample space of a uniform probability space with n equally likely outcomes. Then the probability of each outcome is 1/n.

Let $O_1, O_2, \ldots, O_n$ denote the n outcomes of the experiment. Then

$$P(O_1) + P(O_2) + \ldots + P(O_n) = 1$$

Since all the outcomes are equally likely, $P(O_1) = P(O_2) = \ldots = P(O_n)$. Substituting this into the above equation, we obtain

$$P(O_1) + P(O_1) + \ldots + P(O_1) = 1$$

$n\ P(O_1) = 1$

$P(O_1) = 1/n$

Since all the outcomes are equally likely, the probability of each outcome is 1/n.

When a fair die, which has 6 sides, is rolled, the probability of rolling a 4 is 1/6 (as is the probability of rolling a 1, or a 2, etc.). When a card is drawn from a deck of 52 cards, the probability of drawing the ace of spades is 1/52, which is the same as the probability of drawing the king of spades, or the deuce of clubs.

In the story at the start of the chapter, Pete actually made a subtle error! Once March sees a boy, since he has no way of knowing whether he is seeing the younger or older child, it is twice as likely that he is seeing a family with two boys as either the BG or the GB family. There are four equally likely families: BB (he sees the younger boy), BB (he sees the older boy), BG and GB – and in half of those families the child March does not see is a girl. So Di Stefano was offering odds of 11 to 10 on a 50-50 bet – but got lucky!

In the story, Pete was interested in the probability that the man that whom March asked had a sister. This is not just a single outcome, such as BG, because we do not know whether we are asking the first-born child or not. What Pete was interested in was whether the actual outcome belonged to a predetermined set of outcomes, which in this instance consisted of both the outcomes BG and GB.

An event in a sample space is a set of outcomes; in other words, it is a subset of the sample space. In order to compute the probability that the actual outcome of an experiment belongs to a particular event, one adds the probabilities of all the outcomes in the event, just as Pete added P(BG) + P(GB) = 1/3 + 1/3 = 2/3 to (erroneously) get a probability of 2/3 that the man's sibling was a girl. It is customary to use capital letters to denote events. If we use the event from the story we have been discussing (the sibling of the man March asked was a girl), we might let A = {BG, GB}. So Pete

computed that P(A) = P(BG) + P(GB) = 2/3. Generalizing this example, the probability of an event E, written P(E), is the sum of the probabilities of all the outcomes in the event E.

The Probability of an Event

An **event** E is a subset of a sample space. In other words, an event is a set of outcomes. The **probability of an event** *E*, written P(E), is the sum of the probabilities of all the outcomes in the event E.

Example 3 - A used-car dealership sells Toyotas, Nissans, Volvos, and Jaguars. It has found that the probability P(T) that a customer who buys a car will buy a Toyota is .18. The other probabilities are P(N) = .28, P(V) = .31, and P(J) = .23 . What is the probability that a customer who buys a car will buy a Japanese car?

Solution: If the event E is the purchase of a Japanese car, then E = { T, N }, and P(E) = P(T) + P(N) = .18 + .28 = .46 . The percentage interpretation is that 46% of the customers who buy cars buy Japanese cars. ■

The null set, ∅, also represents a collection of outcomes: no outcomes, to be precise. When we perform an experiment, it is understood that something will be observed, so the probability that no outcome will be observed is 0. The null set is sometimes called the **impossible event**, or the **null event**, and P(∅) = 0.

If S is the entire sample space, the sum of the probabilities of all the outcomes in S is equal to 1, and so P(S) = 1. The entire sample space is sometimes called the **certain event**, because it is certain that, when an experiment is performed, one of the outcomes will be observed.

Many probabilistic computations are simply exercises in counting. This is especially true in uniform probability spaces.

The Probability of an Event in a Uniform Probability Space

Let A be an event in a uniform probability space X. Suppose further that the sample space X has n different outcomes, and that k of these outcomes constitute the event A. Each outcome has probability 1/n, and since there are k of them in A, P(A) can be computed by adding 1/n k times. Therefore, P(A) = k/n.

P(A) = number of outcomes in A/ number of outcomes in X

Example 4 - Let X denote the sample space associated with drawing a single card from a deck of cards. Let F be the event consisting of drawing a face card. Let S be the event consisting of drawing a spade. Let R be the event consisting of drawing a red face card. Compute the probabilities of these events. Is it more likely that you will draw a spade or a face card?

Solution: This is simply a matter of counting, since X is a uniform probability space with 52 equally likely outcomes. There are 12 face cards (four jacks, four queens, and four kings), so P(F) = 12/52 = 3/13. There are 13 spades, so P(S) = 13/52 = 1/4. Notice that P(S) > P(F), so you are more likely to draw a spade than a face card. Finally, there are six red face cards (the jacks, queens, and kings of hearts and diamonds), so P(R) = 6/52 = 3/26. ■

Here's a slightly more complicated example.

Example 5 - A committee consisting of 8 women and 4 men chooses 3 people at random to serve as its Board of Directors. What is the probability that all the members of the Board are women?

Solution: In this instance, the uniform probability space consists of all possible 3-person Boards of Directors chosen from the 12 committee members. We know from the previous chapter that each Board of Directors is a combination of 12 things taken 3 at a time, so the number of different outcomes is

C(12,3) = 12! / (3! x 9!) = 220

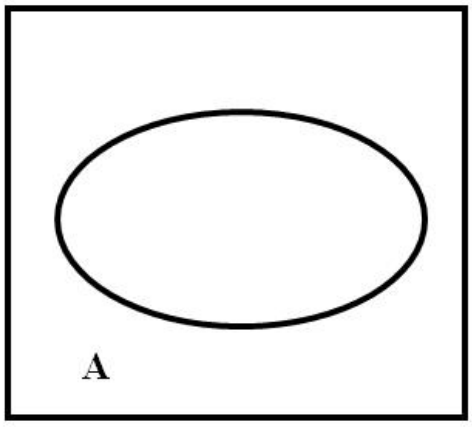
If we let W be the event consisting of all-female Boards of Directors, then each outcome in the event is a combination of 3 women chosen from the 8 women on the committee. Therefore, the number of outcomes in the event W is

$$C(8,3) = 8! / (3! \times 5!) = 56$$

Therefore, P(W) = 56/220, which is slightly more than 1/4. ■

A Geometrical Interpretation of the Probability of an Event

One of the most useful visual tools for understanding probability is to envision the sample space S as a square 1 unit on each side, so the area of the square is 1. We recall that an event is a subset of a sample space. An event A is represented by the interior of a closed curve contained in the square S. The probability of A, P(A), is the area of the interior of the closed curve. Low-probability events take up very little area in the square, and high-probability events take up a great deal of area.



As we shall see, this geometrical interpretation of probability supplies a very useful model for probability, because we can use well-known properties of area to explain many of the laws of probability.

The Law of Large Numbers

Suppose we flip a coin 4 times. We know that the probability of heads is .5, so we would expect that 50% of the time the coin would land heads. However, we wouldn't be astounded if the coin landed heads 3 times, or in 75% of the experiments. However, if we flipped the coin 20 times and it landed heads 75% of the time we would probably be a little surprised, and if we flipped the coin 100 times and it landed heads 75% of the time we would probably conclude that the coin was not a fair coin, and was weighted to land heads.

Intuitively, the more times we perform the same experiment, the closer we expect the actual outcomes to conform to the predictions of probability. When one is dealing with a theoretical model, such as a uniform probability space, this can actually be shown to be true. The theorem demonstrating this is known as the **Law of Large Numbers**, and is due to Carl Friedrich Gauss, one of the great names of mathematics.

Empirical Probabilities and Decision-making

When the probabilities are determined empirically, we still expect the Law of Large Numbers to hold. At least baseball managers do, because they send in left-hand pitchers to pitch against left-handed batters, as empirical probabilities have shown that left-handed batters generally fare less well against left-handed pitchers than against right-handed ones. This is known as playing the percentages, and as we shall see, it forms the basis for many decisions that impact one's everyday life.

There may well be a valid explanation that a given empirical probability is observed, but it may be beyond our abilities to discern that reason. Thousands of years before the invention of probability theory, gamblers were aware that a flip of a fair coin resulted in heads approximately 50% of the time. Even though the mathematics to explain this was not in existence, it was reasonable to make decisions based on this knowledge. Similarly, there are many decisions made today both by individuals and societies which are based on empirical probabilities, rather than on a

deep understanding of the underlying phenomena. Indeed, many models of phenomena are constructed using empirically-determined probabilities as a guide.

However, it sometimes happens that empirically-determined probabilities change over time. For instance, prior to the discovery of the Salk and Sabin polio vaccines, the probability that a child would contract polio was small but nonetheless positive. The Law of Large Numbers would thus predict that there would be many thousands of cases of polio in a country as large as the United States, a prediction which fortunately became completely wrong when the vaccines were developed. In cases such as these, the Law of Large Numbers is still used as a predictor of events with empirically-determined probabilities, but the fact that the empirically-determined probabilities change will be reflected in current predictions. Insurance companies continually update their data bases to reflect changing conditions and ensure more accurate accident probabilities.

Simulations

Even with the computational rules for probability which we shall study later in the chapter, it is sometimes too complicated to compute exact probabilities. For instance, if we know the batting averages and distribution of hits (singles, doubles, etc.) of all the players in the batting order of a baseball team, could we estimate the probability that the team would score 5 runs? Over the course of a nine-inning game, there are an astronomical number of ways that the team could score 5 runs. To come up with an estimate, one could **simulate** a baseball game on a computer.

Many computer software packages include a **random number generator**, which selects an integer at random from a set containing any number of integers. To see how this could be used to simulate an at-bat for a baseball player, let's look at a hitter who had 30 singles, 15 doubles, 5 triples, and 10 home runs in 200 plate appearances. One would use the random-number generator to select a number at random from the integers between 1 and 200 (this is the computer equivalent of putting 200 balls in a jar with the numbers 1 through 200 on them, shaking the jar, and picking one of them). If the number chosen were between 1 and 30, then the batter gets a 'simulated' single. If

the number were between 31 and 45, the batter gets a 'simulated' double. Between 46 and 50, he gets a 'simulated' triple, and between 51 and 60, a 'simulated' home run, with all other numbers resulting in 'simulated' outs.

Because modern computers (even home PCs!) can do this a large number of times in a single second, it is possible to play nine 'simulated' innings in less than a second. One could therefore play thousands of 'simulated' games in an hour, and then compute the fraction of games in which the team scored 5 runs. This would serve as the estimated probability of the team scoring 5 runs.

The table of random digits (Table 9-1) can also be used for the same purpose. To simulate an at-bat for our imaginary batter, choose a random location in the table to start. For instance, if the time is 12:27, you might start at the 27th number in row 12. Take the 3 digits immediately following that location (551 if you use the 27th number in row 12 as a starting point). Divide this number by 5 to obtain a random number between 0 and 200 (in the example case, 110.2). If the number is between 1 and 30, then the batter gets a single, between 31 and 45, a double, etc. In the example just illustrated, where the number was 110.2, the batter makes an out. Before inexpensive computers made random-number generation easy to do, books containing millions of random digits were published to enable simulations to be performed.

3	4	8	6	4	2	4	1	5	8	3	8	0	7	4	2	5	2	2	0	5	2	1	8	5	7	2	2	4	4	9	8	2	3	7	1	5	6	5	0	6	6	2	6	3	1	9	3	4	2
2	4	4	8	6	4	3	8	8	0	0	0	1	8	0	1	0	4	3	7	5	2	9	1	6	2	6	6	1	5	6	7	2	6	7	4	3	1	3	9	1	5	9	8	8	9	8	9	7	6
9	8	0	2	2	0	9	5	8	1	6	5	9	1	2	7	6	7	6	3	9	3	6	9	4	4	7	2	9	9	8	5	9	2	7	6	5	2	3	7	9	2	9	2	9	9	6	4	5	4
5	9	6	1	8	4	2	6	8	5	3	6	5	9	8	0	5	6	1	9	9	3	5	8	2	9	1	4	5	8	9	0	0	7	5	0	5	1	2	0	5	3	0	6	0	9	3	1	4	9
7	0	8	5	9	7	2	6	3	2	1	8	2	2	1	3	4	8	7	4	1	5	3	8	6	6	2	2	3	5	6	3	1	2	0	5	1	8	7	8	9	6	8	1	7	8	5	8	2	1
2	0	2	1	9	9	2	3	1	6	0	8	0	3	5	7	5	2	7	4	2	0	3	1	9	1	6	9	7	7	0	7	4	4	9	5	6	7	7	7	0	5	9	8	5	8	8	1	9	6
6	1	8	7	5	7	7	8	7	2	1	2	0	4	8	5	7	7	6	6	1	9	8	2	1	6	5	1	7	4	6	0	6	7	3	5	9	2	4	2	4	5	2	0	0	4	3	0	4	2
1	1	6	3	0	7	8	0	4	4	8	1	3	1	8	8	0	3	7	0	9	7	7	4	2	2	0	6	4	9	9	6	8	2	7	6	3	2	7	9	9	0	9	4	0	8	8	7	4	1
5	9	8	8	6	8	2	2	0	3	9	8	9	8	3	0	3	3	7	1	7	5	5	0	4	3	9	4	2	1	4	1	8	0	6	1	5	5	7	5	2	4	0	9	6	8	8	9	6	3
7	8	4	2	0	6	7	2	3	5	4	5	3	1	9	6	1	5	5	7	6	3	8	1	2	9	9	2	3	7	5	1	2	7	6	5	1	7	7	5	8	3	9	4	9	2	8	2	1	8
2	9	6	1	8	1	8	9	3	8	3	4	2	4	4	6	9	0	0	5	3	9	8	4	6	4	4	5	3	8	1	0	6	4	1	7	6	6	5	7	1	8	9	1	4	0	7	5	4	1
1	5	5	6	4	9	2	9	3	5	7	1	0	7	7	1	0	3	0	8	2	3	0	1	0	5	5	5	1	3	9	2	2	1	6	6	5	1	9	1	1	0	0	9	7	8	9	3	6	8
3	2	5	9	5	7	3	0	0	5	7	3	9	0	1	2	9	0	9	1	4	2	7	6	0	0	0	5	9	2	7	2	2	6	9	7	3	9	4	9	0	7	9	6	9	6	4	1	4	7
6	4	4	8	4	7	2	8	6	1	8	0	0	9	1	8	6	6	4	1	9	2	7	9	1	2	2	9	3	6	0	3	9	8	4	4	7	0	5	9	8	0	4	2	8	0	3	4	9	7
0	8	2	3	1	2	2	2	2	1	5	4	8	9	6	5	9	8	9	3	4	4	8	3	7	2	0	2	7	3	8	8	8	9	1	3	4	8	0	0	0	7	7	4	2	0	7	4	6	1
7	6	9	5	7	8	7	6	7	1	3	7	4	2	8	0	1	7	4	5	8	9	0	2	3	7	8	2	7	5	6	0	0	1	8	0	2	6	8	2	3	5	3	5	2	5	4	9	7	3
1	8	0	0	9	7	4	5	6	3	7	1	8	6	1	3	3	9	6	1	6	8	8	9	2	1	3	3	0	2	3	5	7	7	3	4	8	8	8	1	2	2	8	4	9	9	9	0	6	5
9	8	6	0	7	2	5	0	2	1	1	9	6	1	9	4	8	9	7	8	1	7	3	2	0	1	7	5	8	4	2	1	3	7	8	2	3	5	6	3	6	5	4	5	2	9	3	2	3	0
9	1	5	7	1	4	0	3	6	9	4	2	2	0	6	1	6	6	1	6	7	3	8	5	0	5	3	0	9	0	1	9	1	1	6	4	1	3	9	6	5	3	8	4	3	6	9	1	1	1
5	9	6	9	9	8	2	7	9	8	0	1	7	4	4	4	0	6	1	8	9	4	7	0	7	0	4	5	0	5	6	5	9	7	6	3	8	1	0	1	8	3	0	8	6	8	6	7	0	7

Table of 1000 Random Digits

Many important engineering, scientific, and business problems require computation of probabilities, but the sample spaces are so large that exact computations are impossible. Modern computers enable all sorts of problems to be analyzed by simulation, including such critical applications as the spread of epidemics (such as AIDS). Computational simulation of complex events is now an important part of such diverse fields of study as aeronautical engineering, cosmology (in which entire universes are simulated to evolve!), and chemistry. These techniques are so reliable that many vital decisions, which were previously merely guesses, can now be made with confidence.

Section 2 - Computing Probabilities of Events

Some of the principles of counting appear in much the same form as principles of probability. This is not so surprising when we consider that in one of the most important types of sample space, the uniform probability space, probabilities are determined by counting. If we use the symbol N(E) to denote the number of outcomes in an event E that is a subset of a uniform probability space S, then the probability P(E) is given by the formula

$$P(E) = N(E)/N(S)$$

Suppose that the sample space S is the set of all outcomes corresponding to drawing a single card from a deck of cards. Let E be the event consisting of drawing either an ace or a black card. While we could simply count the outcomes, let's analyze it by using a counting principle. If A denotes the "ace" event (drawing an ace), and B denotes the "black card" event (drawing a black card), by the counting principle

(1) $$N(A \cup B) = N(A) + N(B) - N(A \cap B)$$

Dividing both sides by N(S) we get

(2) $N(A \cup B)/N(S) = N(A)/N(S) + N(B)/N(S) - N(A \cap B)/N(S)$

Therefore

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

It is easy to see that $P(A) = 4/52$ and $P(B) = 26/52$. Because $A \cap B$ has only two outcomes, the spade and club aces, we see that $P(A \cap B) = 2/52$, and so $P(A \cup B) = 4/52 + 26/52 - 2/52 = 28/52 = 7/13$.

The algebra in equations (1)-(3) does not depend on the specific events involved, but only on the fact that the sample space is a uniform probability space. Therefore, in a uniform probability space, equation (3) applies to the computation of probabilities for any two events A and B.

Although the above computation was done in a uniform probability space, equation (3) actually holds for any probability function and any sample space. This can be proved fairly easily, but it is probably easiest to see simply by using the 'area model' for probability.

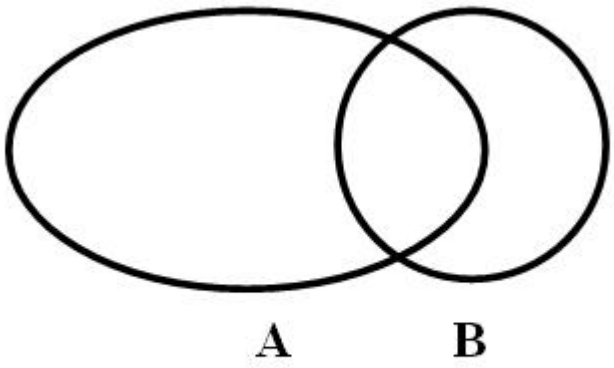
Unions and Intersections in Sample Spaces

Let A and B be two events in a sample space S associated with a particular experiment. The event $A \cup B$ corresponds to the outcome of the experiment belonging to either A, or B, or both; and the event $A \cap B$ corresponds to the outcome of the experiment belonging to both A and B. The probabilities of $A \cup B$ and $A \cap B$ are related by

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

This can be seen from the following diagram, in which the area corresponding to

$A \cup B$ is the sum of areas of A and B, minus the area of $A \cap B$, which has been added twice (once in the area of A, and once in the area of B). You may recall this "overcounting" argument from Chapter 7.



Example 1 - The probability that a car dealer will advertise in either a newspaper or on radio is .94. If the probability that he will advertise in a newspaper is .73 and on radio is .51, what is the probability that he will advertise in both?

Solution: Letting N denote the event 'advertise in a newspaper', and R the event 'advertise on radio', we have

$$P(R \cup N) = P(R) + P(N) - P(R \cap N)$$

$$.94 = .51 + .73 - P(R \cap N) = 1.24 - P(R \cap N)$$

So $P(R \cap N) = 1.24 - .94 = .3$. ■

Two events A and B are said to be **disjoint**, or **mutually exclusive**, if they have no outcomes in common, i.e. $A \cap B = \varnothing$. If this is the case, $P(A \cap B) = 0$, and so the above law reduces to a simpler form.

Probabilities of Disjoint Events

If A and B are disjoint events in a sample space S,

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

In this case, since $A \cap B = \varnothing$, $A \cap B$ has no area.

This formula can be extended to more than 2 disjoint events, as can be seen in the following examples.

Example 2 - In 200 previous plate appearances, a batter has amassed 30 singles, 15 doubles, 5 triples, and 10 home runs. Assuming the batter performs in the future as he has in the past, what is his probability of getting an extra-base hit?

Solution: There are two equally valid ways to compute the probability. One could look at the event E, the set of all extra-base hits. Since N(E) = 15 + 5 + 10 = 30, the probability that the batter will get an extra-base hit is N(E)/N(S) = 30/200 = .15 .

Alternatively, one could consider the events D, the set of all doubles, T, the set of all triples, and H, the set of all home runs. $E = D \cup T \cup H$, and D, T, and H are disjoint. Since N(D) = 15, P(D) = 15/200 = .075. Similarly, P(T) = 5/200 = .025, and P(H) = 10/200 = .05 . Adding the probabilities of these disjoint events, we arrive at P(E) = .075 + .025 +.05 = .15 . ■

Example 3 - The Board of Directors of Software Services consists of 4 men and 6 women. If a five-person committee is selected at random from the Board, what is the probability that a majority of its members will be women?

Solution: The sample space is a uniform probability space consisting of C(10,5) possible five-person subcommittees. The event E, consisting of a majority of women on the committee, is the union of three disjoint events: A = { 3 women, 2 men on the committee}, B = { 4 women, 1 man on the committee}, and C = { 5 women on the committee }. Since the choice of men and women is independent, the Chinese Restaurant Principle can be used to see that N(A) = C(6,3) x C(4,2), and so P(A) = C(6,3) x C(4,2) / C(10,5). Similarly,

P(B) = C(6,4) x C(4,1) / C(10,5)

P(C) = C(6,5) x C(4,0) / C(10,5) = C(6,5)/C(10,5) (why?)

Computing, we obtain C(10,5) = 10! / 5! x 5! = 252, and

P(A) = (6! / 3! x 3!) x (4! / 2! x 2!)/252 = 120/252

P(B) = (6! / 4! x 2!) x (4! / 1! x 3!)/252 = 60/252

P(C) = (6! / 5! x 1!)/252 = 6/252

Therefore, P(E) = 120/252 + 60/252 + 6/252 = 186/252, or approximately .74 . ■

Most of us have used the computation of the probability of disjoint events in connection with daily weather forecasts. If the weatherman tells us that the probability that it will rain tomorrow is 10% (remember that weathermen -- and women -- habitually express probabilities in percentages), we know that the probability that it will not rain is 90%. This is a special case of the disjoint events formula, which occurs when A = E and B = E', the complementary event to E. In this case, $E \cup E' = S$, and so P(E) + P(E') = P(S) = 1. Therefore P(E') = 1 - P(E).

Probability of Complementary Events

If E is an event in a sample space S, and E' is the complementary event, then

$$P(E') = 1 - P(E)$$

Example 4 - If 85% of the viewing audience watch network news telecasts, 70% watch cable news telecasts, and 92% of the TV audience watch at least one of the two, what is the probability that a member of the TV audience will not watch *both* network and cable news?

Solution: If we let N denote the 'watch network news' event, and C the 'watch cable news' event, then we are interested in the probability of $(C \cap N)'$. We can compute $P(C \cap N)$ from the formula

$$P(C \cup N) = P(C) + P(N) - P(C \cap N)$$

$$.92 = .7 + .85 - P(C \cap N) = 1.55 - P(C \cap N)$$

So $P(C \cap N) = 1.55 - .92 = .63$ (these people are news junkies!), and $P ((C \cap N)') =$

$1 - P(C \cap N) = 1 - .63 = .37$. ■

Example 5 - In a class consisting of 25 people, what is the probability that two people were born on the same day of the year? Assume for simplicity that no one is born on February 29th.

Solution: This is a problem where it is easier to compute the probability of the complementary event, which in this case is the probability that no two people were born on the same day. By the Chinese Restaurant Principle, the number of possible birthdates for 25 people is 365^{25}. To compute the number of possible ways that 25 people could be born on different days, notice that the first person could be born on any one of 365 days. However, then the second person only has 364 possible birthdays (without duplicating the birthday of the first person), the third person only has 363 possible birthdays (without duplicating the birthday of either the first or second person), and so on down to the 25th person, who only has 341 possible birthdays (without duplicating a birthday of the first 24 people). Therefore, the probability of all 25 people being born on different days is

$$(365 \times 364 \times 363 \times ... \times 341)/365^{25} = .431$$

The probability that at least two people were born on the same day is $1 - .431 = .569$. ■

The following table indicates the probability that at least two people in a randomly-selected group have the same birthday.

Number of People in Group	Probability of Duplicated Birthday
5	.027
10	.117
15	.253

20	.411
23	.507
25	.569
30	.706
40	.891
50	.970
60	.994

With 23 people, the probability is a little more than 1/2 that at least two have the same birthday. This result is very surprising to most people. When asked to guess how many people one must collect before the chances of having two with the same birthday are roughly half, most people choose a much higher number than 23.

Section 3 - Expectation

Now that we know how to compute probabilities, how do we use them to make plans? Many situations arise in which payoffs are associated with the occurrence of an outcome or an event. Sometimes these payoffs are measured in terms of money, as in the bet between March and DiStefano. Sometimes they are measured in other units. One can, for instance, compute the probability of a worker being sick, and measure the payoffs in terms of hours worked, or the probability of a shipment of items containing defective items, and measure the payoffs in terms of defective items.

Let's start with a simple example. Suppose that you draw a card from a deck of 52, and a benevolent individual offers to pay you $5 for each time you select a face card, provided that you

pay him $1 every time you don't. The game goes as follows: you draw a card, pay or get paid, put the card back in the deck, shuffle the deck, and draw again.

If you were to play this game 52 times, and happened to draw each card in the deck once, your balance sheet would show a gain of $5 for each of the 12 face cards, and a loss of $1 for each of the 40 other cards. Your net would be 12 x $5 - 40 x $1 = $20. Since you played the game 52 times, your average win per play would be $20 / 52, which is approximately $.38 . We say that your ***expected gain per play*** is $.38 .

Let's look at a different way of computing the same number, $.38 . If we let F denote the event 'draw a face card' and N denote the event 'draw a non-face card', we see that our expected gain per play, which we know is (12 x $5 - 40 x $1) / 52, can also be written

(12 x $5 - 40 x $1) / 52 = 12/52 x $5 + 40/52 x (- $1)

= P(F) x V(F) + P(N) x V(N)

where V(F) denotes the value of getting a face card ($5), and V(N) denotes the value of getting a non-face card (-$1).

This gives us the "recipe" for computing the average gain per play associated with any game. Break the game up into events such that each outcome in a selected event has the same payoff. In the above game, the two events (F and N) were determined by the fact that each outcome in F had a payoff of $5, and each outcome in N had a payoff of -$1 . Multiply the probability of each of the events by its associated payoff, and add the result. This average gain per play is called the **expectation** of the game, and is usually denoted by E. A game which has an expected value of 0 is called a **fair** game.

Definition of Expectation

Suppose that an experiment has outcomes $x_1, \ldots, x_N$.

If P(x) denotes the probability associated with outcome x, and V(x) denotes the payoff (value) associated with outcome x, then the expectation E is defined by

$$E = P(x_1)\, V(x_1) + \ldots + P(x_N)\, V(x_N)$$

Payoffs are positive from the point of view of the person or persons for whom the expectation is being computed. If several different outcomes have the same payoff, it is natural to group these together as an event. The above formula would then apply, with x_1 denoting the first event, etc.

Once again, this is what Pete did in the story when he explained March's expectation by assuming that March bet $10 three separate times, winning once and losing twice . We can denote the two events as S (the sex of the sibling is the same as the person being asked), and O (the sex of the sibling is opposite to the person being asked). According to the rules of the game, if March bet $10, V(S) = $11 and V(O) = -$10 . We determined previously that P(S) = 1/3 and P(O) = 2/3. Therefore the expectation is

$$E = P(S) \times V(S) + P(O) \times V(O) = 1/3 \times \$11 + 2/3 \times (- \$10) = -\$3$$

You may recall that Pete described the expectation as 30% (from DiStefano's point of view). It wasn't clear whether each bet was $10, $100, or $1000, so one might just as well assume that each bet was 10 units. Then the expected loss per bet would be 3 units. Considering that 10 units was the amount of the bet, the percentage expected loss per unit bet would be 30%. Percentages therefore give a convenient way to describe the expectation of a game.

Example 1 - You pay $3 to roll a single die. You are awarded the same number of dollars as the number on the face of the die. What is your expectation?

Solution: The probability of each number appearing is 1/6, as the sample space of this experiment is a uniform probability space. Therefore, the expectation is

$$E = 1/6 \times 1 + 1/6 \times 2 + ... + 1/6 \times 6 - 3 = 3.5 - 3 = .5$$

Your expectation is therefore \$.50 . Since you are paying \$3 to play, your percentage expectation is 100 x \$.5/\$3 = 16.67 %. ▉

Example 2 - A shipment of 50 fluorescent light bulbs has a probability of 0 of having more than 3 defective light bulbs. The probability of having 1 defective bulb is .04, the probability of having 2 defective bulbs is .02, and the probability of having 3 defective bulbs is .01 . What is the expected number of defective bulbs per shipment? What is the percentage expectation, and how is it to be interpreted?

Solution: E = 1 x .04 + 2 x .02 + 3 x .01 = .11 defective bulbs per shipment. Since the shipment contains 50 bulbs, the percentage expectation is 100 x .11 / 50 = 0.22% . This is the percentage of defective bulbs. ▉

In Example 2, one would expect that 22 bulbs out of a shipment of 10,000 bulbs would be defective. This does not mean that in a shipment of 10,000 bulbs there will be precisely 22 defective bulbs -- there might 19 defective bulbs, or there might be 30. However, in the long run, a shipment of 10,000 bulbs will contain an average of 22 defective bulbs.

Odds

When a game has only two outcomes, sometimes the payoffs are quoted in terms of how much the player will win as compared with how much the player will lose. Odds of 3-to-1 means that the player stands to win 3 units if he or she wins, but will lose 1 unit if he or she loses. Odds of 3-to-1 describes those situations in which the player stands to win three times as much as he or she stands to lose.

Odds of 2-to-5 means that the player stands to win 2 units, but risks losing 5 units. It is customary to quote odds in whole numbers at the simplest possible ratio. Odds of 1-to-1 are sometimes called **even money**.

Example 3 - A card is drawn from a deck of 52 cards. You are offered odds of 3 to 1 that the card is a face card; in other words, if it is a face card you would win 3 units, if not you would lose 1 unit. What is your percentage expectation?

Solution: Since the probability that the card is a face card is 12/52 = 3/13, your expectation on a $1 bet is

$$E = 3/13 \times \$3 + 10/13 \times (-\$1) = -1/13$$

Your percentage expectation is therefore 100 x -1/13 = -7.69% . ■

Example 4 - A British bookmaker (bookmaking is not only legal in England, but respectable) has had punters (British for bettors) bet 300 pounds on Oxford and 200 pounds on Cambridge on the upcoming Boat Race. If he plans to take 50 pounds for his commission, what odds should he offer on each school?

Solution: A total of 500 pounds has been bet, and after he takes his commission, 450 pounds are left. Those who have bet on Cambridge stand to lose 300 pounds, but they could win 450 pounds. So the odds the bookmaker should offer for a bet on Cambridge are 450-to-300, or 3-to-2. Similarly, since 200 pounds were bet on Oxford, the odds on Oxford are 450-to-200, or 9-to-4. ■

A reasonable assumption in Example 4 is that, because .6 of the total money wagered was bet on Oxford, the probability that Oxford will win the race is .6 . In this case, the expectation of the Oxford bettors is

$$E = .6 \times 150 + .4 \times -300 = -30$$

Notice that, in this case, the percentage expectation of the Oxford bettors is therefore

100 x -30/300 = -10% .

But it is not just games that have expectations associated with them. When an insurance company sells a policy, or a company brings out a new product, they compute expectations. In order to decide whether an investment is a good one, they must calculate the expectation involved. Sometimes the probabilities or payoffs must be estimated due to lack of complete knowledge. A company will usually bring out a new product only if it can project a positive expectation.

Example 5 - An insurance company projects that, in a single year, married females over 25 years of age have a probability of .08 of having an accident. 40% of these accidents are minor, averaging $250 in costs to the company, and 60% are major, averaging $1500 in costs. What is the expected value of an accident? What is the expected payment to a married female policyholder over 25 years of age who has an accident?

Solution: If C is the event 'minor (cheap) accident', and E is the event 'major (expensive) accident', then the expected value of an accident is P(C) x V(C) + P(E) x V(E) = .4 x $250 + .6 x $1500 = $1000. The expected payment to a married female policyholder over 25 years of age who has an accident is the probability of the accident times its expected cost, .08 x $1000 = $80 . ■

Example 6 - Rutabaga Foods is considering investing $10,000,000 in nacho-flavored pretzels. A market survey indicates a probability of 20% that the investment will lose $3,000,000, a probability of 50% that the investment will make $1,500,000 and a probability of 30% that the investment will make $2,500,000. What is the percentage expectation of the investment?

Solution: The total expectation is

E = .2 x -$3,000,000 + .5 x $1,500,000 + .3 x $2,500,000

= $900,000

The percentage expectation is therefore 100 x $900,000/$10,000,000 = 9%

(Note: it is very unlikely that Rutabaga Foods would decide to make the investment. While 9% is a satisfactory return on a safe investment, on a risky one a higher expectation is usually required to allow for the possibility of a disaster.) ■

Chapter 10 - Conditional Probability and Bayes' Theorem

Introduction

Let's analyze Julie's dilemma (to stand pat or to switch her choice) from the story at the start of this chapter. Recall that Julie has initially selected Wyatt as the man she thinks Debbie will marry.

We'll look at two separate cases. In the first case, a scriptwriter is *specifically told* to involve either Ellison or Lowell in the pile-up on I-5. In the second case, a recently-hired scriptwriter comes in and is given the following instruction: write a scene in which one of Debbie's three suitors gets involved in a car crash. The scriptwriter then just happens to pick Ellison or Lowell. Notice the difference between the two cases: in the first case, the scriptwriter is explicitly told to choose Ellison or Lowell, and in the second case he or she is told to choose a victim at random, and that victim just happens to be Ellison or Lowell.

Since we have no advance information which one Debbie is likely to marry, we can assume that she is equally likely to marry Ellison, Wyatt, or Lowell. We first analyze the situation in which the screenwriter is specifically told to involve Ellison or Lowell.

Case 1 - Debbie marries Wyatt. The scriptwriter can involve either Ellison or Lowell in the car crash. It is wrong for Julie to switch (she would lose the $5,000 cost to switch, as well as the grand prize of $100,000).

Case 2 - Debbie marries Lowell. The scriptwriter must involve Ellison in the crash, because otherwise Lowell would be in a coma in Monterey, and unable to take the marriage vows. Now Julie wins the $100,000 by switching.

Case 3 - Debbie marries Ellison. As in case 2, the scriptwriter must involve Lowell in the crash. Again, it is right for Julie to switch.

We can therefore see that, in two out of these three cases, it is right for Julie to pay the $5,000 and switch choices.

Now suppose instead that a scriptwriter who is told to write a scene involving one of the suitors in the car crash happens to write a scene in which either Lowell or Ellison (the two suitors that Julie did not select) is involved. There are six a priori possible situations, summarized in the table below, all of which have equal probability. (This information could also be presented in the form of a tree.)

Case	Debbie Will Marry	Party in Car Crash	
1	Wyatt	Ellison	
2	Wyatt	Lowell	
3	Ellison	Ellison	*
4	Ellison	Lowell	
5	Lowell	Ellison	
6	Lowell	Lowell	*

However, cases 3 and 6, which are marked by asterisks, could not have occurred. If they did, the screenwriter would have been told to go back and choose another victim, because the chosen victim was the one that Debbie had been slated to marry. As a result of this special knowledge, there are only four equally probable situations. In cases 1 and 2, it does not pay to switch, since Julie has already selected the winner. In cases 4 and 5, Julie would clearly benefit from switching. If you and a friend decide to bet a dollar on whether a coin is flipped heads or tails, but you have to pay a nickel to choose tails, you would certainly choose heads -- they are equally likely, so why pay the nickel? Julie is now in exactly the same situation, and so there is no point in switching her choice.

The difference between the two situations is that in the first instance the scriptwriter was given the following *condition*: involve Ellison or Lowell in a car crash. In so doing, the sample space of the experiment has been modified, from the original sample space (any of the three suitors might end up in a car crash), to a subset of the original sample space (only Ellison or Lowell ends up in the car crash). The study of sample spaces which have been modified by a condition forms the subject of conditional probability.

Incidentally, the argument that Julie should switch is far from obvious, and indeed goes against the grain for most people. When a variation of this problem (known by mathematicians as the Monty Hall Problem) was posed in a nationally-syndicated magazine column, the author of the column mentioned that it generated incredible amounts of write-in commentary, and some very highly-educated people came to the wrong conclusion!

Section 1 - Conditional Probability

The poker game of 5-card stud, which was the *de facto* money poker game for nearly a century until televised big-stakes poker made Texas Hold-'Em the game of choice.

In Texas Hold-'Em, each player is initially dealt two cards face down, so that only the player to whom those cards are dealt can see them. After a round of betting, three cards are simultaneously turned face up in the center of the table (this is called 'the flop'), and each player regards those three cards as augmenting the two they were originally dealt.

In 5-card stud, each player is initially dealt one card face down, which only the player can see, and one card face up which everyone can see. A round of betting ensues. Then each player is dealt a card face up, and another round of betting ensues. This continues either until all but one player has 'folded' (refused to match a bet made by another player), or until the remaining players have a total of five cards, one face down and four face up.

Suppose that in a game of 5-card stud, you are dealt the ace of clubs face down, the ace of hearts and the four, five, and six of diamonds face up -- a pair of aces. Doc, your opponent, has been dealt the five, seven, nine, and queen of spades face up. You quickly observe that Doc can only win if he has a spade face down to give him a flush, and you also quickly calculate that there are forty-three unseen cards of which nine are spades. Based on only this information, Doc's chances of winning are therefore 9/43, and yours are a healthy 1 - 9/43 = 34/43.

Doc has, however, made a big mistake: he is sitting with his back to a large mirror, and as he glances at his down card you can see a flash in the mirror -- not enough to know for certain what that card is, but enough to convey some information to you. Let's look at three different cases.

Case 1 - You catch a glimpse of black. Uh-oh. Doc's card must now be one of 21 unseen black cards (you have seen your ace of clubs and Doc's four spades of the 26 black cards in the deck). Since nine of them are spades, Doc's chances of winning have increased to 9/21 = 3/7, and yours have decreased to 1 - 3/7 = 4/7.

Case 2 - You see a flash of red. Doc's card must now be one of 22 unseen red cards (you have seen your ace of hearts and four, five, and six of diamonds), and none of them are spades! Doc's chances of winning are 0/22 = 0, so you are certain to win! Gleefully you watch Doc finger his chips, hoping he'll try to bluff you out of the pot.

Case 3 - You see the markings indicating a face card. Doc's card must now be one of 11 unseen face cards (you have seen Doc's queen of spades), two of which are spade face cards (the jack or the king). Doc's chances of winning are therefore 2/11, and yours are 1 - 2/11 = 9/11.

Each of these situations represents a study in conditional probability. In the original situation, you had no information on the nature of Doc's card. The sample space S for the experiment was therefore all 43 unseen cards in the deck, and the event D (Doc has a winning hand) consisted of all nine unseen spades. Since this is a uniform probability space, P(D) = N(D) / N(S) = 9/43.

In each of the above three cases, information came to you which changed the sample space for the experiment. The new sample space was a subset of the original sample space S. The event that Doc wins was also altered by the outcomes available in the new sample space. In Case 3, for instance, the new sample space was F, the set of all unseen face cards, and the event that Doc wins consisted of all spades that were also unseen face cards. The eight of spades, which was a winning card for Doc in the original situation, was no longer a winning card in light of the fact that it could not belong to F, the revised sample space.

Let's look at this from a more general standpoint. Suppose that we have two events A and B in a sample space S. We now define P(A | B) to be the probability of the event A, given that the event B has already occurred (the symbol A | B is read 'A given B'). The conditional probability P(A | B) is given by the following formula.

Conditional Probability of an Event

Let A and B be events in a sample space. Then the probability of A occurring, given that B has already occurred, is defined by

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

Notice that this definition gives the same results as the computations we have already done in the three cases discussed in the poker problem. For instance, in Case 3, if we let D be the event that Doc wins (Doc's unseen card is a spade) and B the event that Doc's unseen card is a face card, then $D \cap B$ is the event that Doc's unseen card is a spade face card. We had already computed above that $P(D \mid B) = 2/11$. Since $N(B) = 11$ and $N(D \cap B) = 2$, in the original sample space S (no information about the unseen card) we see that $P(B) = 11/43$ and $P(D \cap B) = 2/43$. Therefore, the computational rule $P(D \mid B) = P(D \cap B) / P(B) = (2/43) / (11/43) = 2/11$.

Example 1 - What is the probability that, if at least one child in a two-child family is a girl, the other is a boy?

Solution: We've analyzed this logically in the story in Chapter 9, but let's use the definition of conditional probability given above. The sample space S is the space of all two-child families: S = { BG, BB, GB, GG } (we have written the children older first, as usual). The event A consists of one child of each sex: A = { BG, GB }. The event B consists of at least one girl: B = { BG, GB, GG }. In this case, $A \cap B = A$, and $P (A \cap B) = P(A) = 2/4$, and $P(B) = 3/4$. So

$$P (A \mid B) = P (A \cap B) / P(B) = (2/4) / (3/4) = 2/3$$ ■

The attempt to acquire additional information in the world of business in order to estimate conditional probabilities more accurately is an activity which is generally favorably regarded. On the other hand, similar activities conducted within the confines of a poker game are often viewed as cause for hostile action.

Example 2 - A small town contains 80 men and 60 women. Half of the men, and two-thirds of the women, are Democrats. A pollster interviewed a Republican. What is the probability that the interviewee was a man?

Solution: One way to solve this problem is by counting. There are obviously 40 Republican men and 20 Republican women, so the probability that the interviewee was a man is 40/60, or two-thirds.

The second way is to use the formula for conditional probability. Let R denote the event that a Republican is interviewed, and M the event that a man is interviewed. Then there are 40 male Republicans and 20 female Republicans. So P(R) = 60/140, and

$P(M \cap R) = 40/140$. Therefore

$$P(M \mid R) = P(M \cap R) / P(R)$$

$$= (40/140) / (60/140)$$

$$= 40/60$$

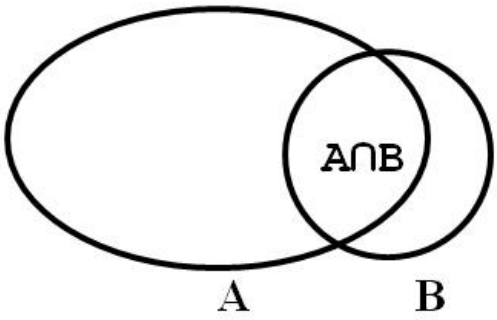
= 2/3

The number 40/140 is the probability that a random person in the town is both Republican AND a man; the number 60/140 is the probability that a random person is a Republican. ■

We have seen that defining P(A | B) = P (A ∩ B) / P(B) yields the correct answers in our examples. Another way to understand the motivation for this definition can be seen in the area model for probability. In this model, the area of the square representing the sample space is 1. Since P(A) = P(A) / 1 = P(A) / P(S), the probability P(A) is the ratio of the area of the event A to the area of the entire sample space.

When we assume that event B has already occurred, the event B becomes the new sample space, and only those outcomes in A that also belong to B can be considered as part of the event A | B. So the probability of A given B, P(A | B), must be the ratio of the area of the allowed outcomes A ∩ B to the area of the new sample space B, or P(A ∩ B) / P(B).

Area Model for Conditional Probability

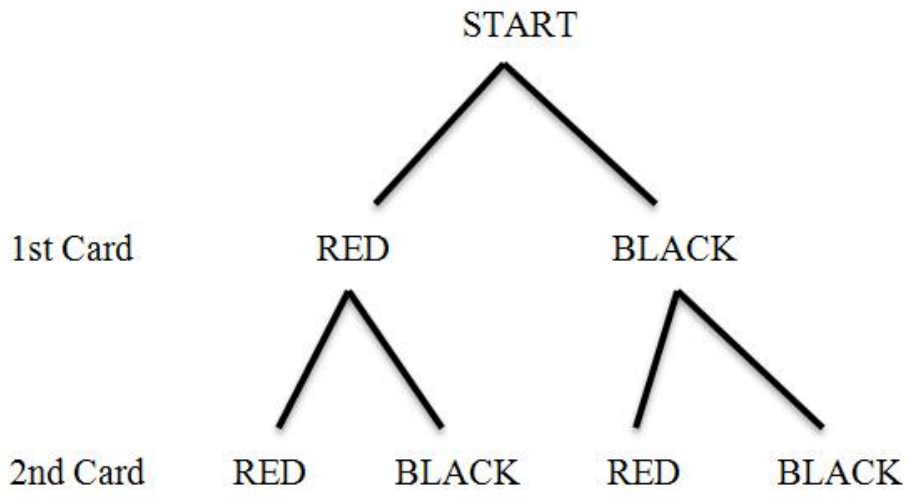


Probability Trees

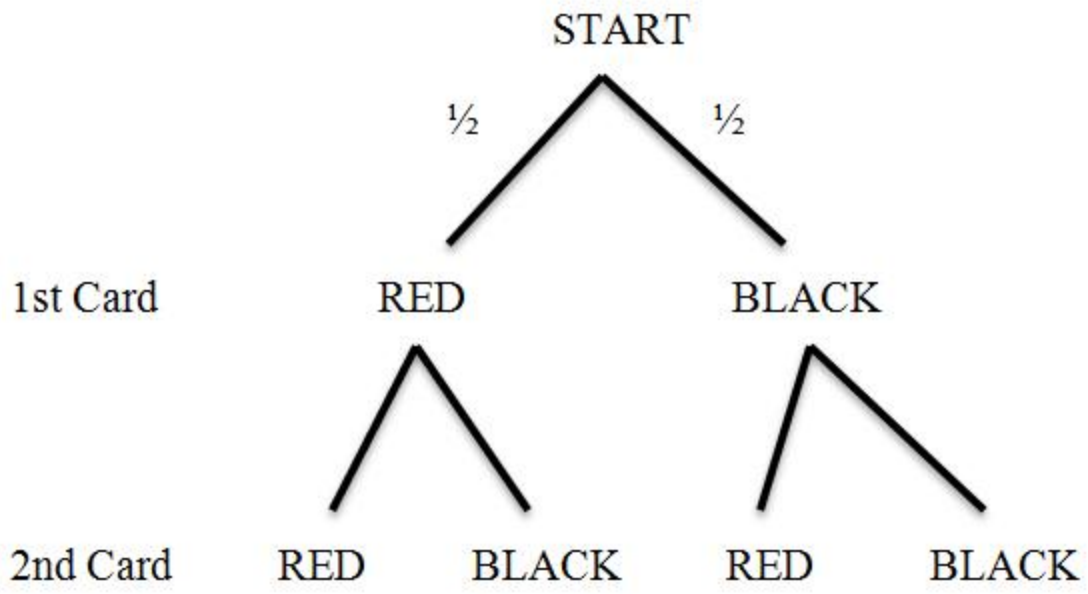
If P(B) is not 0, the formula P (A | B) = P (A ∩ B) / P(B) is equivalent to the formula P (A ∩ B) = P (A | B) P(B). While the first formula is used to compute conditional probabilities, the second formula is often used to compute the probability that two events, A and B, will <u>both</u> occur, and this is how we will interpret P(A ∩ B).

We can handle more complex problems by modifying the counting trees we have used in combinatorics problems. We illustrate this with the following problem. Suppose that two cards are drawn from a deck of cards without replacement. What is the probability that one is black and one is red?

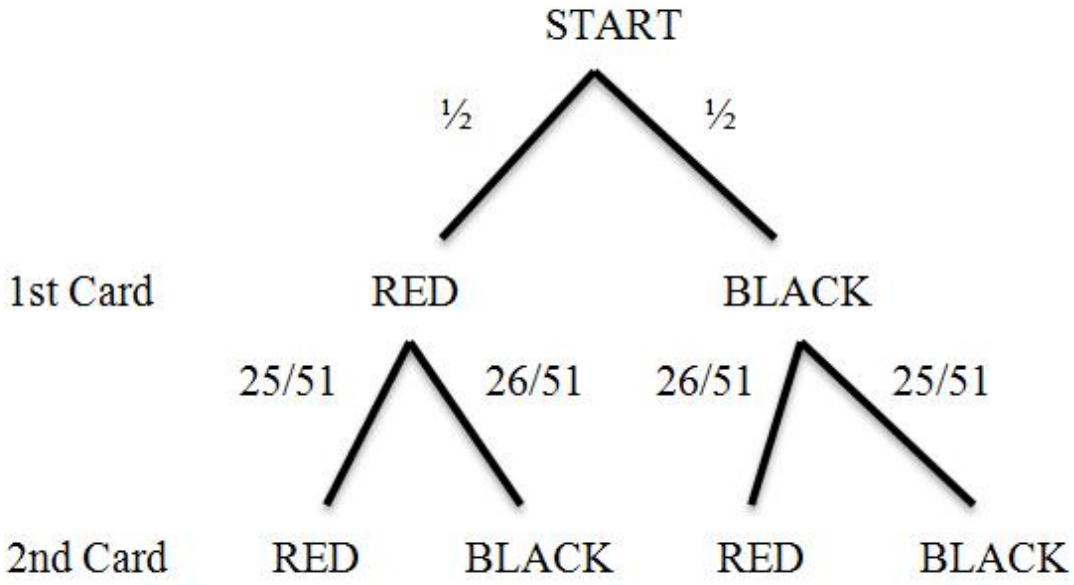
We start by drawing a tree to depict all the possible outcomes. When an event is connected via a straight line (called a branch) to an event in the row above it, the event in the lower row is assumed to have occurred given that the event in the upper row has occurred.



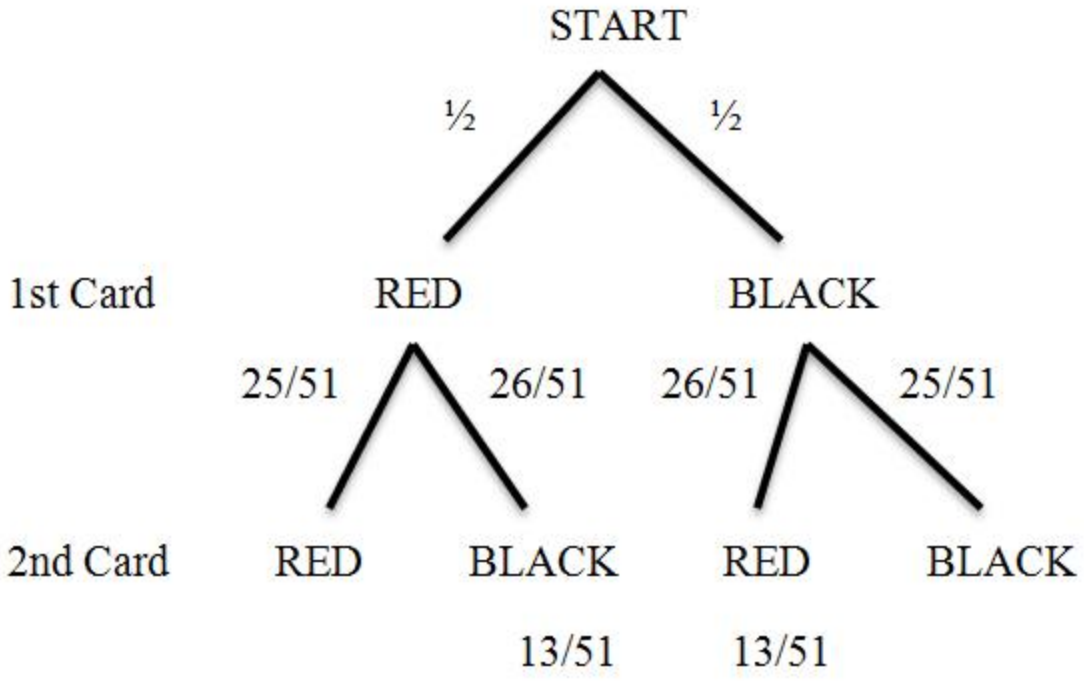
We know that the probability that the first card will be red is 1/2, and the probability that it will be black is 1/2, so we fill in this information on the relevant branches of the tree.



Given that the first card is red, the deck consists of 25 red cards and 26 black cards, so the probability of drawing a red card as the second card is 25/51, and the probability of drawing a black card as the second card is 26/51 . Similarly, given that the first card is black, the deck consists of 26 red cards and 25 black ones, so the probability of drawing a black card as the second card is 25/51, and the probability of drawing a red card as the second card is 26/51. We fill in this information along the various branches.



The probability of getting a red card as the first card and a black card as the second (given that the first card was red) is obtained by multiplying the numbers down that branch of the tree, getting 1/2 x 26/51 = 13/51. Similarly, the probability of getting a black card as the first card and a red card as the second (given that the first card was black) is also 13/51 . These are the only two outcomes relevant to the problem of computing the probability of getting one red card and one black card, and we fill in this information (underlined here) below the relevant branches of the tree.



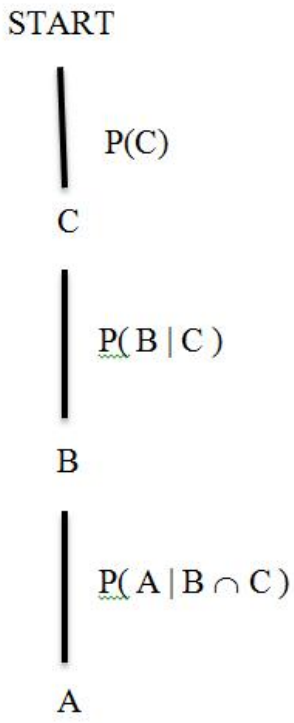
Therefore, the probability of drawing one red card and one black card is 13/51 + 13/51 = 26/51.

This technique of finding the probability of two events both occurring by using conditional probabilities can be extended to finding the probability of three (or more) events occurring. The basic formula used is

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

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

If we look at the relevant branch of a tree, we would have



Since A is below both B and C, P(A | B ∩ C) is the probability of A, given that both B and C have already occurred.

Example 3 - A card is drawn from a deck of cards, and then a second card is drawn without replacing the first card in the deck. What is the probability that both cards are spades?

Solution: It is easy to compute the probability that the first card is a spade, as this is a standard problem in a uniform probability space. 13 of the 52 cards in the deck are spades, so the probability that the first card is a spade is 13/52 = 1/4.

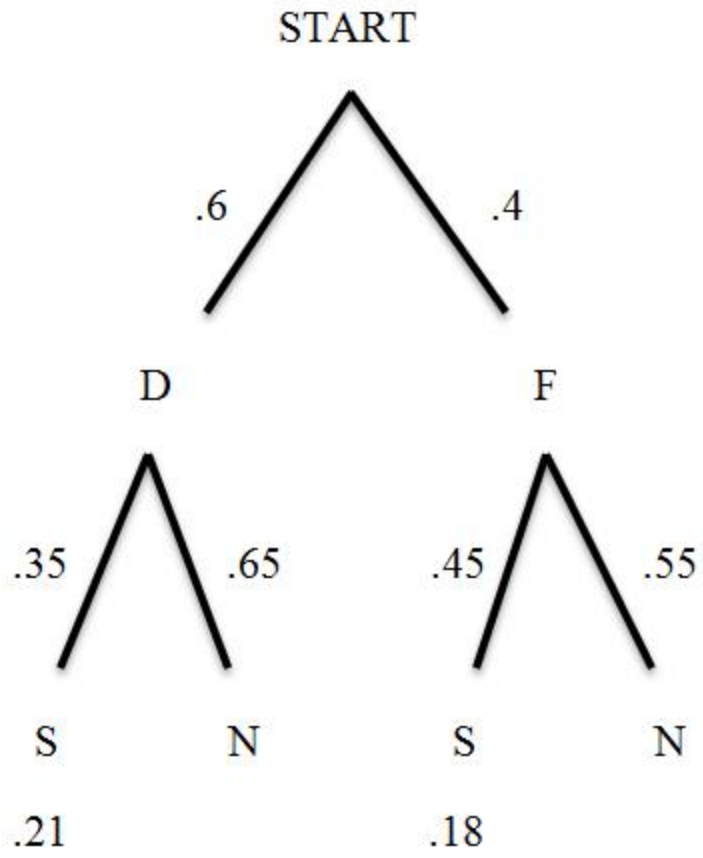
Suppose now that the first card was a spade. The deck now consists of 51 cards, 12 of which are spades, since we have removed one spade from the original deck. The probability that the second card is a spade, given that the first card is a spade, is therefore 12/51.

Suppose we let A be the event that the second card is a spade, and let B be the event that the first card is a spade. The analysis in the previous paragraph shows that P(B) = 1/4 and P (A | B) = 12/51 . The event A ∩ B is the event that both cards are spades. We know that P (A ∩ B) = P (A | B) P(B), and so the probability that both cards are spades is 1/4 x 12/51 = 3/51 = 1/17 .

We could also have done this example by using combination formulas. P(A ∩ B) = C(13,2)/C(52,2) = 1/17. ■

Example 4 - An auto dealer has found that 60% of his customers buy domestic cars and 40% buy foreign cars. 35 percent of the domestic cars purchased, and 45% of the foreign cars, are sports cars. What percentage of his customers buy sports cars?

Solution: Let D denote the purchase of a domestic car, F a foreign car, S a sports car, and N a car that is not a sports car. The statement "60% of his customers buy domestic cars" can be rephrased in the language of probability to read "the probability that a customer will buy a domestic car is .6". Drawing the tree, we obtain



So the probability that a customer will buy a sports car is .21 + .18 = .39. When phrased in the language of percentages, 39% of the customers buy sports cars. ▉

Independence

One of the intriguing aspects of probability theory is that some of its basic concepts almost seem to belong to the realm of philosophy! One such concept is that of independence.

Two events A and B in a sample space S are said to be **independent** if P (A | B) = P(A). In other words, if B occurs, it does not change the probability that A will occur.

In deciding whether two events are independent, we are often required to exercise a particular philosophical point of view. Suppose that we flip a coin twice, and let A denote the event of getting a head on the second flip, and B the event of getting a head on the first flip. Mathematicians and

scientists would shudder at the possibility that somehow the result of the second flip was influenced (had its probability altered) by the result of the first flip.

As a result, mathematicians and scientists adopt the following point of view: if one cannot find a causal explanation for dependence, one must assume independence. Notice that, if one draws two cards WITHOUT replacement from a deck of cards, the event that the second card is red is NOT independent of the event that the first card is red. If we do not know the color of the first card, the probability that the second card is red is 1/2. If we know the first card is red, the probability that the second card is red is 25/51. Moreover, we can explain this: the composition of the deck has changed.

Once the philosophical preliminaries are out of the way, we can concentrate on the mathematical consequences, which are of considerable importance. If A and B are independent, then

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

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

This result is extremely important; it states that the probability of two independent events both occurring can be obtained by multiplying the probabilities of the two events. This result is easily extended to three (or more) independent events. The probability of several independent events all occurring is the product of the probabilities of all the events.

Example 5 - Three dice are rolled. What is the probability that 6 comes up on all three dice?

Solution: The probability of rolling a 6 on any one die is 1/6. Since these events are independent, the probability of all three occurring is 1/6 x 1/6 x 1/6 = 1/216. ■

In the clear-cut world of theoretical probability, it is usually fairly obvious when events are independent. This is not always the case when one is dealing with empirical probabilities. In this case, the formula $P (A \cap B) = P(A) P(B)$ can be used to check whether the events are independent

by computing all three quantities, P(A), P(B), and P($A \cap B$), and seeing whether the above equation is satisfied. Of course, one would not expect empirical probabilities to result in an exact equality. Even if two events are known to be independent (such as two flips of a coin coming up heads), the empirical probabilities gathered from flipping coins are unlikely to satisfy this formula exactly. Techniques in statistics can enable us to determine whether independence is a reasonable assumption, based on a given set of empirical probabilities.

In the game of American roulette (European roulette is slightly different), a rotating wheel is divided into 37 equally-sized compartments. 18 of these compartments are red, 18 are black, and one is green. A bettor can place an even-money bet on red or black, and numerous other types of bets are available. However, if a player bets on red, there are 18 winning red compartments into which the ball can fall, but 19 losing compartments (18 black and one green). The house makes its money from offering inadequate odds; it offers odds of 1-1 when a fair bet would offer odds of 19-18.

The casinos of the world are filled with people who commit the classic gambler's fallacy of waiting for red to come up four times in a row, and then betting black on the fifth time. Of course, there are also those who will keep a record of all the spins of a particular wheel, hoping to detect a statistically significant fluctuation from the average which will indicate that the wheel has been improperly balanced. However, even if a wheel has been improperly balanced, the spins of the wheel are still independent events, but the probability of red may no longer be the 18/37 that characterizes a perfectly-balanced wheel.

Example 6 - a) What is the probability that the ball will fall into a red compartment on a roulette wheel four times in a row?

b) If the ball <u>does</u> fall into a red compartment four times in a row, what is the probability that it will fall into a red compartment on the next spin of the wheel?

Solution: a) The probability of the ball falling into a red compartment is 18/37. Since each spin of the wheel is independent of the results of the other spins (the wheel has no memory), the probability of four consecutive reds is $(18/37)^4$, which is approximately .056 .

b) The fifth spin of the wheel is independent of the preceding four spins, so the probability of the ball falling into a red compartment is *still* 18/37! ▉

Independence bears on the question of cause-and-effect. We shall investigate this problem further in the next section.

Section 2 - **Bayes' Theorem**

Let's assume that Bayside Hospital (an appropriate name, considering the topic for this section) has gone over its patient records. It has found that 20% of the patients admitted to the hospital have pneumonia, and 90% of pneumonia patients complain of a sore throat. Of the 80% of the patients who do not have pneumonia, 10% complain of a sore throat.

Using the methods of the previous section, it is easy to compute that the probability that a patient will complain of a sore throat is .2 x .9 + .8 x .1 = .26 . However, the problem that a doctor will encounter in the real world is not to compute the probability that a patient will complain of a sore throat. The actual problem the doctor will face is to decide whether a patient who complains of a sore throat has pneumonia.

We can use the information above to compute the probability that a patient who complains of a sore throat has pneumonia. If 100 patients are admitted to the hospital, 20 will have pneumonia and 80 won't. Of the 20 who have pneumonia, 18 will complain of a sore throat. Of the 80 who don't have pneumonia, 8 will complain of a sore throat. Therefore, 18 + 8 = 26 people will complain of a sore throat, and 18 of these will have pneumonia. Consequently, the probability that a patient who complains of a sore throat will have pneumonia is 18/26 = 9/13 , or about .69 .

We have both cause and effect in the previous problem: pneumonia is a cause, and a sore throat is an effect. However, there are other potential causes which produce the same effect. When we found the probability that a patient admitted to the hospital would complain of a sore throat, we computed the probability that a certain effect would occur. When we found the probability that a patient who complained of a sore throat had pneumonia, we computed the probability that a given effect could be ascribed to a particular cause.

Bayes' Theorem, discovered by the Reverend Thomas Bayes, an 18th-century British clergyman, computes the probability that a given effect can be ascribed to a particular cause. The actual statement of Bayes' Theorem does not talk about causes and effects, but rather about probabilities. Many of the most important applications of Bayes' Theorem occur in situations where effects are observed, and one wishes to compute the probabilities concerning the causes responsible.

Bayes' Theorem

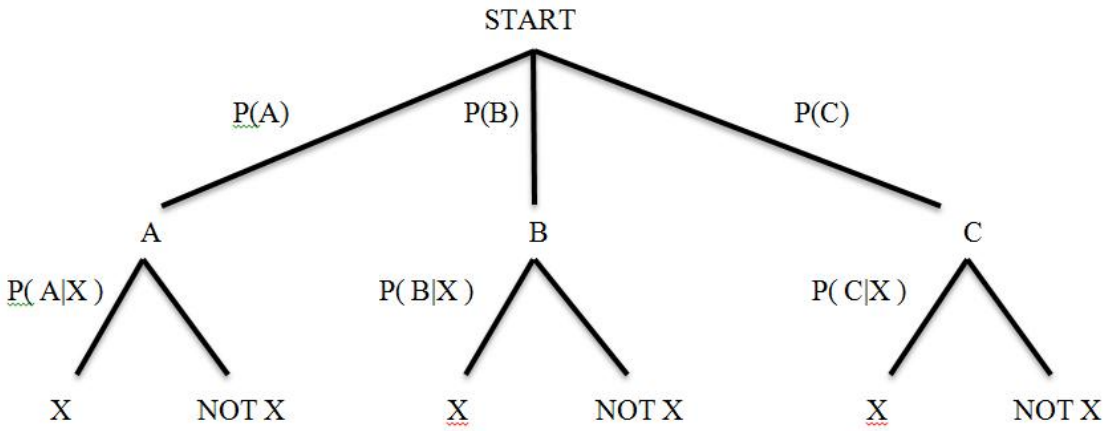
Let S be a sample space, and let A, B, C and X be events in S. Then

$$P(A|X) = \frac{P(X|A)P(A)}{P(X|A)P(A) + P(X|B)P(B) + P(X|C)P(C)}$$

$$= \frac{P(A \cap X)}{P(A \cap X) + P(B \cap X) + P(C \cap X)}$$

The above formula can be extended to any number of events, rather than just A, B, and C. If the events are labeled $A_1, \ldots, A_n$, where the role of A in the first equation is now being played by A_1, then the denominator of the fraction in the first equation is $P(X \mid A_1)\, P(A_1) + \ldots + P(X \mid A_n)\, P(A_n)$.

We can draw a tree to represent Bayes' Theorem. If the possible causes of effect X are A, B, and C, then a tree depicting this would be

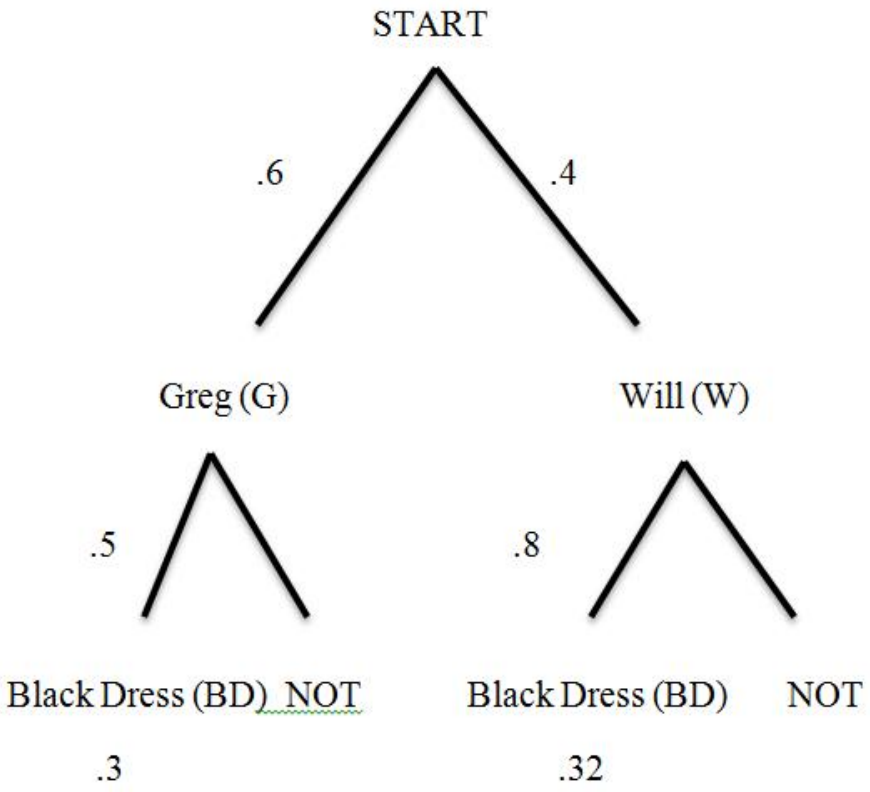


This suggests that Bayes' Theorem can be used to analyze situations in which the events A, B, C, ... precede X in time, and earlier events are often seen as causes of later ones.

Bayes' Theorem has many potential applications. We study a few of them in the following examples.

Example 1 - Ellen has a 60% chance of going to the Homecoming Dance with Greg, and a 40% chance of going with Will. If she goes with Greg there is a 50% chance she will wear her black dress, but if she goes out with Will, there is an 80% chance she will do so. You run into Ellen during a break, and she is wearing the black dress. What is the probability that she is with Will?

Solution: We'll analyze this by drawing a tree.



So Ellen's probability of being with Will is

$$P(W \mid BD) = P(W \cap BD) / (P(W \cap BD) + P (G \cap BD))$$

$$= .32 / .62 = 16/31 .$$

Notice that it was not necessary to fill in the probabilities next to irrelevant branches of the tree. ■

Example 2 - A factory has three machines producing microprocessors. Machine A contributes 40% of the total production, machine B contributes 35%, and machine C contributes 25%. Machine A produces a defective microprocessor 0.08% of the time, machine B produces a defective 0.1% of the time, and machine C produces a defective 0.12% of the time. Every time a defective is produced, the machine producing it must be fixed. Assuming it requires an equal amount of time to locate and fix each machine, in what order should the machines be repaired?

Solution: If we let A, B, and C denote the events in which machines A, B, and C produce microprocessors, and X the event in which a defective microprocessor is produced, then

$P(A \mid X) = .4 \times .0008 / (.4 \times .0008 + .35 \times .001 + .25 \times .0012)$

$= 32/97$

$P(B \mid X) = .35 \times .001 / (.4 \times .0008 + .35 \times .001 + .25 \times .0012)$

$= 35/97$

$P(C \mid X) = .25 \times .0012 / (.4 \times .0008 + .35 \times .001 + .25 \times .0012)$

$= 30/97$

So one should repair B first, as it is the most likely to have contributed the defective microprocessor. If the problem is not fixed, then repair A, and finally C. ■

Example 2 represents one of the basic applications of Bayes' Theorem to industry. Notice that this application also emphasizes the 'cause-and-effect' theme: having isolated an effect (a defective microprocessor), find the cause most likely to have led to that effect.

Example 2 again reiterates the principle that one can deduce causes from effects, at least on a probabilistic level. The type of situation discussed in this example obviously can be applied to any type of behind-the-scenes maneuvering in which a certain decision is reached.

Example 3 - Adams, Baker, and Corey are locked in a struggle for control of Rutabaga Pharmaceuticals. Adams controls 45% of the stock, Baker 35%, and Corey 20%. They are considering bringing a new drug, Rutabagol, onto the market. If Adams gains control, there is a .5 probability that he would develop Rutabagol. Baker only has a .3 probability of developing Rutabagol, and Corey has a .6 probability of developing Rutabagol. If Rutabagol is not developed, what is the probability that Corey has gained control of the company?

Solution: Let A denote the event that Adams gains control of the company (similarly for B and C), and let R be the event that the company decides to develop Rutabagol. In the absence of any additional information, the best estimate of P(A) is .45, the fraction of stock owned by Adams (similarly, P(B) = .35 and P(C) = .2). We also see that P(NOT R | A) = .5,

P(NOT R | B) = .7, P(NOT R | C) = .4. Therefore, by Bayes' Theorem

P(C | NOT R) = .2 x .4 / (.45 x .5 + .35 x .7 + .2 x .4) = 8/55, or about .145 . ■

It is worth noting that Bayes' Theorem is somewhat controversial, or at least as much as a mathematical theorem can be controversial. There are those who believe that, once the effect has occurred, any discussion of the probability of prior causes is meaningless. This is a minority view, as two types of decisions from this section -- medical diagnosis and maintenance scheduling -- are commonly made, and made profitably, from a Bayesian perspective. Bayes' Theorem also plays a key role in forensic analysis.

Chapter 11 – Statistics

Introduction - Lies, Damned Lies, and Statistics

We live in a world in which we are continually exposed to statistics. On a typical day, we might receive statistical information from the stock market (the Dow-Jones averages), the entertainment industry (the Nielsen ratings), sports (baseball batting averages), economic reports (cost-of-living indexes), and a multitude of other areas. Everybody uses statistics to make a point, generally the point they want to make.

As a result, statistics could probably use some good P.R., because many people feel that statistics are used to promote a particular point of view and cover up the truth. That viewpoint is expressed in the sentiment so brilliantly expressed by Disraeli; there are three kinds of lies: lies, damned lies, and statistics.

Statistics is frequently used to summarize data so that the data becomes more useful. In a world increasingly devoted to collecting and processing data, we are literally inundated with data. It is not possible to understand a mass of data in 'raw' form, not because it is intrinsically incomprehensible, but simply because there is so much of it. To understand the results of the last presidential election (2012), we do not want to know how all 100 million voters, as individuals, voted. We do not even want to know the vote totals for Obama and Romney. We want to know the percentage of voters who voted for each candidate.

These percentages constitute one of the fundamental tools of statistics: the **probability distribution**. We shall define this more accurately later, but the term itself indicates that there is a connection between statistics and probability. We recall that probability had both an empirical and a predictive aspect, and the same can be said of statistics. The two basic problems of statistics are how to summarize data, and which statistical measures of data can be used to predict future behavior.

Section 1 - Random Variables, Distributions, and Graphs

Throughout this section, S will be the sample space of an experiment whose possible outcomes are $O_1, O_2, \ldots, O_n$. During a basketball game one evening, LeBron James made 8 2-point field goal attempts and missed 7, and made 2 3-point field goal attempts and missed 3. The sample space of this experiment is an attempted field goal by LeBron James, and the possible outcomes are

O_1 = made a 2-point field goal

O_2 = missed a 2-point field goal

O_3 = made a 3-point field goal

O_4 = missed a 3-point field goal

This information can be obtained from the box score in the paper next day. A mathematician might summarize it in terms of a function X whose domain consists of the numbers 1, 2, 3, and 4 (the possible outcomes), and whose values are given by

$$X(1) = 8 \quad X(2) = 7 \quad X(3) = 2 \quad X(4) = 3$$

The function X we have defined, which is associated with the sample space S, is called a **random variable**. A random variable is simply a function whose domain (the allowable inputs) are outcomes from a sample space, and whose range (the outputs of the function) are numbers. It is customary to use capital letters at the end of the alphabet, such as X, Y, and Z, to denote random variables.

When the range of a random variable consists of non-negative whole numbers, we use the term **frequency distribution** to describe the random variable. The random variable X defined above is a frequency distribution. X(1) is the frequency of 2-point field goals LeBron James made, X(2) is the frequency of 2-point field goals LeBron James missed, etc.

In the basketball game to which we have been referring, X(1) + X(2) + X(3) + X(4) = 8 + 7 + 2 + 3 = 20, which is the number of field goals LeBron James attempted. If we define Y(1) = X(1)/20 = 8/20 = .4, a probabilistic interpretation of Y(1) is that, if one were to select a LeBron James field goal attempt at random, that attempt would be a successful 2-point field goal with probability .4 . (The percentage interpretation would be that 40% of LeBron James's field goal attempts were successful 2-pointers.) Similarly, if Y(2) = X(2)/20 = .35, Y(3) = X(3)/20 = .1, and Y(4) = X(4)/20 = .15, then Y(1) + Y(2) + Y(3) + Y(4) = .4 + .35 + .1 + .15 = 1. The function Y is simultaneously a random variable and a probability function. A random variable that is also a probability function is called a **probability distribution**.

Let X be the random variable associated with LeBron James's field goal attempts. By borrowing the letter P from probability, we can discuss the various probabilities associated with LeBron James's field goal attempts without having to introduce another letter (such as Y). The notation

$$P(X = 1) = .4$$

is read "the probability that the random variable X assumes the value 1 is .4". This notation is commonly used to describe probability distributions associated with frequency distributions.

Example 1 - Rutabaga Biotech has 27 employees whose salaries are less than $40,000 a year, 16 employees whose salaries are between $40,000 and $80,000 a year, and 7 employees whose salaries are more than $80,000 a year. Describe the sample space, frequency distribution, and probability distribution associated with this experiment.

Solution: O_1 = an employee has a salary of less than $40,000 a year, O_2 = an employee has a salary of between $40,000 and $80,000 a year, and O_3 = an employee has a salary of more than $80,000 a year. The frequency distribution is the random variable X such that

$$X(1) = 27 \quad X(2) = 16 \quad X(3) = 7$$

Since 27 + 16 + 7 = 50, the probability distribution associated with this random variable is P(X = 1) = 27/50 = .54, P(X = 2) = 16/50 = .32, and P(X = 3) = 7/50 = .14 . ■

Notice that in Example 1, we had a choice of how to describe the outcomes of the sample space. One possibility was to describe the possible outcomes as O_1 = a salary of \$1/year, O_2 = a salary of \$2/year, etc., up through the maximum salary that an employee at Rutabaga Biotech makes. This has the obvious inconvenience of having an experiment with over 80,000 different outcomes. Alternatively, we might only have used the actual salaries as possible outcomes, which would limit the sample space to 50 outcomes if everyone had a different salary. The actual procedure we selected, using a range of possible values as a particular outcome, is known as **binning** (shorthand for 'to place in a bin'). In Example 1, a judicious choice of bins has made it relatively clear what the salary structure at Rutabaga Biotech is, using a sample space with only 3 different outcomes.

Example 2 - A deck of cards is shuffled and a card selected. It is then replaced, and this procedure is repeated ten times. The cards selected were: KD, 3H, 4S, AC, JD, 7S, 9C, 4H, QD, 8D. Describe the sample spaces, frequency distributions, and probability distributions if binning is done

a) by suit

b) by rank

Solution: a) The sample space is O_1 = clubs, O_2 = diamonds, O_3 = hearts, and O_4 = spades. The frequency distribution is X(1) = 2, X(2) = 4, X(3) = 2, X(4) = 2. The probability distribution is P(X = 1) = .2, P(X = 2) = .4, P(X = 3) = .4, P(X = 4) = .2 .

b) The sample space is O_1 = ace, O_2 = deuce, ... , O_{13} = king. We list the frequency and probability distributions in a more convenient form below.

k	1	2	3	4	5	6	7	8	9	10	11	12	13

X(k)	1	0	1	2	0	0	1	1	1	0	1	1	1
P(X = k)	.1	0	.1	.2	0	0	.1	.1	.1	0	.1	.1	.1

Visual Display of Statistical Information

Statistical information is often packaged in one of several different visual formats, as it is easier to comprehend the entire picture when we can see it all at once.

Line Graphs

A **line graph** displays a random variable as a sequence of points connected by line segments, as in Fig. 10-2, which shows the afternoon temperature on a pleasant summer afternoon in Antarctica.

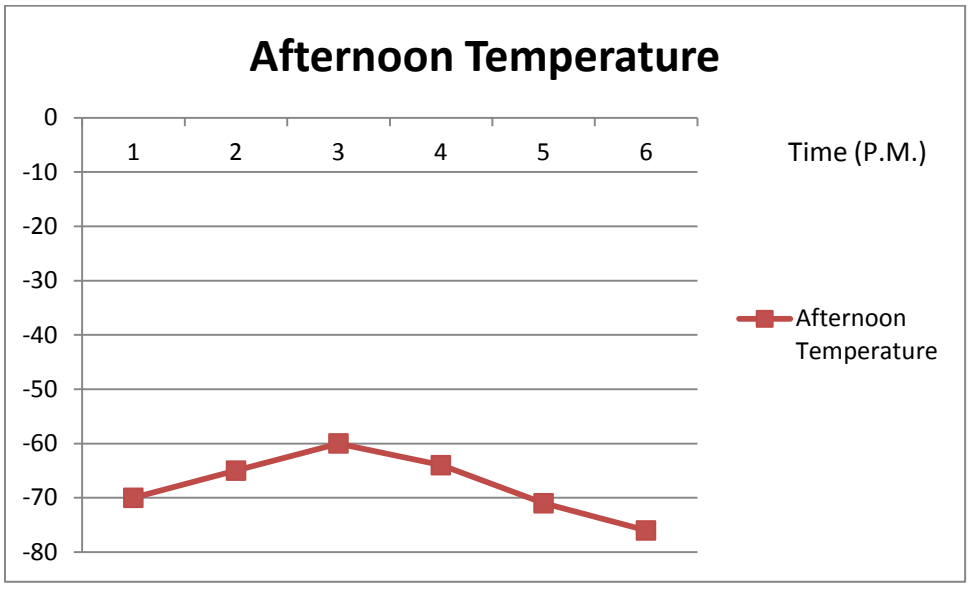


Fig. 10-1

Brr! Well, at least it's pleasant for the penguins. Notice that, in a line graph, it is clearly evident that a random variable is a function. Two perpendicular axes are used to graph the information.

Line graphs are very useful when the outcomes can be described as successive values of an independent variable. In Fig. 10-1, for example, the successive values are the hours 1 P.M., 2 P.M., 3 P.M., etc. The lines that connect successive points on the graph imply a progression of some sort between the successive points.

Bar Charts

A **bar chart** displays a random variable as a separated collection of parallel bars (both vertical and horizontal parallel bars are used in drawing bar charts). Fig. 10-3 shows a bar chart for the following probability distribution.

Example 3 – A sample probability distribution to illustrate types of charts.

X	1	2	3	4	5
P(X)	0.2	0.3	0	0.1	0.4

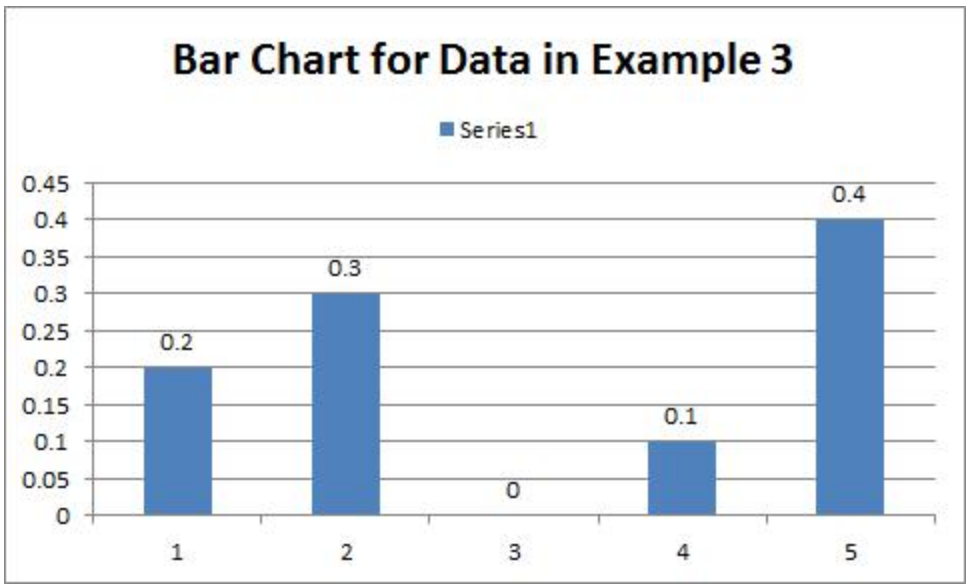


Fig. 10-2

Pie Charts

A **pie chart** is often used to display random variables in which the values of the random variable are measured in percentages. The size of the 'slice' of the pie corresponds to the percentage assigned to a particular outcome in the sample space. The following is a pie chart for the same probability distribution used for the bar chart in Fig. 10-2.

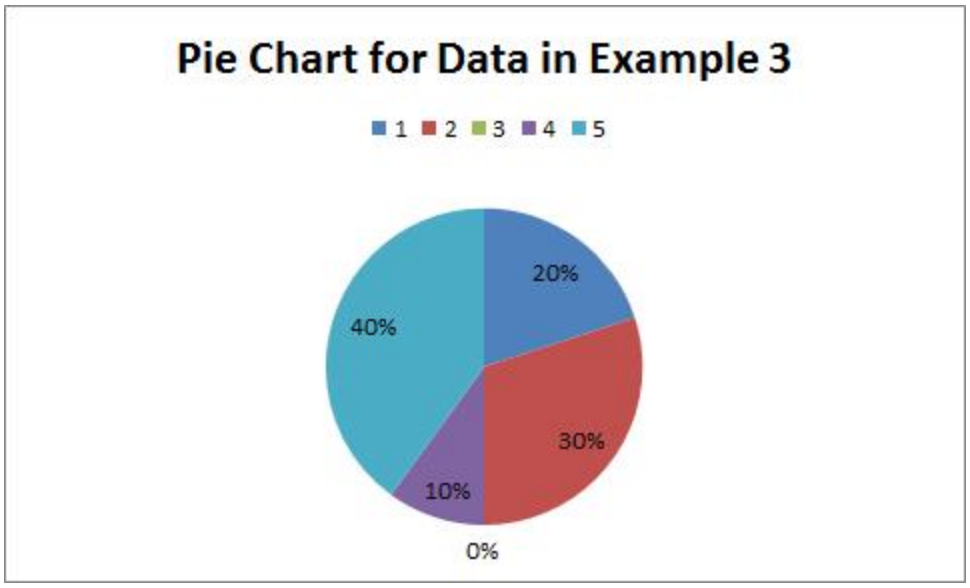


Fig. 10-3

Subterfuge With Statistics

Statistics offer the unscrupulous and the semi-scrupulous a wonderful opportunity to present one's case in a favorable light. The two most frequent ways that statistical information is used to convey an erroneous impression are by selection of specific statistics and misleading presentation of a statistical package.

When one wishes to present data to make a point, one has an almost limitless choice of statistical information. Were the 1980s a period of economic growth? The Republicans can point to 30 million new jobs created, the Democrats to a widening gap between rich and poor.

The complexities of any economy generate masses of data, and there are usually statistics available to emphasize any position. This statistical profusion may be part of the reason for the staggering increase in the salaries of mediocre baseball players. After all, it is certainly possible that a .215 hitter may have a batting average of .380 with men on base and two out in the eighth and ninth innings.

Section 2 - Measures of Central Tendency and Dispersion

During the height of a political campaign, one is deluged with 'sound bites', those little morsels which condense an extremely complex position into one memorable phrase or slogan. Because frequency distributions can also be complex, there is substantial interest in finding numbers which can serve as 'data bites', compressing much of the information of the distribution into a very few quantities. There are two basic types of 'data bites': measures of **central tendency**, which locate the middle of the distribution, and measures of **dispersion**, which tell how tightly packed the distribution is around its middle.

Measures of Central Tendency: The Mean, The Median, and the Mode

The **mean** of a distribution is just our old friend, the average. If we are given numbers X_1, ... , X_n, the mean, denoted by μ (the Greek letter mu), is simply

$$\mu = (X_1 + ... + X_n)/n$$

The **median** is the 'middle' number of the distribution; it is the number such that 50% of the distribution lies above it, and 50% below it. To compute the median, arrange the data low to high. If there are an odd number of data points, the median is the middle data point. If there are an even number of data points, the median is the average of the two middle data points. Thus, if the data points are

14, 17, 23, 26, 27, 31, 38 (7 data points)

↑

middle data point

the median is 26 (the same number of data points are above 26 as are below it). If the data points are

14, 17, 17, 21, 24, 29, 29, 29 (8 data points)

↑ ↑

middle data points

the median is 22.5, the average of 21 and 24. Note that the value 17 has frequency 2 and the value 29 has frequency 3. In determining the median all data points are listed, listing each value the number of times indicated by its frequency.

The **mode** of a distribution is the value which has the highest frequency. A distribution can have several modes. For example, the distribution

8, 12, 15, 15, 21, 23, 23, 28, 32, 32, 37

has modes 15, 23, and 32. If all values in a distribution have the same frequency, the distribution has no mode.

Example 1 - In 6 consecutive home games, the Chicago Cub pitching staff allowed 4, 8, 9, 6, 6, and 9 earned runs (the wind was blowing out in Wrigley Field). Calculate the mean, median, and mode of this distribution.

Solution: The mean $\mu = (4 + 8 + 9 + 6 + 6 + 9) / 6 = 7$. The median is found by arranging the data points in increasing order

4, 6, 6, 8, 9, 9

↑ ↑

middle data points

Since there are 6 data points, the average of the two middle values 6 and 8 is 7, so the median is 7. The values 6 and 9 both have frequency 2, so the distribution is **bimodal** (having 2 modes), with modes 6 and 9. ■

A ranking of the importance of these three quantities would undoubtedly have the mean first, followed by the median, with the mode finishing dead last. To a statistician, the value of a statistical 'data bite' is the reliability and strength of the statements that can be made about it. The formula for computing the mean never changes (unlike the median), and it is always a single number (unlike the mode). As a result, it is easier to make precise mathematical statements about the mean than about either the median or the mode.

Example 2 - Maria's scores on 5 chemistry quizzes were 95, 40, 92, 88, and 95. Compute the mean, median, and mode of the distribution. Which one seems to be the most reliable measure of the 'middle' of the distribution?

The mean $\mu = (95 + 40 + 92 + 88 + 95) / 5 = 82$. Arranging the scores in order, we get

40, 88, 92, 95, 95

There are an odd number of data points, so the median is the middle score 92. Finally, the number 95 has frequency 2, so it is the mode. ■

Clearly, the mode is not a good indicator of the middle of the distribution. If we assume that Maria had an 'off' day when she scored 40, we would be most likely to select the median score of 92 as most representative of Maria's abilities as a chemistry student. Medians are useful in cases such as this, when a few extreme scores distort the mean.

The mean has a useful physical interpretation. If we imagine a score of 80 in a distribution as a weight of 1 pound suspended 80 centimeters from the left-hand side of a long rod, the entire distribution resembles a collection of 1 pound weights suspended at different locations on the rod. The mean of the distribution is where to place a fulcrum so that the rod is in perfect balance.

Measures of Dispersion: Range and Standard Deviation

If we look at Maria's chemistry scores in Example 2, we observe that her scores ranged from awful (40) to excellent (95). The difference between the highest and lowest numbers in a data set is called the **range**. In this case, the range of Maria's scores is 95 - 40 = 55.

If the range is small, the data points are obviously packed close to the mean. However, it is possible for the range to give an inaccurate picture of the distribution. In the entertainment industry, for example, most actors are extras who do not make enough at the job to support themselves, but the superstars make megabucks. The range is huge, but most of the data points are clustered at the very low end of the distribution. Obviously, a better method of evaluating the dispersion of a distribution than the range is needed.

Again, let's look at Maria's chemistry scores in Example 2: 40, 88, 92, 95, and 95. As we have seen, the distribution has a mean of 82. The **deviation** of each score is the difference between that score and the mean. The deviation of 40 is 40 - 82 = -42. The deviation of 88 is 88 - 82 = 6. The deviation of a score (it is actually the score's 'deviation from the mean') is negative if the score is less than the mean and positive if the score is greater than the mean. Of course, if the score is exactly equal to the mean, its deviation is 0.

We now make a table Maria's scores and their deviations.

Score	Deviation
40	-42
88	6
92	10
95	13
95	13

If we add all the deviations, the total is -42 + 6 + 10 + 13 + 13 = 0. This is always true, and can be easily shown using fairly simple algebra. In fact, it characterizes the mean, which is the unique number around which the sum of the differences between that number and the scores totals 0.

When analyzing dispersion, we are interested in how far away a number is from the mean. In a distribution with a mean of 80, 88 is 8 points *above* the mean, 72 is 8 points *below* the mean, but both 88 and 72 are 8 units *away* from the mean. When we regarded the mean as the point where we place the fulcrum to make the rod balance, we were interested in how far away a weight is from the fulcrum, not whether it is to the left of the fulcrum or to the right of it.

One obvious measure of dispersion is the average distance of the points from the mean. Again using Maria's chemistry quizzes, the distances of the scores from the mean are 42, 6, 10, 13, and 13. Their average is (42 + 6 + 10 + 13 + 13) / 5 = 16.8 . When the average distance from the mean is small, the data points in the distribution are closely packed around the mean. Conversely, when it is large, the data points tend to be far from the mean.

The average distance from the mean is a natural measure of dispersion. Unfortunately, while it is useful qualitatively, it is not so useful quantitatively: it is difficult to make predictions (mathematical statements about probabilities) based on the average distance from the mean.

More useful is the **standard deviation**, which is denoted by σ (the Greek letter lower-case sigma), and is defined according to the equation

$$\sigma = \sqrt{ [((X_1 - \mu)^2 + ... + (X_n - \mu)^2)/ (n-1)] }$$

This rather-imposing looking formula is calculated by the following procedure.

Computing the Standard Deviation of a Sample

Step 1 - Find the mean.

Step 2 - Compute the sum of the squares of all the deviations.

Step 3 - Divide the result of *Step 2* by 1 less than the number of data points.

Step 4 - σ is the square root of the result of *Step 3*.

The Nobel-Prize winning physicist Isador Rabi was once confronted with experimental evidence for a bizarre subatomic particle. His response was, "Who ordered that?" A formula such as the one for the standard deviation often provokes a similar response in students. However, computers can calculate standard deviations very easily, and it is not much bother with a hand calculator. In fact, many hand calculators are programmed to find the standard deviation of a data set. One simply keys in the data points, and the result appears.

Example 3 - What is the standard deviation of Maria's scores on chemistry quizzes in Example 2?

Solution: We have already computed the mean to be 82, and the deviations to be -42, 6, 10, 13, and 13. The squares of the deviations are 1764, 36, 100, 169, and 169. The sum of the squares of the deviations are 2,238. Since there are 5 data points in the sample, we divide by 5 - 1 = 4, obtaining 559.5 . Finally, we take the square root of this number, obtaining σ = 23.65 .

Although the standard deviation appears unnaturally complicated, it has the great advantage (over the range or the average distance from the mean) that predictions can be made from it.

Section 3 - The Normal Distribution

When Pete sent Freddy to report on how many free throws Theresa Middlebury made out of 100, he knew Freddy would report that Theresa would have made a whole number of free throws. She would not have made 72.1 or 86.437 free throws. The number of free throws Theresa made out of 100 is an example of a **discrete** random variable; a random variable that can only assume a fixed, finite number of values (in Theresa's case, between 0 and 100).

In contrast, had Pete sent Freddy to measure Theresa's height, any value would conceivably have been possible, such as 64.18 or 67.304 inches, assuming that Freddy had a sufficiently finely-calibrated ruler. Admittedly, we normally measure height to within the nearest inch or half-inch, but this is by choice -- it is certainly within our capability to measure more accurately. A random variable which can assume any value (within a given range) is said to be **continuous**.

Many continuous random variables have a distribution shaped like the one in Fig. 10-4 below, which is a hypothetical distribution of the heights of basketball players. The curve is known as a **probability density function**. It has the property that the total area under the curve is 1. The probability that a randomly-selected basketball player, from a distribution with a mean of 78 inches and a standard deviation of 4 inches, has a height between 74 and 81 inches is the shaded area in Fig. 10-4. This could also be interpreted as the fraction of basketball players in the distribution between 74 and 81

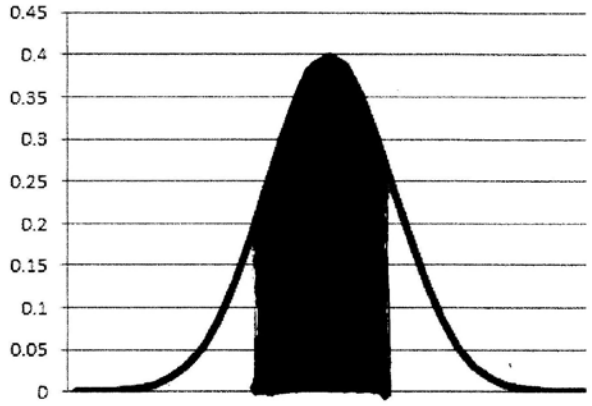


Fig. 10-4

The bell-shaped curve in Fig. 10-4 is a special type of distribution known as a **normal** distribution. Normal distributions are extremely important because not only are they characteristic of many continuous random variables, they also provide excellent approximations to many discrete random variables (Pete talked about this in the story).

We recall from the previous section that, if we are given two numbers μ and σ, there can be many different distributions with mean μ and standard deviation σ. However, there is only <u>one</u> normal distribution with mean μ and standard deviation σ. Consequently, when we know the mean and standard deviation of a normal distribution, we know precisely which bell-shaped curve to draw, since there is only one. Once the curve is drawn, we can answer any question about the distribution.

Suppose that we discover that basketball players have a mean height of 76 inches with a standard deviation of 4 inches. *Every* height can be expressed in terms of a number of standard deviations above or below the mean. For instance, 74 inches is 2 inches less than the mean of 76 inches, and 2 inches is $2/4 = 0.5$ standard deviations. Therefore, $74 = \mu - 0.5\,\sigma$. Similarly, 81 inches is 5 inches more than the mean of 76 inches, and 5 inches is $5/4 = 1.25$ standard

deviations. Therefore, $81 = \mu + 1.25\,\sigma$. In order to find the fraction of basketball players between 74 and 81 inches, we must find the fraction of the normal curve that lies between $\mu - 0.5\,\sigma$ and $\mu + 1.25\,\sigma$.

The Language of Normal Distributions

Any normal distribution is completely characterized by its mean μ and standard deviation σ. Any measurement from a normally-distributed population can be expressed as $\mu + z\,\sigma$, where z is positive if the measurement is above the mean, and negative if it is below the mean.

Example 1 - The weights of defensive linemen are normally distributed, with a mean of 260 pounds and a standard deviation of 20 pounds. Express the weights of a 283-pound lineman and a 231-pound lineman in the form $\mu + z\,\sigma$.

Solution: We have $283 = 260 + 23 = 260 + 23/20 \times 20 = 260 + 1.15 \times 20 = \mu + 1.15\,\sigma$. We say that the z-value of 283 is 1.15 in this case. Similarly, $231 = 260 - 29 = 260 - 29/20 \times 20 = \mu - 1.45\,\sigma$. In this case, the z-value of 231 is -1.45. Note that we divided and multiplied by 20 because $\sigma = 20$ for this distribution. ■

Example 2 - In the distribution in Example 1, what is the actual weight of a lineman whose z-value is 2.85?

Solution: The actual weight is $\mu + 2.85\,\sigma = 260 + 2.85 \times 20 = 260 + 57 = 317$ pounds. ■

The table below gives the cumulative probability that a number chosen from a normally-distributed population will have a z-value less than or equal to a specific value (this table is abridged; more complete tables exist in profusion on the Web). It can be used in a variety of situations, one of which has been alluded to earlier and is illustrated in Example 3.

Table of Cumulative Probabilities for a Normal Distribution

z-value	Cumulative Prob.
-3.00	0.0013
-2.75	0.0030
-2.50	0.0062
-2.25	0.0122
-2.00	0.0228
-1.75	0.0401
-1.50	0.0668
-1.25	0.1056
-1.00	0.1587
-0.75	0.2266
-0.50	0.3085
-0.25	0.4013
0.00	0.5000
0.25	0.5987
0.50	0.6915

0.75	0.7734
1.00	0.8413
1.25	0.8944
1.50	0.9332
1.75	0.9599
2.00	0.9772
2.25	0.9878
2.50	0.9938
2.75	0.9970
3.00	0.9987

The above table is only for z-values that end in .00, .25, .50, or .75. More complete tables are easily found online simply by entering "normal distribution table" into a search engine.

Example 3 – The heights of basketball players are normally distributed with a mean of 78 inches and a standard deviation of 4 inches. What percentage of basketball players have heights between 74 inches and 81 inches?

Solution: Earlier, we discovered that in order to find the fraction of basketball players between 74 and 81 inches, we must find the fraction of the normal curve that lies between $\mu - 0.5\,\sigma$ and $\mu + 1.25\,\sigma$.

From the table, we see that the fraction of the normal curve that lies to the left of $\mu - 0.5\,\sigma$ is 0.3085, and the fraction of the normal curve that lies to the left of $\mu + 1.25\,\sigma$ is 0.8944. So the

fraction of the curve lying between $\mu - 0.5\,\sigma$ and $\mu + 1.25\,\sigma$ is $0.8944 - 0.3085 = .5859$. So 58.59% of basketball players have heights between 74 and 81 inches. ▉

Notice that it is not possible to say what percentage of the players are *exactly* 6'9" tall, because this would correspond to the area of a single vertical line, which is 0. This is not so surprising, because it is highly unlikely that *any* basketball player is exactly 6'9" tall -- they are more likely to be 6'8.982" tall or 6'9.071" tall. However, if we agree that 6'9" tall (81 inches) is to be interpreted as between 80.5 and 81.5 inches, then this represents an actual area under the curve, which we can find by using a sufficiently detailed normal distribution table.

The Central Limit Theorem

Not all populations are normally distributed. One example would be the revenue generated by movies. A movie is generally a hit or a flop, and is much more likely to be a flop than a hit. However, if one takes any distribution whatsoever, and takes samples of size 30 or more, the distribution of the means of those samples will be normally distributed! For example, if one were to choose 30 movies at random, and compute their mean revenue, then another 30 movies at random, and compute their mean revenue, etc., these mean revenues would be normally distributed. This delicious result is known as the **Central Limit Theorem**, and is due to Gauss.

Hypothesis Testing - Part 1

The scientists at Rutabaga Pharmaceuticals have recently developed Rutabagol, a new drug which they believe will increase life expectancy. At least, it seems to work on hamsters. Hamsters have a mean life of 300 days. They have raised 25 hamsters on Rutabagol, and the mean life expectancy of those 25 hamsters is 315 days, with a standard deviation of 20 days. How significant a result is this?

Admittedly, the scientists might have gotten lucky and just picked a bunch of healthy hamsters which would have lived longer than average whether they were given Rutabagol or not. How

lucky would they have had to be? We make two assumptions, both of which can be experimentally and mathematically substantiated.

- First, we assume that the life spans of hamsters are normally distributed.

- Second, we assume that the standard deviation of those life spans is the standard deviation of the sample (20 days) divided by the square root of the number of hamsters in the sample (25). This new standard deviation is known as the **standard error of the mean**, and is the standard deviation that is actually used in such problems. This gives a standard deviation for hamster life spans of $20/\sqrt{25} = 20/5 = 4$. Since $315 = 300 + 15 = 300 + 15/4 \times 4 = \mu + 3.75\ \sigma$, we can tell from the table of z-values that this would happen by chance less than 1 time in 1000.

When statistics is used to test a hypothesis, it does so by deciding what level of probability is needed before it is agreed that the result of the experiment didn't happen simply by chance. Some hypotheses are accepted if they would only have happened 1 time in 20 due to chance alone (this is called a 5% **significance level**). Social sciences usually accept hypotheses at a 5% significance level. Other disciplines need a 1% significance level, or even less, before they are willing to accept a hypothesis. In California, courts require a 1% significance level in paternity testing before an alleged father is relieved of responsibility for the child -- DNA testing must show that an alleged father has less than a 1% chance of being the father.

We see continual evidence of hypothesis testing in our daily lives: in the market surveys which determine which products companies will produce and how they will advertise them, in the Nielsen ratings which determine what TV shows we will see, and in the opinion polls which influence the activities of the politicians. In the next section we will discuss these examples of hypothesis testing.

Hypothesis Testing for Normally-Distributed Measurements

Suppose we wish to test whether measurements taken on a sample from a normally-distributed population are unusual. We must first decide on what we mean by unusual; 1 chance in 20 (the 5% significance level), 1 chance in 100 (the 1% significance level), or any other value.

Next we take measurements on a sample of size n, computing the sample mean m and the sample standard deviation s according to the computational rules of the previous section. The standard deviation σ of the normal curve is given by $s/\sqrt{n}$. If we know the mean μ of the normal distribution, we can compute the z-value of the sample mean m, and then decide whether this z-value is unusual (at our predetermined significance level) by simply looking up the area under the normal curve corresponding to this z-value.

If we do not know the mean of the normal distribution, we can choose any number we like, but this choice must be included as part of the hypothesis. With the hamsters, if we did not know that the mean life span was 300 days, we could have gone through the computations for a mean of 300 days. Our conclusion would be: if hamster life span has a mean of 300 days, then the sample mean occurs at $z = 3.75$ (most normal curve tables only go up to z-values of at most 3.5, so when a higher z-value occurs, one simply states the z-value).

In the example of the hamsters, we cannot measure the life spans of all hamsters, so we cannot compute μ and σ exactly. However, advanced statistical techniques enable the determination of these values from means and standard deviations of samples.

Section 4 - Binomial Distributions

In the story which began this chapter, we learned that Theresa Middlebury was an 80% foul shooter (reasonably good, even by NBA standards!). What is the probability that she would make precisely 1 out of 2 free throws?

When we analyze this problem, we assume that each free throw is independent -- no matter whether she hits or misses the first, her probability of sinking the second will still be .8 . When two events are independent, we recall that the probability of *both* events occurring is obtained by multiplying the probability that *each* event will occur. The probability that she will make the first shot and miss the second is .8 x .2 = .16 . Similarly, the probability that she will miss the first shot and make the second is .2 x .8 = .16 . So the probability that she will make precisely 1 out of 2 is 2 x .16 = .32 .

Now let's consider the probability of Theresa's making exactly 80 out of 100 free throws. One way she could do this is to make the first 80 and miss the last 20. Since these free throws are all independent, the probability is obtained by multiplying .8 x ... x .8 x .2 x ... x .2 = $.8^{80}$ x $.2^{20}$. Another way that she could make exactly 80 out of 100 free throws would be to miss the first 20 and then make the remaining 80. The probability of this happening is .2 x ... x .2 x .8 x ... x .8 = $.2^{20}$ x $.8^{80}$ = $.8^{80}$ x $.2^{20}$. A third way this could happen would be for her to miss the first 10, make 80 in a row, and then miss the last 10. The probability of this happening is .2 x ... x .2 x .8 x ... x .8 x .2 x ... x .2 = $.2^{10}$ x $.8^{80}$ x $.2^{10}$ = $.8^{80}$ x $.2^{20}$. No matter in which specific order we arrange for Theresa to make 80 and miss 20, the probability that she will shoot the free throws in that exact order is $.8^{80}$ x $.2^{20}$.

In order to compute the probability that Theresa will make precisely 80 out of 100 free throws, we must therefore add $.8^{80}$ x $.2^{20}$ once for each of the different orders in which she could make 80 and miss 20. How many different such orders are there? If we imagine that we have 100 numbered balls in a jar, and that we choose 80 of them, we could simply decree that Theresa makes the free throws whose numbers are on the balls that have been chosen, and misses the others. Therefore, if we choose the 80 balls numbered 11 through 90, Theresa would miss the first 10, make shots 11 through 90, and miss the last 10.

The number of ways of choosing 80 numbered balls from 100 is C(100,80) = 100!/(80! x 20!). Therefore, Theresa's probability of making exactly 80 of 100 free throws is $C(100,80) \times .8^{80} \times .2^{20}$, which is approximately .0993 .

There are really only three essential numbers in the above formula, as the other numbers can be computed from knowing these three. The first number is 100, which represents the number of free throws Theresa shoots. The next number is 80, which represents the number of free throws Theresa makes. The third number is .8, which represents the probability of Theresa making a free throw. The number .2, which is the probability of Theresa missing a free throw, is 1 - .8, and the number 20, which is the number of free throws Theresa misses, is 100 - 80.

Now let's generalize this a little. Suppose that we send a random free-throw shooter to the line N times. We assume that her probability of making each free throw is p, and that each shot is independent of the others. The probability that she will make exactly k free throws (and miss N-k) is given by the formula

$$C(N,k) \times p^k \times (1-p)^{N-k}$$

This formula, which is known as the **binomial distribution formula,** can be used in a much wider context. The binomial distribution is the theoretical probability distribution in which

$$P(X = k) = C(N,k) \times p^k \times (1-p)^{N-k}$$

Binomial Distributions (a.k.a. Bernoulli Trials)

Suppose that we conduct an experiment which has only two outcomes, which we shall call success, which occurs with probability p, and failure, which occurs with probability 1 - p. Suppose further that each trial of this experiment is independent of the other trials. The probability of getting *exactly* k successes in N trials is given by

$$P(X = k) = C(N,k) \times p^k \times (1-p)^{N-k}$$

Example 1 - A single die is rolled 3 times.

a) What is the probability of rolling exactly 2 sixes?

b) What is the probability of rolling at least 2 sixes?

Solution: a) This is clearly an example of Bernoulli Trials, because 'a die has no memory' (much like a roulette wheel). Therefore, the success probability p = 1/6, the number of trials N = 3, the number of success k = 2, and the formula gives

$$C(3,2) \times (1/6)^2 \times (5/6)^1 = 3 \times 1/36 \times 5/6 = 5/72$$

b) The event 'at least 2 sixes' consists of the two outcomes 'exactly 2 sixes' and 'exactly 3 sixes'. The probability of exactly 3 sixes is

$$C(3,3) \times (1/6)^3 \times (5/6)^0 = 1/216$$

So the probability of at least 2 sixes is 5/72 + 1/216 = 2/27 . ■

Recall that earlier we obtained the probability of Theresa making exactly 80 of 100 free throws as $C(100,80) \times .8^{80} \times .2^{20} = .0993$. There is a tremendous amount of calculation involved in obtaining the number .0993 (WARNING!!! - even with a pocket calculator this is time-consuming. Additionally, if the numbers are not computed in the correct order, it will take you beyond the limit of your calculator; 100! is far beyond the limit of a typical pocket calculator).

Example 2 - Assume that Theresa is an 80% foul shooter. What is the probability that she will make at least 80 out of 100 free throws?

Solution: Using the same reasoning as in Example 1, it is easy to see that this probability is the sum of the probabilities that she will make exactly 80 out of 100, and the probability that she will make exactly 81 out of 100, ... , and the probability that she will make exactly 100 out of 100 (the NBA record for consecutive free throws, as of 2015, is 97). This number is

$$C(100,80) \times .8^{80} \times .2^{20} + \ldots + C(100,100) \times .8^{100} \times .2^{0}$$

Computing the value of this number is such a strain that we will temporarily abandon the project until we find a shortcut. ▉

The Normal Approximation to a Binomial Distribution

Mathematicians are no different from anyone else; they like to do as little grunt work as possible. It was discovered over a century ago that the normal curve provided a very good approximation to the binomial distribution as long as both Np and N(1-p) are both greater than 5. The normal curve which best matches the binomial distribution has mean $\mu = Np$ and standard deviation $\sigma = \sqrt{[N \times p \times (1-p)]}$.

Recall that this is exactly what Pete did in the story. Freddy reported back that Theresa had made 66 out of 100 free throws. Freddy's reaction was, "Oh, well, Theresa's just having an off day." Pete, however, approximated the binomial distribution with a normal distribution whose mean $\mu = 100 \times .8 = 80$, and whose standard deviation $\sigma = \sqrt{(100 \times .8 \times .2)} = \sqrt{16} = 4$. This was valid, because $Np = 100 \times .8 = 80 > 5$, and $N(1-p) = 100 \times .2 = 20 > 5$. Since 66 was 14 below the mean of 80, and 14/4 = 3.5 standard deviations, Pete realized that the hypothesis 'Everything's fine with Theresa' had to be rejected even at the 1% significance level.

Example 3 - Use the normal approximation to the binomial distribution to find the probability that Theresa sinks exactly 80 out of 100 free throws. What is the probability that she will sink at least 80 free throws?

Solution: Because the normal curve involves a continuous random variable and the binomial distribution involves a discrete random variable, the phrase 'exactly 80' for the discrete variable is replaced by 'between 79.5 and 80.5' for a continuous variable. As we have seen, the approximating normal distribution has $\mu = 80$ and $\sigma = 4$. Since $79.5 = 80 - .5 = \mu - 0.125\,\sigma$ and $80.5 = \mu + 0.125\,\sigma$, we have to find the area under the normal curve between $\mu - 0.125\,\sigma$ and

$\mu + 0.125\ \sigma$. This is .0995, which is very close to the exact value of .0993 .

The probability that she will sink at least 80 free throws is approximated on the normal curve by the probability that she will sink at least 79.5 free throws. The area under the normal curve above $\mu - 0.125\ \sigma$ is .5497. ■

Hypothesis Testing - Part 2

We have already touched on an example of hypothesis testing, when Pete decided that something was wrong with Theresa. The usual procedure in hypothesis testing is to 'knock down a straw man'. Pete set up the hypothesis 'Everything is fine with Theresa', and then proceeded to knock it down by showing that it was not true at the 1% significance level. The hypothesis that plays the part of the straw man to be knocked down is called the **null hypothesis**. We say that the **null hypothesis can be rejected at such-and-such significance level**. Notice that this procedure does *not* establish the truth of what we wish to show (in the case of the story, that something was wrong with Theresa), but only establishes that the opposite of what we wish to show is highly unlikely.

Example 4 - A manufacturer of microprocessors claims that the defect rate of his product is 1%. In a shipment of 10,000 microprocessors, 120 are found to be defective. Can you reject the manufacturer's claim at the 5% significance level?

Solution: The null hypothesis would be 'Of 10,000 microprocessors, the defective microprocessors constitute a binomial distribution with N = 10,000 and p = .01'. We approximate the binomial distribution by a normal distribution with mean $\mu = 10{,}000 \times .01 = 100$ and standard deviation $\sigma = \sqrt{\ (\ 10{,}000 \times .01 \times .99\ )\ } = \sqrt{99} = 9.95$. The probability of finding at least 120 defective microprocessors would be the area under the normal curve to the right of 119.5, which is 19.5 / 9.95 = 1.96 standard deviations above the mean. We know that claims are rejected at the 5% significance level if they exceed $1.65\ \sigma$, so we must reject this one. ■

Why Is So Much Faith Placed in Polls?

There are times when it seems as if the political and economic infrastructure of the country, to say nothing of the entertainment we see, is run by polls. Why is it that a poll of 300 voters, or the TV preferences of 1200 Nielsen families, is sufficient to determine actions that affect so many of us? The following example illustrates some of the thinking that goes into statistically-based decision-making.

Example 5 - Suppose that an election between Smith and Jones is viewed as a tossup. What is the probability that a poll of 300 voters will find at least 55% favoring Smith?

Solution: If the success probability p = .5, then we approximate the binomial distribution with a normal distribution in which $\mu = 150$ and $\sigma = \sqrt{(300 \times .5 \times .5)} = \sqrt{75} = 8.66$. The probability that at least 55% of the voters, or 165 voters, will favor Smith is that portion of the normal curve lying above 164.5, which is 14.5/8.66 = 1.67 standard deviations. According to the table, the probability of this is .0475 . ■

In Example 5 above, a pollster working for Jones would tell him (or her) that the situation was desperate. 300 people may seem like a very small number when compared with an electorate that could conceivably be millions of voters, but one of the reasons statistics is such an important discipline is that relatively small samples are quite likely to accurately reflect the entire population. In this example, if you were to assume that a voter was just as likely to vote for one candidate as the other, the probability of finding that at least 55% of the voters favor Smith is less than 5%. There are a number of high-profile cases in which the polls got it wrong (pollsters working for the Republican campaign in 2012 were convinced Romney was going to win), but many more in which they get it right, which is why they continue to be used.

Chapter 12 - Game Theory

Introduction

In the story, Freddy found himself faced with a choice of two possible actions, or **strategies**: he could either call Lisa, or he could decide not to call her. How his choice worked out did not depend solely on what strategy he chose, but also on how Lisa felt. Lisa either wanted Freddy to call or she didn't, but it is probably incorrect to say that she is an opponent bent upon making life as miserable for Freddy as possible. When the opposing actions are not the result of a conscious choice, they are usually referred to as **states**, rather than strategies. The **payoff** Freddy received (which is measured in what might be called relative happiness points, with a high of 10 and a low of 0) depended on the strategy he chose and the state Lisa happened to be in. Since each of the two principals has two possible choices, this is called a **2 x 2 game**. If Freddy expanded his list of strategies to include a surprise visit to New York, the result would be termed a **3 x 2 game**. It is customary to use the first of the two numbers to denote the number of strategies available to the player from whose standpoint we measure the payoffs.

Game theory was started in the 1920s by the mathematician Emile Borel, although the first significant work was a book published in 1944 called *The Theory of Games and Economic Behavior,* by John von Neumann and Oskar Morgenstern. Because many conflict situations arising in war can be conveniently formulated in the framework of game theory, it is not surprising that it was intensively investigated during World War II. However, game theory has found a large number of applications to areas probably not considered by its founders. As we have mentioned, this is characteristic of mathematics -- applications pop up in surprising places.

Section 1 - 2 x 2 Games

Saddlepoints

Let’s take another look at the diagram that Pete made for Dolores.

		High Price (thousands)	Reserve Price (thousands)
Auctioned at	Classic	100	20
	Vintage	70	30

If Dolores auctions the card at Classic, the worst that can happen is that she receives 20 points (recall that each point is equal to $1,000). This number is called a **row minimum** (it is the minimum of all the numbers in the first row). Similarly, if she auctions the card at Vintage, the worst that can happen is that she receives 30 points. The 'best of the worsts', the larger of the two row minimums, is 30 points. This number is called the **maxmin**, which is an abbreviation for the maximum of the row minimums. The maxmin of 30 points occurs when she auctions the card at Vintage.

We turn now to the problem of whether the card will be purchased for the reserve price, or for something higher. If it is auctioned at the reserve price, the worst that could happen is if the card was auctioned at Vintage, and it would cost the buyer 30 points. This number is called a **column maximum** (it is the maximum of all the numbers in the first column). If the card is purchased for a high price, column maximum is 100 points. From the standpoint of potential buyers, who

do not know whether the card will go for a reserve price or a high price, the 'best of the worsts', the smaller of the two column maximums, is 30 points. This number is called the **minmax**, which is an abbreviation for the minimum of the column maximums. The minmax of 30 points occurs when the card goes for a reserve price.

Originally, game theory was devised under the assumption that two intelligent rivals each could choose a particular strategy. The contestants were simply described as Red and Blue, and the strategies as Red 1 and Red 2, Blue 1 and Blue 2. Each side could view the diagram and decide which strategy to choose, much as the two contestants in rock-paper-scissors are free to make their choice. Dolores is in such a position; she is free to sell her card at either auction house.

This is a situation in which the opponent Dolores faces is not a flesh-and-blood rival, but a situation. She has no way of knowing whether the card will sell for a high price or a reserve one. However, if she looks at the game as if the world were choosing the best way to thwart her desires, this is the way she would analyze it. The world knows that that Dolores' maxmin calls for her to auction the card at Vintage, and she knows that its minmax occurs if the card goes for a reserve price. Even if she tells them she will auction the card at Vintage, she cannot be prevented from making at least $30,000. Even if the world lets us know that the card is going for the reserve price, she cannot make more than $30,000. This number, $30,000, is the **value** of the game.

You may be a little worried that the real world will not conspire in such a fashion as to do its best to ensure that the card will sell at the reserve price. That's certainly true. Nonetheless, there's no way she can tell – the card may not interest any buyer. Markets for collectibles are frequently misjudged. Dolores has to choose her strategy based on the assumption that she has a rational opponent (that was one of the original assumptions of game theory) – and she will accrue added profit if her opponent deviates from the best strategy. No one ever does well in a

game by depending on one's opponent to make a mistake. You can hope your opponent makes a mistake, but it's better to plan on the assumption that they won't.

Notice that if each side plays its best strategy (Dolores auctioning the card at Vintage for the reserve price), if either side deviates while the other side plays its best strategy, the side that deviates loses. If Dolores sells at Classic for the reserve price, she loses by doing so. If the card commands a high price at Vintage, Dolores gains (and the world loses) as a result. This is characteristic of games in which the maxmin is equal to the minmax; the side that deviates from its best strategy loses.

When the maxmin and the minmax have the same value, the game is said to have a **saddlepoint** (the rider of a horse sits at a place that is simultaneously a maximum and a minimum -- the lowest point from the back of the horse to the front, and the highest point from the left side of the horse to the right). In this case, each player will elect to follow the single strategy (known as the **pure strategy)** which guarantees that the maxmin and minmax will both be achieved, and the value of the game will always be the maxmin (or the minmax, depending on which of the two opponents you are).

Mixed Strategies

Now let's look at the situation that Freddy encountered. The diagram is repeated below for convenience, but the analysis will be a little clearer if we just look at the classic situation in which there are two intelligent opponents, Red and Blue, each of whom has a choice of two strategies. Remember that high scores are good for Red, low scores are good for Blue.

		Blue	
		1	2
Red	1	10	0
	2	2	7

The first row minimum is 0, the second is 2, so the maxmin is 2. The first column maximum is 10, the second is 7, so the minmax is 7. Let's suppose that Red elects Strategy 2 in order to choose his maxmin strategy and Blue also elects Strategy 2, choosing his minmax strategy. The payoff is 7 points – but Blue can improve his score by switching from Strategy 2 to Strategy 1 if Red sticks with Strategy 2. In fact, no matter which of the four possible choices are made (Red 1 vs. Blue 2, etc.), one side will *always* benefit from deviating if the other side continues to play the same strategy. This is the type of thing that's seen in repeated plays of rock-paper-scissors; one side will always benefit from switching its strategy if the other side continues to play the same strategy.

This type of situation was analyzed in the story when Freddy was trying to decide whether or not to call Lisa. If there is no saddlepoint, each side's best long-term procedure is to adopt a strategy whose expectation is the same against either of its opponent's options. Freddy played the role of Red in the story, so let's suppose Red elects to play Strategy 1 randomly with probability p, and Strategy 2 randomly with probability 1-p. We can compute Red's expectation against either of Blue's strategies.

Against Blue 1, Red wins 10 points with probability p and 2 points with probability of 1-p, for an expectation of $10p + 2(1-p) = 8p+2$.

Against Blue 2, Red wins 0 points with probability p and 7 points with probability of 1-p, for an expectation of $0p + 7(1-p) = 7-7p$.

If we equate these two expectations, nothing Blue can do will prevent Red in the long run from achieving the equated expectation. Solving $8p+2 = 7-7p$, we see that $p = 1/3$, and that the expectation for Red against either Blue 1 or Blue 2 is 4 2/3 points, just as in the story.

Analyzing 2 x 2 Games

Step 1 - Find the minmax and the maxmin. If these two are equal, the game has a saddlepoint, and each side should play a pure strategy. The row player should play the strategy with the higher minimum, and the column player should play the strategy with the lower maximum. The value of the game is the maxmin (or the minmax, as they are the same).

Step 2 - If the game does not have a saddlepoint, assume that the row player plays one row with probability p and the other with probability 1-p. Compute the expectation of this strategy against each of the column player's two strategies. Equate these two expectations to determine the value of p. The value of the game can be determined by computing the expectation against either strategy using the value of p just determined.

The column player should go through a similar analysis.

Section 2 - Non-Zero-Sum Games; Prisoner's Dilemma and Chicken

The cops have taken Doc and Fingers in for questioning just after they stashed the take from the Metropolitan bank job. Doc has no priors (prior convictions), but Fingers had done a previous stretch in the slammer, and so Fingers will end up with a stiffer sentence if convicted. Doc and Fingers are being interrogated separately.

"Listen," the lieutenant tells Doc, "we know you and Fingers knocked over the bank. However, the guy we really want is Fingers. Turn state's evidence, and we'll let you off. It'll be

a lot tougher if you don't." He leaves the room, and Doc takes out a pencil and paper, not to write a confession or turn state's evidence, but to draw the following diagram.

		Fingers	
		Squeals	Clams Up
Doc	Squeals	D3 F6	D0 F8
	Clams Up	D5 F0	D1 F3

Yes, it's a 2 x 2 game matrix, but it's different from the ones we examined in the previous section. Each box has two payoffs; the one starting with D is the number of years of imprisonment Doc estimates he will receive, and the one starting with F is the number of years of imprisonment that Doc estimates Fingers will receive. Thus, in the situation where Doc squeals on Fingers but Fingers clams up, Doc estimates that he will get off without any jail time (0 years), but that Fingers will be sentenced to 8 years. Note that Doc is no fool; the payoffs in the case where Doc clams up and Fingers squeals indicates that Doc realizes another cop is offering Fingers the same sort of deal the lieutenant offered Doc.

In the games we studied in the previous section, whatever was won by one player was lost by the other. If we regard wins as positive and losses as negative, then if one player wins $5 the other will win -$5. The sum, $5 + -$5, is 0. Games of this type are called **zero-sum games**.

The situation with Doc and Fingers is substantially different. No longer is it true that what Doc wins, Fingers loses (and vice-versa), so this is a **non-zero-sum game**. As we shall see, the analysis of non-zero-sum games is substantially more complicated than the analysis of zero-sum games.

Let's go back and look at the situation that confronts Doc. First, Doc must decide what type of game he is playing. There are three basic types.

In the first type of game, Doc is only interested in minimizing the amount of time that he spends in jail. Games of this type are called **individual** or, for obvious reasons, **selfish**.

Another possibility is that Doc and Fingers together are interested in minimizing the total amount of time that the two of them spend in jail. Games of this type are called **co-operative**.

Finally, it is possible that Doc wants to serve a shorter time than Fingers, possibly so that he can dig up the Metropolitan haul and head for Rio while Fingers is still in jail. Games of this type are called **competitive**.

Notice that the optimum payoff for Doc in the individual game occurs when he squeals and Fingers clams up, for then Doc serves no jail time. The optimum payoff in the co-operative game occurs when both Doc and Fingers clam up, for then they spend a combined four years in jail. Finally, the optimum payoff in the competitive game also occurs when Doc squeals and Fingers clams up, for then Fingers is in jail for eight more years than Doc. Payoffs in competitive games are measured in terms of the differences in payoffs to the two players.

The situation involving Doc and Fingers is called the **Prisoner's Dilemma**. Suppose that each of the two is playing a selfish game. From examining the payoff matrix, we observe that, if each player knows what the other will do, that player is better off by squealing. For instance, if Doc knows that Fingers will squeal, Doc can reduce his 5-year sentence to 3 years by squealing. Alternatively, if Fingers knows that Doc will clam up, Fingers can reduce his 3-year sentence to 0 years by squealing. You should examine the other two cases, where Doc knows that Fingers will clam up, or that Fingers knows Doc will squeal, to see that each player will still benefit by squealing.

Thus, each player is faced with an overwhelming temptation to squeal. Yet, if they both squeal, each receives his next-to-worst result. Therefore, if they have discussed this situation in advance, or if each has faith in the other, they might agree that they should both clam up, so as to avoid near-disasters to each of them. They would certainly trust each other to clam up if both are playing a co-operative game, for this would guarantee the best result.

But can Doc trust Fingers (and vice-versa)? Paradoxically, if Doc trusts Fingers to keep his word and clam up, then Doc should squeal (at least in the individual or competitive game)!

Notice that, when both players elect to squeal, each is punished if he deviates from that strategy if the other maintains it. Thus, Doc's term would increase from 3 years to 5 years if he clammed up while Fingers squealed, and Fingers' term would increase from 6 to 8 years under the same circumstances. This is somewhat analogous to a saddlepoint in a zero-sum game, and is called an **equilibrium point**.

Notice also that the type of game being played is very sensitive to the actual numbers of the payoffs. If Fingers has no prior convictions, so that he can be expected to receive the same sentence as Doc, the payoff matrix would be

		Fingers	
		Squeals	Clams Up
Doc	Squeals	D3 F3	D0 F5
	Clams Up	D5 F0	D1 F1

Now each player has substantially more incentive to clam up, since both benefit equally by doing so. As a result, Doc is far more likely to trust Fingers in this situation than in the other.

This analysis serves to point out that, unlike the zero-sum game, there is a definite psychological component to the Prisoner's Dilemma. As a result, it has attracted attention not only from mathematicians, but from psychologists.

Repeated Plays

As we observed in the first section, playing the proper mixed strategy in a zero-sum 2 x 2 game without a saddlepoint will insure the greatest long-term expectation. The proper strategy in individual Prisoner's Dilemma is not as amenable to mathematical analysis as is the zero-sum game, so to study the various possible strategies a 'computerized tournament' was conducted some years ago in which each entrant concocted a strategy, and the various strategies went 'head-to-head' against each other. Each entrant played the role of Doc in a fixed number of repetitions of the game with the same payoff matrix, and then played the role of Fingers.

When playing the role of Doc, each entrant's strategy consisted of a plan determining Doc's action in the first case, and then Doc's actions in later trials was based on what Fingers did. For instance, one possible strategy is STOOL PIGEON; so-called because a stool pigeon always squeals. Another possible strategy is 'Doc starts by clamming up, and then only squeals if Fingers has squealed more than he clammed up'. Thus, if in the first 10 games Fingers had squealed 4 times and clammed up 6 times, then Doc would clam up in the 11th game.

This contest attracted considerable attention from both game theorists and psychologists. The winning strategy, devised by Anatol Rapaport of the University of Michigan, was the deceptively simple strategy described as 'TIT FOR TAT', in which Doc clammed up the first time and then always did what Fingers did the time before. Thus, if Fingers squealed on the seventh time the game was played, Doc squealed on the eighth time the game was played.

A little thought will enable one to appreciate the value of TIT FOR TAT, assuming that the other player is rational and is playing a selfish game. STOOL PIGEON, for example, always squeals. As a result, STOOL PIGEON's opponent will quickly discover that any attempt to co-operate by clamming up is futile, and will therefore be forced to squeal in self-defense. As a result, when STOOL PIGEON is one of the participants, the game will quickly gravitate to a situation in which both players always squeal.

Against TIT FOR TAT, however, clamming up is immediately rewarded. A rational player will eventually realize this, and the game will enter a period where both players clam up. At some stage, TIT FOR TAT's opponent will decide that he can gain an immediate advantage by squealing, believing that TIT FOR TAT is in clam-up mode. TIT FOR TAT immediately retaliates, and continues to squeal until TIT FOR TAT's opponent 'atones' for squealing by clamming up. This redresses the immediate advantage the opponent gained from the first squeal, and the game re-enters a phase where both players clam up.

It has not been mathematically proven that TIT FOR TAT is the optimal strategy, but it habitually wins in such tournaments. Fame and fortune (well, fame, anyway) await you if you can devise a strategy that does better.

Prisoner's Dilemma in the Real World

Prisoner's Dilemma has attracted a great deal of analysis because it can be shown to model many situations that exist in the real world. Two of them are presented below.

Example 1 (It's a MAD, MAD World) - During the Cold War, the prevailing military philosophy between the two superpowers was that of Mutually Assured Destruction (MAD). The idea was that each side should maintain a nuclear arsenal invulnerable to destruction from a first strike from the other side, so that any attempt to start a war would result in destruction for both sides. However, the proliferation of nuclear weapons made it substantially more likely that a war might

start accidentally or through malign intent on the part of a misguided subordinate (see the movie, *Dr. Strangelove*, for an example). As a result, both sides could see the value of simultaneously disarming, but neither could risk unilateral disarmament. This led to the following Prisoner's Dilemma matrix, in which 4 represents the most favorable result, 1 the least favorable result.

		Russia	
		Disarms	Maintains Stockpile
United States	Disarms	US3 R3	US1 R4
	Maintains Stockpile	US4 R1	US2 R2

Using the numbers 1 to 4 to represent the spectrum of the worst alternative up to the best alternative retains the essence of the Prisoner's Dilemma situation without the problem of quantifying such situations as mutual annihilation. This method of describing games is called **ordinal evaluation**.

Example 2 (It's an Ad, Ad World) - TerrifiCola has an annual profit of $100 million, and MagiCola has an annual profit of $85 million. Agency bigwigs are trying to persuade each company's management to launch a $20 million advertising campaign featuring a rock superstar. Assuming that one company does and the other doesn't, $40 million in cola profits can be swung to the company's ledger (less the campaign costs). On the other hand, if both companies enter into a battle of the rock superstars, the only ones to gain will be the superstars and the ad execs. This leads to the following payoff matrix.

		MagiCola	
	` Campaign		No Campaign
TerrifiCola	Campaign	T80 M65	T120 M45
	No Campaign	T60 M105	T100 M85

If the two companies could be persuaded to co-operate, they would undoubtedly choose not to run advertising campaigns. But can they trust each other?

Notice also that the cost of the campaign relative to the profit margin of the company can make a difference. If an expensive ad campaign would force MagiCola to lose money, TerrifiCola might adopt such a policy even though its profits might be hurt as a result. Such a situation can also exist in other Prisoner's Dilemma scenarios. One school of thought holds that, even though the arms race cost the United States a fortune, the expense of trying to keep up eventually led to the dissolution of the Soviet Union.

The Game of Chicken

In the classic film, *Rebel Without a Cause,* James Dean plays a gang member who is trying to acquire status by engaging another gang member in a game of Chicken. Both men get in cars, and drive towards a cliff, simultaneously keeping an eye on the other. The object of the game is to bail out (by diving out of the car) after your opponent has 'chickened out' by bailing out first.

There are three alternatives for each player: bails out first, bails out last, or fails to bail out in time (and thus goes over the cliff). This would lead to a 3 x 3 game, and so a simplified game of Chicken has been constructed which adheres to the original spirit, yet only has two strategies

available for each driver. They drive towards one another, and each has the option of swerving to avoid a head-on crash.

It is difficult to accurately quantify the payoffs in the game of Chicken, so ordinal evaluation can be used to describe the game instead.

		Opponent	
		Swerves	Doesn't Swerve
You	Swerve	Y3 O3	Y2 O4
	Don't Swerve	Y4 O2	Y1 O1

The worst result is clearly when neither you nor your opponent swerve; the result could be fatal. If one swerves and the other doesn't, both at least survive, but the one that swerves loses the respect of the other gang members. If both swerve, they obviously survive and neither loses any respect relative to the other.

Like Prisoner's Dilemma, the game has an obvious co-operative solution; both of the contestants will agree to swerve. The competitive and selfish versions of the game are not so clear, and the actual result can depend upon how the players quantify their alternatives.

Chicken, too, occurs in real-life situations.

Example 3 (The Cuban Missile Crisis) - In October, 1962, the United States announced that it had discovered missile bases in Cuba. Russian ships were headed towards Cuba to re-supply them, and an American naval blockade was set up. The payoff matrix for the game was

Russian Ships

	Turn Back	Run Blockade
Allow Passage	A3 R3	A2 R4
American Ships		
Intercept	A4 R2	A1 R1

This is the exact same payoff matrix as the one for the game of Chicken, but played for much higher stakes. This was perhaps the closest the Cold War came to turning into a hot one, but the Russian ships turned back, and the missile sites in Cuba were dismantled. This game of Chicken was described by the Secretary of State, Dean Rusk, as "We're eyeball to eyeball, and I think the other guy just blinked."

Chapter 13 - Elections

Introduction - The Limits of Knowledge

It's been an aphorism for a long time that knowledge is power. There is an unspoken corollary, that if you could secure unlimited knowledge, you would consequently have unlimited power.

Of course, it has been known for thousands of years that there were some things you couldn't know. The Greeks knew, for instance, that you couldn't trisect an angle using just a straight-edge and a compass. However, that wasn't a limitation of *knowledge*; it was an example of how knowledge could save you a lot of time trying to accomplish the impossible.

It has only been during the 20th century that we have actually started to come to grips with the realization that knowledge itself might be intrinsically limited. In the late 1920s, the physicist Werner Heisenberg proved his famous Uncertainty Principle, which stated that if you knew where an electron was, you couldn't know where it was going, and if you knew where it was going, you couldn't know where it was. Less than a decade later, the mathematician Kurt Godel proved that in any axiomatic system of sufficient complexity, there existed **undecidable** results: propositions that could be neither proved nor disproved! (He actually illustrated this with a proposition about arithmetic.)

It has long been a dream of social scientists to come up with an ideal system for translating individual preferences to the preferences of the society. In this chapter, we shall study Arrow's Theorem, which shows that this cannot be accomplished.

Section 1 - Voting Methods and Arrow's Theorem

We have already met with three different voting schemes in the Bankers' Club Election. We will assume that every voter submits a ballot which has a preference listing among all the candidates. For example, if a voter submits a ballot which has Ackroyd as the first choice, Williams as the second, and Morris as the third, then if Ackroyd is not elected, the voter prefers Williams to Morris. Ties are not permitted.

For convenience, we reprint the results of the Bankers' Club election.

1st: Ackroyd 2nd: Morris 3rd: Williams # of ballots - 24

1st: Williams 2nd: Morris 3rd: Ackroyd # of ballots - 18

1st: Morris 2nd: Williams 3rd: Ackroyd # of ballots - 12

Election Scheme 1: *Plurality* - The candidate who receives the most first-place vote wins.

In the story, this was Forrest Ackroyd's preferred voting method. This has the advantage that it is generally easy to compute, and rarely results in ties (especially if there are a large number of voters). The disadvantage of this scheme is that it is possible for the winner to be loved by a minority of the electorate, and vigorously detested by everyone else. It appears that this is how the electorate at the Bankers' Club felt about Ackroyd.

Election Scheme 2: *Runoff* - Eliminate all candidates except those who have the two highest 1st-place totals, and then have a secondary election between them.

This was how Helen Williams felt the election should be decided. One of the difficulties with this method is that it lends itself to **insincere voting**. An example of insincere voting could be seen in a four-person race decided by the runoff method, in which there is a clear front-runner, a close contest for the second slot, and a splinter candidate. The splinter candidate could wield a lot of clout by telling his supporters to throw their votes to one of the two candidates who are in contention for the second position on the post-runoff ballot. This frequently happens in the real

world. **Sincere voting** occurs when each individual lists his preferences, letting the chips fall where they may.

Election Scheme 3: *Borda Count* - An arithmetical weighting scheme is devised for 1st place, 2nd place, 3rd place, ... , with more points given for higher placings (1st place gets more points than 2nd, etc.). The winner has the highest total (or highest average per voter).

In its favor, Borda counts reflect how each voter feels about each candidate. One disadvantage of Borda counts is that different Borda counting schemes can result in different winners!

Example 1 - Suppose there are three candidates in an election, whom we shall call A, B, and C, and the balloting produces the following results:

1st - A 2nd - B 3rd - C # of ballots = 11

1st - B 2nd - C 3rd - A # of ballots = 8

1st - C 2nd - B 3rd - A # of ballots = 17

Compute the results of the election by the Borda count method (a) if the weighting scheme is 3-2-1, and (b) if the weighting scheme is 5-3-2.

Solution: This is basically simple arithmetic. With a 3-2-1 weighting scheme, the point totals are

$$A = 11 \times 3 + 8 \times 1 + 17 \times 1 = 58$$

$$B = 11 \times 2 + 8 \times 3 + 17 \times 2 = 80$$

$$C = 11 \times 1 + 8 \times 2 + 17 \times 3 = 78$$

So B is the winner. However, with a 5-3-2 weighting scheme, the point totals are

A = 11 x 5 + 8 x 2 + 17 x 2 = 105

B = 11 x 3 + 8 x 5 + 17 x 3 = 124

C = 11 x 2 + 8 x 3 + 17 x 5 = 131

In this case, C is the winner. ■

Let's take another look at the election in Example 1.

1st - A 2nd - B 3rd - C # of ballots = 11

1st - B 2nd - C 3rd - A # of ballots = 8

1st - C 2nd - B 3rd - A # of ballots = 17

Suppose we looked at how the three one-on-one races (A vs. B, A vs. C, and B vs. C) would fare.

A vs. B: A is preferred to B on 11 ballots, but B is preferred to A on 25 ballots. *B wins*.

A vs. C: A is preferred to C on 11 ballots, but C is preferred to A on 25 ballots. *C wins*.

B vs. C: B is preferred to C on 19 ballots, but C is preferred to B on 17 ballots. *B wins*.

B claims victory in the three-person race because B bests all the other candidates when they go one-on-one. This gives us yet another voting method.

Election Scheme 4: *One-on-one* (a.k.a. Condorcet voting): Each possible pair of candidates is matched. If one candidate wins against all other candidates, that candidate is declared the winner.

One-on-one voting runs the risk that there may not be a winner.

Example 2 - Suppose that the outcome of an election with three candidates A, B, and C is

1st - A 2nd - B 3rd - C # of ballots = 11

1st - B 2nd - C 3rd - A # of ballots = 10

1st - C 2nd - A 3rd - B # of ballots = 9

What is the result of the one-on-one voting method?

Solution: A vs. B: 20 prefer A to B, 10 prefer B to A. A wins.

A vs. C: 11 prefer A to C, 19 prefer C to A. C wins.

B vs. C: 21 prefer B to C, 9 prefer C to B. B wins.

Since each candidate wins one of the three one-on-one matches, this voting method fails to produce a winner. ▉

The Paradox of Transitivity

It has been known for at least a century that difficulties can arise in certain situations. One obvious property that individual preferences display is that, if the individual prefers alternative A to alternative B, and that same individual also prefers alternative B to alternative C, then that individual must prefer alternative A to alternative C. This is called **transitivity**. Numbers displays a similar type of transitivity: if $a > b$ and $b > c$, then it follows that $a > c$.

However, transitivity of individual preferences does <u>not</u> produce transitivity of the preferences of a majority of society! To see an example of this, let's look at Example 2 again.

Example 3 - Once again, suppose that A, B, and C are candidates in an election, and that the voters cast their ballots as follows:

1st - A 2nd - B 3rd - C # of ballots = 11

1st - B 2nd - C 3rd - A # of ballots = 10

1st - C 2nd - A 3rd - B # of ballots = 9

Notice that A is preferred to B by a majority of voters (20 to 10), and similarly B is preferred to C by a majority of voters (21 to 9). If an individual exhibited these preferences (A over B and B over C), we would be justified in concluding that he or she preferred A to C. However, when a majority prefers A to B and B to C, we cannot reach the same conclusion, for in this example, a majority prefers C to A (by 19 to 11).

This example simply illustrates that there is something wrong with our intuitive ideas about how the preferences of the majority can be deduced from the preferences of individuals.

Arrow's Impossibility Theorem

Example 3, that majority preference is not transitive, is a little unsettling. Nonetheless, it did not dissuade generations of social scientists from seeking a system which would translate the preferences of individuals into preferences for the society.

Ideally, we would like to take a list of individual preferences, and from this arrive at a list of the preferences of society. Example 3 shows that we cannot expect transitivity to hold for society's preferences, even though it will certainly hold for the preferences of the individual. We would like to construct a 'social preference method', which is derived from a list of individual preferences, which enables society to choose between any two alternatives. Here is a list of some properties, each of which is desirable.

Transitivity: We would like this 'social preference method' to be transitive: if society prefers alternative A to alternative B, and it also prefers alternative B to alternative C, then it should also prefer alternative A to alternative C.

Non-dictatoriality: We would like a society that is non-dictatorial. In other words, there should be no individual whose preferences are automatically adopted by society. In a dictatorship, the dictator's preferences are automatically adopted by society; that's what makes a dictator.

Preservation of Unanimous Preferences: If every member of the society prefers alternative A to alternative B, then the society should prefer alternative A to alternative B.

Independence of irrelevant alternatives: Suppose that the ballot contains at least three alternatives, A, B, and C, and that society prefers alternative A to alternative B. Now suppose that alternative C is eliminated from the ballot. Society should still prefer alternative A to alternative B.

Like transitivity, this is generally obvious for individuals. Let's suppose you are going out for dinner, and steak, fish, chicken and hamburger are on the menu. You select steak. The waiter comes back and tells you that they ran out of fish. Your reaction would undoubtedly be, "Who cares? Bring me my steak!" The absence of fish is an irrelevant alternative; it would only be relevant if you had actually decided to have fish for dinner.

Each of these properties is not only desirable, but seems ostensibly very reasonable. However, Arrow's Impossibility Theorem shows that one cannot construct a social preference method with all of the above properties!

Every time an election involves more than two candidates, there is the possibility that the choice of method may play a critical role in deciding the election, and that the 'will of the people' may be inadvertently or unknowingly subverted. Arrow's Theorem shows that there can be no perfect method.

The implications of this for both social philosophy and real-world democracy should not be underestimated. It might be nice to believe in miracles, and think that there is nothing that we

cannot accomplish if we simply give 110%. It is far more realistic to be aware of what the limitations of what we can do and what we can know are, for that way both individuals and society can plan intelligently and set reasonable goals.

Section 2 - Voting Systems: Who's Got the Power?

In a democracy, each person has one vote, and presumably this implies that each person has the same amount of political power. In reality, voters divide up into various different **blocs**, such as liberals, moderates, and conservatives. While it is not true that every liberal is pro-labor or every conservative is pro-business, nonetheless great importance is attached in an election to getting the liberal vote, or the conservative vote, or the environmental vote. So how can we determine how much power each voting group actually exerts?

Weighted Voting Systems

When shares are issued in a corporation, voting is usually done not by the 'one shareholder, one vote' method that would characterize a democracy, but by the 'one share, one vote' method, in which each individual gets as many votes as the number of shares he or she owns. In such a situation, a shareholder who owns a thousand shares obviously wields more clout than a shareholder with just one share.

For the remainder of this section, we will imagine that there are n voters. A **weighted voting system** consists of the number of votes each of the voters has, and the number of votes needed to pass a piece of legislation. We write $W = \{ w_1, w_2, \ldots , w_n; q \}$ to abbreviate the weighted voting system W. w_k is the number of votes that voter k has, and is called the **weight** of voter k, or the k^{th} weight. q, which stands for quota, is the number of votes needed to pass a bill. We shall assume for simplicity that the quota will always be at least half the total number of votes available.

Quotas of 1/2 are usual, but other quotas are not infrequent. In the United States Congress, a simple majority is required (in both houses) to pass most bills, but to overturn a presidential veto requires 2/3 of the votes cast. In order to amend the Constitution, 3/4 of the states must vote in favor of the amendment.

Coalitions and Swing Voters

If W is a weighted voting system consisting of n voters, a **coalition** is a subset S of the voters (we can regard S simply as a subset of the integers). It is a **winning coalition** if the sum of the votes of the members of the coalition is greater than or equal to the quota q, and is a **losing coalition** otherwise. If a particular coalition is a winning coalition, a particular voter is a **swing voter** for that coalition if, when that particular voter drops out of the coalition, the coalition becomes a losing one.

Example 1 - Suppose that Ann controls 40% of the stock of a corporation, Bob controls 25%, Charles 20%, and Donna 15%, and 51% of the stock is required to approve a decision.

(a) Is { Ann, Bob, Charles } a winning or losing coalition? If it is a winning coalition, who are its swing voters?

(b) Is { Bob, Donna } a winning or losing coalition? If it is a winning coalition, who are its swing voters?

Solution: (a) The subset { Ann, Bob, Charles } is a winning coalition, since they own 40% + 25% + 20% = 85% of the stock. Ann is a swing voter for this coalition, since if she drops out, Bob and Charles together own only 45% of the stock. However, if Bob drops out, Ann and Charles still own 60% of the stock, and if Charles drops out, Ann and Bob still own 65% of the stock, so neither Bob nor Charles is a swing voter for this coalition.

(b) The subset { Bob, Donna }, which controls only 40% of the stock, is a losing coalition, and so has no swing voters.

Notice that in Example 1 both the weight of the voter *and* the quota determine whether or not that voter is a swing voter for a particular coalition. If the quota had been 62%, both Ann and Bob would have been swing voters for { Ann, Bob, Charles }, but Charles would not have been. Increase the quota to 67%, and all are swing voters. In the **unanimous** coalition { Ann, Bob, Charles, Donna }, there are no swing voters.

The Banzhaf Power Index

In a weighted voting system consisting of n voters, there are exactly 2^n coalitions (an easy way to see this is to use the Chinese Restaurant Principle; each of the n voters has 2 choices, whether to be in or out of a particular coalition). The **Banzhaf Power Index** B_k of voter k is the number of coalitions for which voter k is a swing voter. The more coalitions that a voter can swing, the more power that voter has, just as it's better to have more job offers than fewer. The name comes from one of the first individuals to study it, the lawyer and consumer advocate John Banzhaf.

Since the weighted voting system in Example 1 has $2^4 = 16$ possible coalitions, let's defer the analysis of this system for a moment, and look at W = { 40, 35, 25; 62 }. We shall compute the Banzhaf Power Indices by constructing the voting system equivalent of a 'truth table', in which the statements are the voters, and they can assume the truth values 'yes' (they belong to the coalition) or 'no'.

Voter	Weight	Quota = 62							
1	40	Y	Y	Y	Y	N	N	N	N
2	35	Y	Y	N	N	Y	Y	N	N
3	25	Y	N	Y	N	Y	N	Y	N
Total		100	75	65	40	60	35	25	0

Wins?		Y	Y	Y	N	N	N	N	N

Each column of Ys and Ns represents a particular coalition. The third column, which reads Y N Y from the top down, contains voters 1 and 3, but not voter 2. Its total weight is 40 + 25 = 65, and it is a winning coalition.

This represents the first step in computing the Banzhaf Power Indices of the voters. The next step is to examine the winning coalitions to see which voters are swing voters for that coalition. In the first coalition (Y Y Y), only voter 1 is a swing voter. In the second coalition (Y Y N), both voters 1 and 2 are swing voters, and in the third coalition (Y N Y), both voters 1 and 3 are swing voters. We now augment the above table by putting a * by the Y each time a voter is a swing voter, and then compute the Banzhaf Power Index (BPI) simply by counting the number of times a starred Y appears in each row.

Voter	Weight	Quota = 62								BPI
1	40	Y*	Y*	Y*	Y	N	N	N	N	3
2	35	Y	Y*	N	N	Y	Y	N	N	1
3	25	Y	N	Y*	N	Y	N	Y	N	1
Total		100	75	65	40	60	35	25	0	
Wins?		Y	Y	Y	N	N	N	N	N	

Even though we have done this in two stages, of course once a coalition is determined to be winning, one can simply check to see which voters are swing voters for that coalition.

Determining the BPI is a straightforward process. However, because there are precisely 2^n coalitions, as the number of voters increases, the length of time required for the calculation

increases exponentially. With 10 voters, there are 1,024 possible coalitions. With 20 voters, there are over a million!

Example 2 - Determine the BPIs of the various voters in the weighted voting system W = { 50, 30, 20; 75 }.

Voter	Weight	Quota = 75								BPI
1	50	Y*	Y*	Y	Y	N	N	N	N	2
2	30	Y*	Y*	N	N	Y	Y	N	N	2
3	20	Y	N	Y	N	Y	N	Y	N	0
Total		100	80	70	50	50	30	20	0	
Wins?		Y	Y	N	N	N	N	N	N	

Notice that in Example 2, voter 3 is never a swing voter, and so the BPI of voter 3 is 0. This is not to say that voter 3 never participates in any winning coalitions, but simply that voter 3 can never determine the outcome of an election. A voter with a BPI is said to be **powerless** (sometimes such a voter is called a **dummy**).

Let's imagine that, in Example 2, the voters are the shareholders of a corporation which has issued 100 shares of stock. Voter 1 has encountered what are euphemistically called 'cash flow problems', and wants to sell some stock to voter 3.

Example 3 - In the weighted voting system W = { 50, 30, 20; 75 } from Example 2, suppose that each weight unit represents a share of stock. What is the largest number of shares of stock voter 1 can sell to voter 3 without changing the BPIs of each voter?

Solution: Suppose that voter 1 sells S shares to voter 3. Now voter 1 will have 50 - S shares, and voter 3 will have 20 + S.

Voter	Weight	Quota = 75							
1	50-S	Y	Y	Y	Y	N	N	N	N
2	30	Y	Y	N	N	Y	Y	N	N
3	20+S	Y	N	Y	N	Y	N	Y	N
Total		100	80-S	70	50-S	50+S	30	20+S	0

In order for voter 1 to have a BPI of 2, we must look at coalitions 1 through 4 (reading the coalition columns left to right), in which voter 1 participates. Coalitions 3 and 4 can never be winning coalitions, because 70 and 50-S are less than the Quota of 75. Coalition 2 is only a winning coalition if $S \leq 5$. If $S = 5$, both voters 1 and 2 are swing voters in coalitions 1 and 2, and these are the only winning coalitions. Voter 3 is still powerless, so the BPIs remain unchanged. ■

Suppose that, in Example 3, voter 1 sells 6 shares to voter 3. Now the BPI structure becomes

Voter	Weight	Quota = 75								BPI
1	44	Y*	Y	Y	Y	N	N	N	N	1
2	30	Y*	Y	N	N	Y	Y	N	N	1
3	26	Y	N	Y	N	Y	N	Y	N	0
Total		100	74	70	44	56	30	26	0	
Wins?		Y	N	N	N	N	N	N	N	

In Examples 2 and 3, voters 1 and 2 have equal power, and voter 3 is powerless. The same can be said for the example above. This raises the question: is a weighted voting system in which the BPI structure is (2, 2, 0) the same as one in which the BPI structure is (1, 1, 0)?

(Notice that we use parentheses (...) to denote the BPI structure, rather than braces { ... }, because the order of the numbers in the BPI structure corresponds to the order of the voters.)

This is a tricky question. In the sense that the two 'empowered' voters have equal power and the third voter is powerless, the answer is yes. On the other hand, if the most powerful voter has a BPI of 1, it is easier for a powerless voter to achieve clout than if the most powerful voter has a BPI of 2, or higher. This is a subject of some complexity.

The Banzhaf Power Index has been used to study various electoral systems. In particular, it has been used to show that there is a slight bias on the part of the electoral college as a whole toward large states, but this bias is more pronounced when one looks at the role played by the individual voter. For instance, a voter from California is more than three times as likely as a voter from the District of Columbia to be the swing voter who elects the President. Admittedly, the probability that this will happen is extremely small, but nonetheless it demonstrates that the electoral college system contains built-in biases.

Other Applications

As is so often the case in mathematics, a tool invented to study one area can have applications in widely disparate areas. Although we have discussed the Banzhaf Power Index in connection with weighted voting systems, the actual computation of the BPI only involves knowing which coalitions are winning, and which ones are losing.

The words 'winning' and 'losing' can be interpreted in a broader context. Suppose we have a set S containing n things, and we have a way of classifying all the subsets of S as 'good' or 'bad'. For each element $x \in S$ and each subset A of S such that $x \notin A$, we compute the BPI of a particular item x by adding 1 to the BPI of x if A is 'bad' and $A \cup \{x\}$ is 'good'.

This is of special interest when one is examining a system which will function properly when certain subsets of its components are functioning properly, such as airplanes or communications

networks. Computing the BPIs of the various components enables one to isolate the components which are most critical to the functioning of the system.

Chapter 14 – Algorithms, Efficiency and Complexity

Introduction - The Value of Planning

We live in an era in which it is necessary to go places and do things in as efficient a manner as possible. When we go places and do things inefficiently, we waste precious resources. Some of these resources, such as time and money, can be expressed in terms of numbers. As we know, anything that can be expressed in numbers is fair game for analysis.

Most of us are familiar with some of the basic principles of operations research, the branch of mathematics concerned with efficient planning, from our everyday life. For instance, if we need to mail a package at the post office, pick up a book at the library, and leave the car to have the oil changed, and all of these locations are within walking distance of one another, most of us will find out when the library and post office are open so we can mail the package and pick up the book while the oil is being changed. Also, if we have several different locations to visit, we try to visit them in an order which will minimize the time we spend getting from one place to another.

In each day, we have only a few things to do, and a few places to go. It is therefore fairly easy to plan to do these things with some efficiency, because there are only a limited number of ways to do them. However, when there are many things to do and many places to go, the number of possible ways to do them can be astronomically large. For example, in any major construction project, there are a huge number of tasks to be done. A lot of time and money can be wasted if the electricians are sitting around waiting for the wiring conduits to be installed.

In this chapter, we shall investigate some of the mathematics of going places (routes). As we have repeatedly emphasized, mathematics is concerned with the real world. The contributions

that mathematics can make toward the solution of these problems are of tremendous benefit to society.

Section 1 - Going Places (Routes)

All cities, whether large or small, face two common problems involving routes: garbage collection, and the repair of broken traffic lights. A garbage truck starts out from a central location, picks up the garbage at all the buildings on a number of streets, and returns to its starting point. The route is most efficiently designed if the truck does not have to retrace any of the streets.

A traffic light repair crew faces a different problem. Broken traffic lights will undoubtedly occur at different areas of the city, and the route for repairing them will be most efficiently designed if the total distance the repair crew has to travel is kept as small as possible.

Each of the routing problems can be presented in the simplified form of a **graph**, as indicated below (yes, the word 'graph' has a different meaning when discussing functions, but mathematicians are not the only ones to assign multiple meanings to the same word -- look up 'spring' in the dictionary!). A graph consists of points, called **vertices**, connected by lines, called **edges**.

The garbage collection problem is to design routes which will retrace as few edges as possible. The traffic light repair problem is to minimize the total distance traveled in visiting all the vertices.

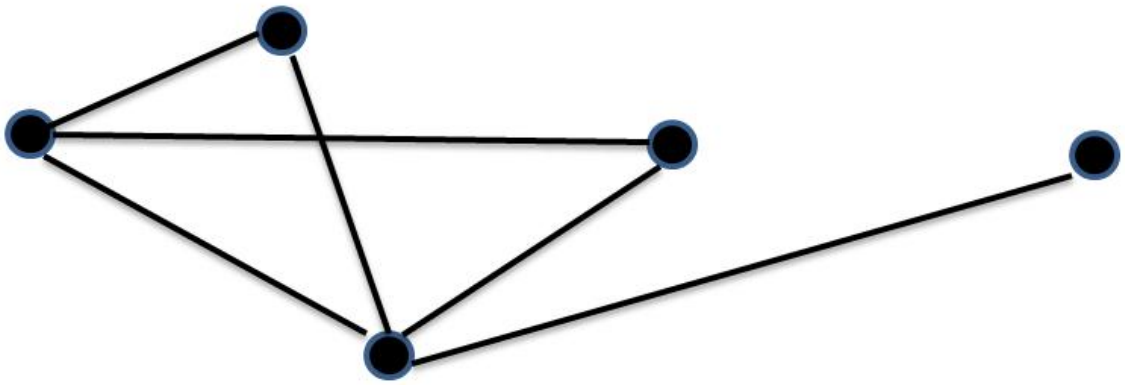


Fig. 13-1 A Graph With Five Vertices and Six Edges

When the Goal is to Retrace as Few Edges as Possible

In the early portion of the 18th century, the citizens of Konigsberg, a small town in Prussia, enjoyed walking across the bridges to two islands in the middle of the River Pregel, which went through the city. The layout of the islands and bridges appears in Fig. 13-2.

The question arose as to whether it was possible to cross all the bridges once without crossing any of the bridges twice. No one was able to find such a route (you might try your hand at it!), but no one was able to show that no such route could be found -- until Leonhard Euler, one of the greatest mathematicians in history, studied the problem.

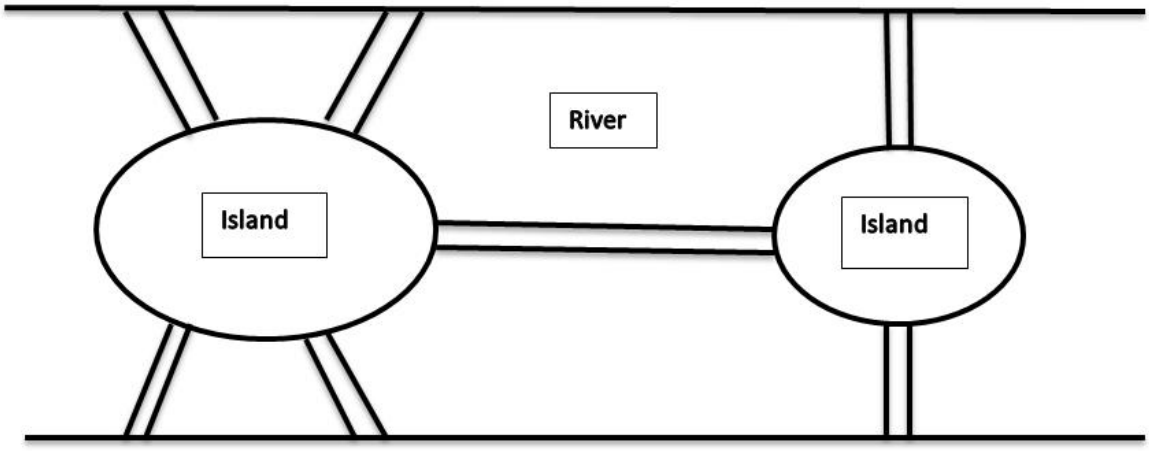


Fig. 13-2 The Bridges of Konigsberg

He first simplified the problem by presenting it in the form of the graph pictured in Fig. 13-3.

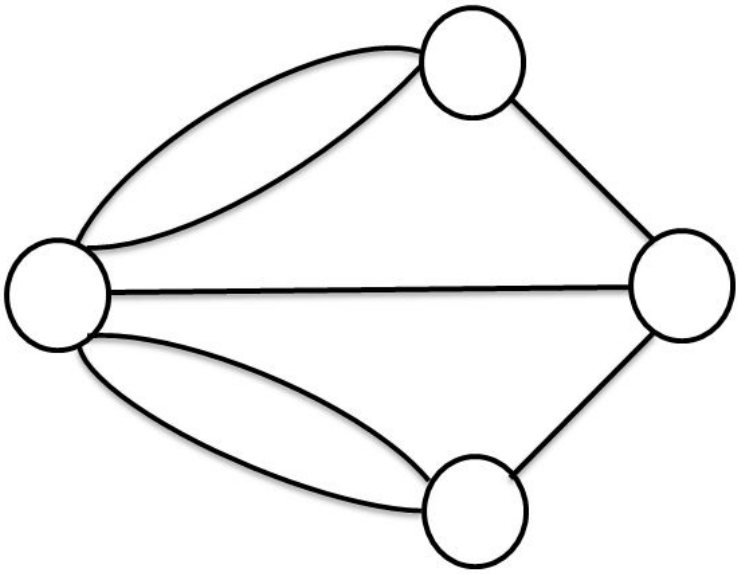


Figure 13-3 – An Euler Graph of the Bridges of Konigsberg

The hollow circles represent the vertices, which are either islands or riverbanks, and the lines represent the edges, which in this case are bridges. Euler then hit upon the key concept of defining the **degree** of a vertex as the number of edges meeting at that vertex. Fig. 13-4 presents a graph with the degrees of each vertex inside the circle that represents the vertex. A vertex is called an **even** vertex if the degree of that vertex is even, and is called an **odd** vertex if the degree of that vertex is odd.

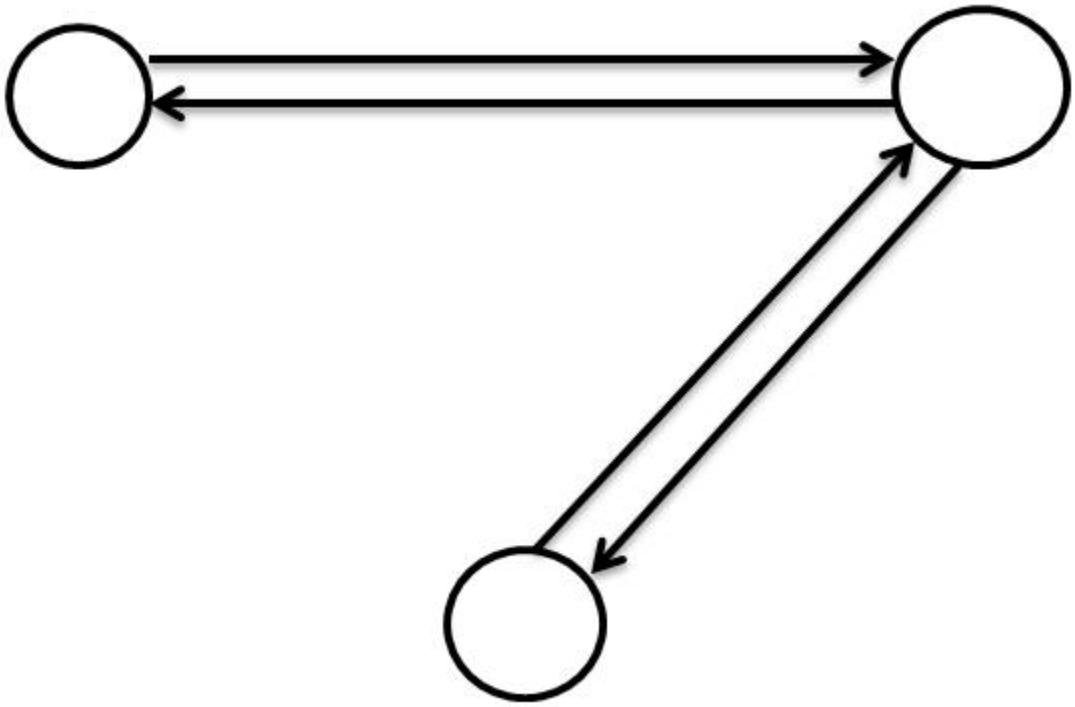


Fig. 13-4 A Graph for Which All Edges Can Be Traversed Precisely Once

Notice that, in Fig. 13-4, there is a route which traverses all the edges precisely once and ends where it started. Start at the top left, follow the arrows to the top right, to the bottom, to the top right and then back to the top left. Notice that the degree of each vertex is even; the top right vertex has degree 4, and the other two vertices have degree 2.

Euler recognized that the traversibility of a graph was determined precisely by the number of odd and even vertices. His rules are summarized in the following table.

Euler's Results on Traversibility of Graphs

1) Any graph has an even number of odd vertices.

2) If a graph has no odd vertices, one can find a route starting from any vertex which traverses every edge exactly once, and ends where it started.

3) If a graph has precisely 2 odd vertices, one can find a route which traverses every edge exactly once, but it must start at one of the odd vertices and end at the other.

4) Unless a graph has 0 or 2 odd vertices, it is impossible to find a route which traverses every edge exactly once.

Of course, all of these can be proved; some more simply than others. For instance, since each edge connects two vertices, the sum of the degrees of all the vertices must be an even number. If there were an odd number of odd vertices, the sum would be odd; this proves (1).

Applying Euler's rules to the Bridges of Konigsberg graph (this time with the degree inside each vertex) in Fig. 13-5, we see why it is impossible to find a route which traverses all the edges exactly once: there are four odd vertices.

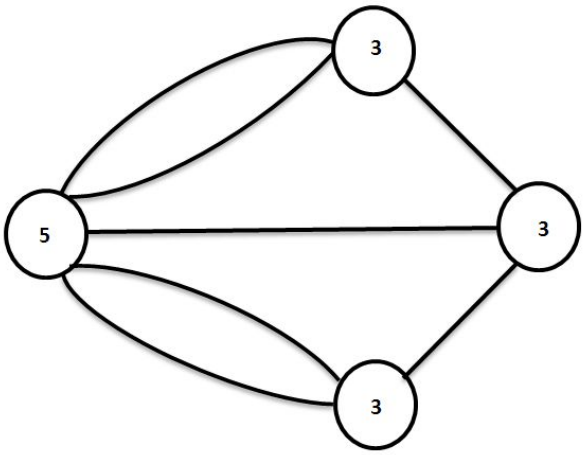


Fig. 13-5 – Euler Graph of Bridges of Konigsberg with Degrees

However, if we were allowed to build another bridge connecting the two islands, each of the islands would now be even vertices, while the two riverbanks would be odd vertices. According to Euler's rules, we should be able to find a path which traverses each edge exactly once, but it must start on one side of the river and end on the other. The graph is shown in Fig. 13-6.

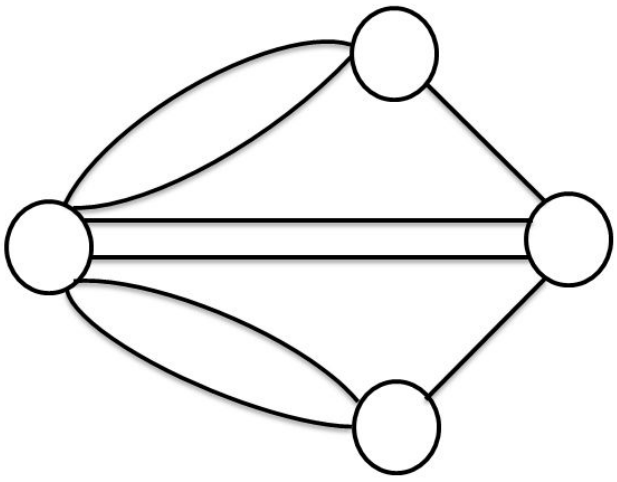


Fig. 13-6 Graph of Bridges of Konigsberg with One Additional Bridge

Finally, if we were to build one more bridge, this time across the river, we would be able to traverse every edge exactly once, and return to the point from which we started. It should not be necessary to draw the extra bridge in Fig. 13-6, as we merely use it to cross back to the original side of the river after we have completed the route in Fig. 13-6.

A route which traverses every edge of the graph exactly once is called an **Euler circuit**. Our constructions in Fig. 13-6 show that it is possible to take a graph which has no Euler circuit and simply add extra edges to create a new graph which (1) contains the old graph, and (2) has an Euler circuit. In fact, when we look at Euler's rules, we see exactly how this should be done: simply connect pairs of odd vertices, just as we did with the Konigsberg bridges. Euler's first rule says that the number of odd vertices is even, and this guarantees us that odd vertices, like socks, always come in pairs.

This process of adding extra edges to a graph to create a new one which has an Euler circuit also shows us what concessions must be made in case we are not allowed to add extra edges (for example, it might not be financially or aesthetically advisable to build extra bridges in Konigsberg). In order to traverse every edge once, we will have to re-use some of the edges. Which edges should we re-use? Obviously, those edges which connect odd vertices.

When the Goal is to Visit All the Vertices

As we saw in the story, the problem of visiting all the vertices in a graph while minimizing the total distance traveled is known as the 'traveling salesman' problem, abbreviated TSP.

If we assume that a salesman has to visit n different cities before returning home, he has a choice of n different cities to visit first. From there, he could go to any one of the (n-1) unvisited cities, and from that city to any one of the (n-2) unvisited cities, etc. By the Chinese Restaurant Principle, we see there are a total of n x (n-1) x (n-2) x ... x 1 = n! different routes that he could take.

We may recall from the chapter on counting that, even for fairly small values of n, such as 25, n! is an astronomically large number. Even the fastest supercomputer would take billions of years to examine the total distances of each of 25! different routes. As a result, mathematicians have examined two different questions.

Question 1 - Is there an algorithm which will enable one to examine only a select handful (this can be defined in mathematical terms, but we won't bother) of routes and still come up with the shortest route?

Question 2 - Is there an algorithm which will enable one to examine only a select handful of routes, and come close to the shortest route?

As Pete points out in the story, as far as Question 1 is concerned, no one even knows whether such an algorithm exists, although the betting in the mathematical community is that it doesn't. The traveling salesman problem is an example of what mathematicians call an **NP-complete** problem. There are many extremely important such problems, and they generally involve a factorial number of possibilities.

Consider, for instance, a scheduling problem that might take place in a typical factory, such as the problem of assembling a car, or a TV, or a circuit board. Many different subtasks have to be performed, and although some must clearly follow others, it is often possible to perform many of the subtasks in any order. Here it is necessary to minimize the total time, or possibly the total cost, of performing the entire job, but it is just the TSP in another guise.

Any problem which involves factorials is troublesome, because a problem with 'factorially many' computations to make requires far too many computations even for fairly small numbers. As we have seen, a traveling salesman visiting 25 cities is way beyond the power of even the fastest supercomputer to handle, and variations of the TSP often have the equivalent of thousands of cities.

In 1971 Stephen Cooke, a mathematician at the University of Toronto, showed that, if one NP-complete problem could be solved, they could all be solved. A result such as this does not tell *how* to solve a problem, but it yields a certain amount of insight. Additionally, it shows that (1) if someone can solve just one of these problems, they can all be solved, and (2) if someone can show that just one of these problems can not be solved, there is no need to "waste time" trying to solve any of the others. To date, no one has solved any NP-complete problem, but no one has shown that they cannot be solved, either.

More progress has been made in answering Question 2. It is fairly easy to describe an algorithm, such as the 'nearest neighbor' algorithm that appears in the story, and to execute it. What is substantially more difficult is to figure out how good that algorithm is. Obviously, any TSP has a 'best' answer -- the number of miles of the shortest route. If one could find an algorithm which guaranteed an answer within, say, 10% of the best answer, this would obviously represent substantial progress. Several algorithms have been devised which give excellent results with problems that occur in the real world, but no algorithm is yet known which guarantees 'coming close' in all cases.

Because of the immense practical value of the TSP and its related NP-complete problems, this is one of the most intensively investigated of all mathematical problems.

Calculating Task Complexity

A task which is 'do-able' in N^2 steps, or N^8 steps, or N^p steps for any fixed integer p is said to be a **polynomially complex task**. Polynomially complex tasks have the following property: the price tag for increasing the number N by "just 1 more" gets smaller and smaller as N gets larger and larger.

Consider a task which is "do-able" in N^3 steps. If N = 10, then the task requires 1000 steps. If N = 11, the task requires 1331 steps, an increase of about 13.3%. However, if N = 100, the

task requires 1,000,000 steps, but if N = 101, the task requires 1,030,301 steps, an increase of only about 3%. This "price tag" gets smaller as N gets larger.

The next stage of task magnitude is the **exponentially complex task**, which might require 2^N steps to complete. For such a task, the price tag of "just one more" is always the same -- the task time doubles. Obviously, any algorithm which can reduce an exponentially complex task to a polynomially complex task represents a tremendous potential savings in time, especially when N is large.

The ultimate horror show in task complexity is the **factorially complex task**, such as the TSP. As we have seen, a traveling salesman who must visit N cities has a choice of N! possible routes. The cost of "just one more" for a factorially complex task gets worse and worse as N gets larger and larger. In fact, since (N+1)! / N! = N+1, we see that the cost of going from N to N+1 increases by an ever-increasing factor of N+1.

Notice that the 'nearest neighbor' algorithm discussed in the story represents reduces a factorially complex problem to a polynomially complex one. Applying the 'nearest neighbor' algorithm to an N-city TSP requires one simply to look at N numbers for the first intercity trip, then N-1 numbers for the second intercity trip, N-2 numbers for the third intercity trip, and so on. This would give a total of N + (N-1) + ... + 1 computational steps, and we know (from the chapter on patterns) that this total is N x (N+1)/2, which is less than N^2 .

The 'nearest neighbor' algorithm is often called a **greedy** algorithm because it decides at each stage what is best according to a certain rule, and then hopes that this step-by-step plan will give the best overall result. This is somewhat akin to an individual who eats the first item of food he sees whenever he is hungry, and hopes that by so doing his nutritional needs will be best satisfied. It is possible, but highly unlikely.