11 Estimating Arithmetic

Numeracy

Most children enjoy their first encounters with arithmetic. Mastering the basic operations is "empowering" – it gives one power over numbers. After years of practicing these operations over and over – adding long columns of numbers or multiplying long strings of digits – it becomes tedious. Students come to associate mathematics with boring drill. Exploration of mathematical concepts bogs down because of the need to perform calculations. Calculators and computers permit one to study mathematical *ideas* and free one from long, error-prone computations. That's good.

On the other hand, the use of calculators often leads people to be alienated from numbers. They can get numeric results quickly, but they often fail to understand what the numbers *mean*. When a student takes out a calculator to figure out what 12% of 200 is, or to figure what percentage 20 is of 1000, she demonstrates innumeracy. That's bad.

To divide one integer by another integer and report the result to eight or ten decimal places is innumerate and ignorant. Such precision is utterly bogus and meaningless, but the calculator or computer gives results to that precision. People learn to trust the calculator, even when it seems to tell them that 42 is 0.251497005988% or 2.51497005988% of 167. Numerate people would notice that these answers are *obviously* wrong.

Numerate people know how to do manual calculations, and how to simplify the task of manual calculation. They should also know how to use calculators, without losing the "hands-on" familiarity and "feel" for numbers that comes from intimate acquaintance. To be numerate requires that one have a "vocabulary" of numbers against which one can compare other numbers that one encounters. One should know (approximately) the population of one's city, one's province or state, one's country, and the world. One should have a bunch of distances (how wide is North America? how far away is Europe? the moon? the sun? the nearest star?). Sizes of commonly-met things (cars, bricks, books, packages of 500 sheets of paper, your hand, your height, the height of one storey of a building) are useful for comparisons.

One should exercise this vocabulary frequently. When driving, estimate (mentally calculate) the distance traveled, the average speed, the distance remaining to the destination. While shopping, practice keeping a running total in one's head. Exercising one's "arithmetic muscles" leads to a familiar ease with numbers, just as frequent reading and talk improve

one's ability to use one's language. Edward MacNeal[1] suggests that you "don't count anything, look up, or ask for any figure without estimating it first."

Most important of all, *don't skip over numerical claims* when reading an article or book. Think about the numbers that are presented. Consider carefully what they mean. Are they credible? Are they relevant to the author's point? Learn to judge the reasonableness of the numbers you read.

John Allen Paulos makes a convincing case for the claim that innumeracy leads to gullibility and makes us prey to demagogues and all kinds of moronic pseudo-science like "biorhythms" and astrology. Innumeracy makes it difficult to distinguish between science and nonsense. Tabloid claims are ranked higher than the measured opinions of scientists because the innumerate can understand the tabloids, while science requires more thinking.

Inability to think clearly about probabilities leads to irrational fears and prejudices that can interfere with our lives. Without the ability to evaluate probability, *any* sort of coincidence may seem meaningful. Statistical reports and the results of polls make up a huge proportion of our news, but most citizens are incapable of making intelligent judgments about what they mean.

Up to this point, this book has emphasized logic and mathematics as abstract sciences. They are products of human imagination and creativity, comparable in beauty and significance to anything humans have produced.

This last section is concerned with mathematics as a practical ability without which nobody should be considered educated.

Using Numbers Intelligently

How many numbers are there? "Infinitely many" is *correct*, but it is not a *good* answer. First, it is vague. As we have seen, there are several "orders" of infinity. There are infinitely many natural numbers, and there are infinitely many real numbers. However, there are more real numbers between 0.000000… and 0.999999… than there are natural numbers altogether. "Infinitely many" tells only part of the story.

Second, the "correct" answer fails to recognize that the question is ambiguous. The answer to "How many numbers are there?" depends on the *purpose* of the numbers.

Everyday non-technical communication does not use infinitely many numbers. As Edmund C. Berkeley[2] explains, non-technical communication uses number-words. "Familiar numbers" are those that can be expressed using no more than two number-words. The first

---

1 Edward MacNeal, *Mathsemantics: Making Numbers Make Sense* (New York: Viking Penguin, 1994).

2 Edmund C. Berkeley, *A Guide to Mathematics for the Intelligent Nonmathematician*, (New York: Simon and Schuster, 1966).

number-words are the common numerical words that count units. Berkeley lists "14 such words expressing 12 numerical ideas: 'one,' 'a,' 'two,' 'three,' 'four,' 'five,' 'six,' 'seven,' 'eight,' 'nine,' 'ten,' 'eleven,' 'twelve,' 'dozen.'" To some of the words from this list we append the suffixes "-teen" or "-ty" (to get words like "seventeen" and "twenty") or words from the list of powers of ten – "hundred," "thousand," "million." "Billion"[3] is becoming a "familiar word" because of concerns with national debts and population. Then we have words like "half," "third," "quarter" and the suffix "-th" (with which we make "fifth," "tenth," "hundredth," etc.). And we have a bunch of words for 0 ("none," "zero," "naught," "no," etc.). When we put more than two of these terms together to express a number it ceases to be *ordinary* discourse. It becomes *technical*. Non-technical numbers include just those that can be made with pairs of terms from these choices.

Using just the words for numbers from 1 to 12 combined with the powers of ten ("millionth," "thousandth," "hundredth," "tenth," "-teen," "hundreds," "thousands," "millions," and "billions" – 9 terms in all) we can make only $12 \times 9 = 108$ two-word numbers. With the 12 non-combined terms we have 120 numbers. If we include the indefinite numbers ("some," "a few," "many" or the numbers we get by adding the suffix "-s" as in "millions," "tenths") we have about 150 different numbers in non-technical usage. Slightly-technical usage using three terms (as "forty-seven," made from "four"+"-ty"+"seven") increases this to a mere 1500 numbers. Compared with some cultures which use only three numbers ("one," "two," and "many"), our smallish collection is very powerful and expressive. But **there are only about 1500 numbers** in ordinary non-technical use.

How about technical usage? How many numbers are there <u>in science</u>?

In empirical science, the only numbers that are meaningful are those that can result from measurement. This depends on how precisely we can measure things with the best instruments we have. The greatest precision I know of is measurement to eight significant decimal digits. That is, we can measure precisely a value like 1.2345678 (or 12,345,678), but we cannot distinguish (measure any difference between) 1.23456775 and 1.23456784 (or between 12,345,677.5 and 12,345,678.4). **In science, if we cannot measure a difference between two numbers, they are the same number**. There are only 100,000,000 (10^8) different strings of eight digits.

Scientists express numbers as a coefficient times a power of ten. One hundred would be expressed as 1.0×10^2 or as 1.0E2. A million would be 1.0×10^6. One one-thousandth (1/1000) would be 1.0×10^{-3}. Assuming that no more than eight digits of precision are meaningful, there are only 10^8 numbers we can use for coefficients.

The smallest number in science is about 10^{-40}, which is 1 divided by 1 followed by 40 zeros. The largest number is about 10^{120}, or 1 followed by 120 zeros. The range from 10^{-40} to 10^{120} requires 161 powers of ten. Each of them can be used with any of the 10^8 coefficients, so **there are 161×10^8 (more than 16 billion) positive numbers in science, and the same**

3 An American billion is one thousand million (10^9). A British billion is one million million (10^{12}).

number of negative numbers. 32 billion is a lot of numbers, but it is far from infinitely many.

Data for Estimating

Estimating means forming an approximate idea of a quantity (distance, size, cost, count) without actually measuring or counting.

How far is Montreal from Toronto? The speed limit on the highway to Toronto is 100 km/hr. I drive slightly faster than that, but I stop at least once. I estimate my average speed at about 90 km./hr. My usual trip time is between five and six hours. I estimate the distance as something between $90 \times 5 = 450$ km. and $90 \times 6 = 540$ km., and call it about 500 km. A road map says it's 542 km., so 500 km. is not a bad estimate. I made another estimate by measuring the straight-line distance on the map and got 510 km. It's somewhere in that ballpark, and "ballpark figures" are what one aims at in simple estimating.

The two processes in estimating a number are: (1) gather some data (observations, experiences, facts, statistics, reasonable assumptions, etc.) to base the estimate on; (2) use logic and arithmetic to operate on those data to arrive at the number.

Gathering data is much easier if you already have a good numerical vocabulary at your fingertips. If you don't have relevant data in your memory, you'll have to do some research. Two points should be made about data-gathering. First, don't do more research than the estimate is worth. If you're going to do a lot of intensive research, you might as well look up or calculate the number instead of making an estimate. Second, your estimate will never be more precise than the least precise of your data-numbers. Both my time (5-6 hours) and speed (~90 km./hr.) were accurate to only one figure. That is, my time could be anywhere between five and six hours and the average speed could be anywhere between 85 and 95. So my estimate should be expressed with no more than one significant figure (500 km.). To express my estimate as 450 or 540 km. would be misleading.

Keep your eyes and ears open to pick up data and facts for estimating. Notice that a long drive at what seems to be a steady 110 km./hr. is probably at least 10% slower. How many weeks per year does a person usually work? How many pages are there in an average-size book? How many words on an average printed page? How many lines per inch does a computer printer normally print? How many characters are there on an average line of type?

At least some of the following data are worth remembering:

Montreal to Toronto	540 km. (by road)
Halifax to Vancouver	5,500 km. (by road)
New York to Los Angeles	4,500 km. (by road)
diameter of the earth	13,000 km.
circumference of Earth	40,000 km.
earth to moon	390,000 km.
earth to sun	150 million km.
sun to Pluto	6 billion (10^9) km.

speed of light	300,000 km./sec.
1 light-year	9.5 trillion (10^{12}) km.
sun to nearest star	4.3 light-years
500 sheets of paper	3 cm. thick

A minimally-numerate person (or *anyone* who presumes to call himself or herself "educated") should be familiar with the relation between the most-common fractions and their representation as percentages and as decimal numbers. "*Per*" is a Latin word meaning "for each" (as in "one per person"). *Centum* means "hundred." So 1 per cent means 1 out of each hundred. One should be able to convert any fraction of a hundred instantly into a percentage – 24 out of 100 is 24% – right away. One should know that 1/2 is 50% or 0.5; 1/4 is 25% or 0.25; 3/4 is 75% or 0.75; 1/8 is 12.5% or 0.125; 1/3 is ~33%[4] or ~0.33; 2/3 is ~67%; 1/5 is 20%; 2/5 is 40%; 3/5 is 60%; 4/5 is 80%; 1/6 is ~17%; 1/7 is ~14%; 1/9 is ~11%; 1/10 is 10%; and so on. One should be aware that there is a nice relationship so that 1/9 is ~11% and 1/11 is ~9%; 1/5 is 20% and 1/20 is 5%. You should know that a 200% increase results in a number that is three times as big as the original number, and that three times as big means 300% of the original number. Know that increasing a price by 50% and then lowering the resulting price by 50% will not return you to the same number.

You should also have (or have ready access to) common formulas for calculating volumes and areas and converting units (e.g., metric to English) and so on.

Significant Figures and Orders of Magnitude

Scientific notation, using a coefficient and an exponent (power of ten) is useful for recording very small and very large numbers economically. 4×10^{20} is a lot easier to write than 400,000,000,000,000,000,000.

Another advantage of scientific notation is that one can indicate very clearly just how **precise** a number is. When you look at a number like 4 followed by 20 zeroes, it is not clear whether that number is precise down to the last digit or not. But 4×10^{20} announces its **precision** explicitly; the number is somewhere between 3.5×10^{20} and 4.5×10^{20}. If a scientist wrote 4.0×10^{20} she would be announcing that the number was between 3.95×10^{20} and 4.05×10^{20}. It is precise to two digits. 4.00×10^{20} is precise to three digits of accuracy, so we know it is between 3.995×10^{20} and 4.005×10^{20}. And so on.

The coefficient of a number written in scientific notation contains the number's **significant figures**. The power of ten indicates the number's **order of magnitude**. A number is of a **higher order of magnitude** than another when its magnitude is a higher power of ten. So when someone says that one number is "of a different order of magnitude" than another, he is saying that it is at least 10 times as large or small as the other. Or he's just ignorant.

---

4 The "~" symbol here means "approximately."

Simply Useful Arithmetic

Ordinary arithmetic addition is boring and error-prone when the numbers contain many digits. Use a calculator or computer to do ordinary arithmetic addition of multi-digit numbers or long lists of numbers. However, even adding with a calculator can produce errors. One should estimate the sum of the numbers one adds, to check on the result. Use the estimate to verify that the calculated sum is reasonable ("in the ball-park").

In estimating arithmetic, we simplify the numbers and use shortcuts to get an approximate answer. A couple of examples will illustrate different methods.

Suppose we are adding the following prices:

$2.49	$0.84	$5.22	$3.16	$1.89
$4.69	$1.99	$4.71	$5.45	$4.95
$5.49	$5.50	$3.62	$3.16	$1.50
$4.87	$5.16	$1.36	$1.07	$4.43

The average price on the list seems to be about $4.00. There are 20 items, so we'd expect the total to be in the neighborhood of 4×20 or about $80. To estimate this total more accurately, round off the amounts to the nearest dollar, giving 2+1+5+3+2+5+2+5+5+5+5+6+4+3+2+5+5+1+1+4. Adding these single-digit numbers is almost as easy as counting, and gives a sum of $71. The exact sum is $71.55.

As a second example, add this list of large numbers:

1,352,795	3,562,979	2,029,969	3,667,411	3,121,915
486,190	519,808	1,814,640	107,178	1,679,702
3,053,271	1,640,387	674,252	3,176,745	499,858
555,131	811,031	448,264	1,204,260	181,893

There are 20 numbers, and the average seems to be somewhere around one million. A first estimate would be approximately 20 million. A second, more-careful estimate would be to round the numbers off to the nearest million and add. The sum (going from left to right on each row) would be 1+4+2+4+3+0+1+2+0+2+3+2+1+3+0+1+1+0+1+0. The answer we get is 31 million. Another way is to add just the millions digits 1+3+2+3+3+ 1+1+3+1+3+1, getting 22 millions, and then add the number of times that the hundred-thousands digit (the sixth digit from the right) is greater than 4 (9 times). 22+9 gives 31 million. The exact answer is 30,587,679. 31 million is very close than we would ordinarily expect. 20 million is not very close, because our estimate of the average size of the numbers was not good.

Two methods of estimating a sum have been presented. The first (less reliable, unless you're good at estimating the average of a list of numbers) is:

1. **estimate the average** item to be added, **A**;
2. **count** the number of items to be added, **N**;
3. take **A times N** as an estimate of the sum.

This turns a long addition into a short multiplication.

The second (more reliable) method is:

1. **add the most significant digits** only (making sure to use the most-significant figure of the largest number, and only adding digits of the same significance from every number), obtaining **S**;
2. **count** the number of times the digit in the same place as the second-most-significant digit of the larger number is greater than 4, obtaining **H**;
3. then **S plus H** is an estimate of the sum.

As an example of the importance of using the most-significant digit of the largest number and using the same digit for all the numbers, consider the sum of 7,618 plus 21,143,211 plus 562 plus 3,224. Using only the most-significant digit of the largest number, our estimate would be 20 million. The three smaller numbers *have no effect on the estimate*. This may be clearer when we re-write the numbers in scientific notation so that we emphasize the orders of magnitude of the four numbers. In scientific notation they are 7.618×10^3, 2.1143211×10^7, 5.62×10^2, and 3.224×10^3. 10^7 is **four orders of magnitude** greater than 10^3; it is 10^4 (10,000) times as big.

A drop of water contains less than 0.5 cc. of water, so a drop of water is less than $1/10^4$ of the volume of a 5-liter bucket. So a number that is four orders of magnitude smaller than another is literally "a drop in the bucket."

It is sometimes easier to estimate a sum if you re-write the numbers in scientific notation. Start with the largest number, and skip any number that is two or more orders of magnitude smaller. For example, if the biggest number is two million (2×10^6), you'd ignore any number smaller than 1×10^4.

Be careful, however, if the list contains one or two really big numbers and many small numbers. If you were estimating the total of all the salaries of all the employees of a company, where the CEO makes 5 million per year, and a hundred employees make salaries between $20,000 and $50,000, you could not leave out the other employees' salaries, even though they are two orders of magnitude less than the CEO's. In this case, a better procedure would be to estimate the sum of the other employees' salaries using one of our methods (here, the **A** times **N** method would probably be adequate). In this case, the number might be somewhere around $100\times3\times10^4$ or $1.0\times10^2\times3\times10^4 = 3\times10^6$. Added to the CEO's 5×10^6, we get 8×10^6, or 8 million dollars. If we had just ignored the other salaries, we'd have estimated the sum as only 5×10^6 – a pretty bad estimate.

How bad was that estimate? We express the **imprecision of an estimate** as a percentage. We find out how far our estimate was off (in the last case, it would have been off by 3×10^6), and divide that by the actual value 8×10^6, giving 3/8 or about 38%. We were about 38% off in our estimate. There was an error of ~38%.

Subtraction usually just involves two operands rather than a list. The main clue in estimating-arithmetic subtraction is to pay close attention to the order of magnitude of the numbers. When you subtract 30 from 4000, the estimated answer is 4000. In subtraction, precision is important if the operands are approximately equal. For example, 4340 is about 4000 for estimating purposes, as is 3720. But $4340-3720$ is not $4000-4000 = 0$. 0 would be a

very bad estimate of the difference. To one digit of precision, the correct estimate of the difference is ~600.

Multiplication is the most important operation in estimating arithmetic. The first thing to remember is that your answer should never have more digits of precision than the numbers you base it on. If the operands have two significant digits, then your answer must have no more than two significant digits. The second thing is to convert all your multiplicands into scientific notation with no more than two digits of precision, one digit before the decimal point and one (at most) after.

As an example, I estimate how many cigarettes I have smoked. I have been smoking for over 40 years. That's 4.0×10. Each year contains 365 days. Call it 370 or 3.7×10^2 days per year. The total number of days I have smoked is about $4.0\times3.7\times10\times10^2$ or $15\times10^3 = 1.5\times10^4$ days. At different times in my life I have smoked more (about 60-75 cigarettes per day) or less (about 20-30 per day). Call it about 40 per day on average or 4.0×10. That's $4.0\times1.5\times10\times10^4$, or about 6.0×10^5 cigarettes. We won't leave two digits of precision in the answer, because our data estimates weren't very precise. It's about 6×10^5 cigarettes. 600,000 cigarettes! Notice that we add the powers of 10 when we multiply. That is, $10\times10^4 = 10^{1+4} = 10^5$, and $10^2\times10^3\times10 = 10^{2+3+1} = 10^6$, and so on.

To multiply by 25, you can multiply by 100 (10^2) and divide by 4. To multiply by 50, multiply by 100 and divide by 2. To multiply by 5, multiply by 10 and divide by 2. To multiply by 17, multiply by 100 and divide by 6.

Division should also be kept to two digits of precision. Again, translate into scientific notation. Dividing powers of ten uses subtraction. For example, to estimate how much those 600,000 cigarettes cost, we could estimate that, at 20 cigarettes per pack, it represents $6\times10^5/2\times10 = 6/2 \times 10^5/10 = 3\times10^4$ packages of 20 cigarettes. When I started smoking, a pack cost about $0.50. Now it's about $6.00. The average cost was probably about $2.00 per pack. So the cost was about $2\times3\times10^4 = 6\times10^4$ = about $60,000.

Another important trick in estimating arithmetic involving division is to write out all the factors of the dividend (before multiplying) above a horizontal line, and all the factors of the divisor (before multiplying) below the line, and "cancel out" similar factors. This works especially well in eliminating powers of ten.

Another trick: If you have to divide by 25, it is easier to multiply by 4 and divide by 100. To divide by 50, multiply by 2 and divide by 100. And so on.

Examples

1. How many seconds are there in a year? A year is about 365.25 days. A day is 24 hours. An hour is 60 minutes. A minute is 60 seconds. So the answer is $365.25\times24\times60\times60$. Say $4\times10^2\times2\times10\times6\times10\times6\times10 = 4\times2\times6\times6\times10^5 = 8\times40\times10^5 = 320\times10^5 = 3.2\times10^7$ (32 million seconds in a year). My calculator gives 31557600, or 3.16×10^7. Not bad, especially when you consider that I rounded off 365.25 to 400, and 24 to 20, and

6×6 (36) to 40. The error of my estimate was $3.2\times10^7 - 3.16\times10^7 = .04\times10^7 = 4\times10^5$. The percentage error was $.04\times10^7 / 3.16\times10^7$. Canceling the 10^7 from the numerator and denominator, we get .04/3.16, or about 4 in 300, or about 1.3%.

2. How long is a million seconds? A million seconds is 1×10^6 seconds. Divide that by 60 (seconds in a minute) to get $1\times10^6 / 6\times10 = 1\times10^5 / 6$. This is about $\mathbf{\frac{1}{6}\times10^5}$. Remembering that 1/6 is about 0.17, or 1.7×10^{-1} we get $1.7\times10^{-1}\times10^5 \cong 1.7\times10^4$. There are about 17,000 minutes in a million seconds. There are 60 minutes in an hour, so 1.7×10^4 minutes is $(1.7\times10^4)/6\times10 \cong 0.3\times10^3 = 3\times10^2$ or 300 hours. Since there are 24 hours in a day, this is about 12 (rounding 24 off to 25 hours/day) days. The calculated answer is 11.57 days. Another way to calculate a million seconds is to use the answer to the previous question to discover that a year is about 32 million seconds, so a million seconds is about 1/32 of a year. 32 goes into 365 somewhere between 11 and 12 times.

3. How long is a billion seconds? An American billion is a thousand million. A million seconds is 1.16×10 days, so a billion would be $1.16\times10\times10^3 = 1.16\times10^4$ days. Since a year is about 4×10^2 days, this is about $(1\times10^4)/4\times10^2 = (1\times10^2)/4 = 0.25\times10^2$ or 25 years. The calculator gives 31.69 years. Think about these last two numbers. We all have a "feel" for how long a second is. A million seconds is **12 days**! But look how much more a billion is than a (mere) million! A billion seconds is **almost 32 years**!

   If a package of 500 sheets of printer paper is about 3 cm. thick, then 1000 sheets is about 6 cm. A million is a thousand thousand sheets, 6,000 cm. thick. That's 60 meters, or about the height of a 20-storey building. A billion is a thousand million, so a billion sheets would be 60,000 meters or 60 km. high.

4. If you shrink the Earth to the size of a basketball (about 30 cm. diameter), how far above the surface of the oceans would Mount Everest stick up? Everest's peak is about 9 km. above sea level. Earth's diameter is about 13,000 km. So Everest would stick up about $\mathbf{\frac{9\times30}{13000} = \frac{2.7\times10^2}{1.3\times10^4} \cong 2\times10^{-2}}$ cm., or **0.2** millimeters. It would be so small you would have trouble finding it. If the Earth were the size of a billiard ball, Everest would be too small to count as an imperfection.

5. Recently the world population was about five billion people. If we lined everybody up as if they were queued to go through a turnstile (with two people per meter), the length of the line would be $0.5\times5\times10^9 = 2.5\times10^9$ meters $= 2.5\times10^9/1\times10^3 = 2.5\times10^6$ km. The distance to the moon is about 3.9×10^5 km. $(2.5\times10^6)/(3.9\times10^5) \cong 6$. We could make 6 lines of people that reached all the way to the moon. That's a lot of people! If we put every man, woman, and child into military formation one meter apart (so each person took up one square meter), we could put 1000×1000 = one million people per square kilometer. The whole five billion would occupy $5\times10^9 / 1\times10^6 = 5\times10^3 = 5000$ square kilometers. They would all fit into Trinidad.

Internal Consistency

When you read or hear a number, think about whether the number makes sense in that context.

The *Bible* says that the Great Flood covered the highest hills as the result of forty days and nights of rain. Does that make sense? Mount Ararat was one of the hills that were covered. The top of Ararat is about 5200 meters above average (mean) sea level. So 5200 meters of water would have to fall in $40 \times 24 = 960$ hours. Call it 5000 meters in 1000 hours. That's five meters of water per hour. Too much to be possible. Didn't happen.

When science (or other) fiction stories describe gigantic people or insects, think about the numbers. If a woman were 15 meters tall (the "fifty-foot woman"), she would be about ten times as tall as a normal human woman. Her bones would be about ten times as thick. The cross-sectional area of her thighbones would be about $10 \times 10 = 100$ times as great, so they could carry about 100 times as much weight.[5] But the woman's volume (length$\times$ width$\times$height) would be 1000 times as great, as would her weight. Her bones would be ten times as likely to break. Her muscles might be 100 times as strong, but they'd have to move 1000 times as much mass.

An ant that was two meters long (as in the movie *Them*) is about 200 times as long as a large ordinary ant. Its legs would be 200 times as thick. It would weigh 40,000 times as much. Its legs couldn't carry it, and its own weight would cause it to squish.

John Allen Paulos[6] gives a number of examples from newspapers. He quotes Khalid Abdul Muhammad as saying that 600 million African-Americans died because of slavery. Is this consistent with the fact that the total number of slaves brought to the New World was between 8 and 15 million? Louis Farrakhan has said that Jews owned 75 percent of African slaves. At the beginning of the Civil War, the 20,000 southern Jews were only 0.22% of the southern white population of 9 million. Of the 1.9 million slaveholders, the 5,000 Jewish slaveholders were only 0.26 percent. It is highly unlikely that the average Jewish slaveholder had 300 times as many slaves as the average non-Jewish slaveholder, but that's what these numbers would entail.

Edward MacNeal[7] gives an example of a report in the February, 1989 issue of *Soviet Life*, saying that 2,700 Soviet couples get married every year. The number looked wrong to him, as it should have looked wrong to you. There were more Soviets than Americans, and there were 250 million Americans. Assuming an average lifetime of about 70 years, one could estimate that about 1/70 of the population would reach marriageable age each year. If there were 280 million Soviets, 1/70 of the population would be four million people reaching marriageable age. That's about two million marriageable couples. Is it likely that less than 3,000 couples

---

5 Not really, because a thighbone acts like a column and a column ten times as tall will bend and break more easily, even if its diameter is ten times as great.

6 John Allen Paulos, *A Mathematician Reads the Newspaper* (New York: Basic Books, 1995).

7 Edward MacNeal, *Mathsemantics: Making Numbers Make Sense* (New York: Viking Penguin, 1994).

(about three people out of every two thousand reaching marriageable age) got married? One could reasonably assume that the correct figure was 1000 times as large as reported, or about 2,700,000 people (1,350,000 couples).

Adding percentages causes really silly bloopers. A.K. Dewdney in *Scientific American* magazine cited a newspaper headline that claimed "Seven Italians Out of Ten Have Committed Adultery." The survey on which the headline was based had found that 49% of Italian men and 21% of Italian women surveyed confessed to extramarital affairs. If the number of women and men are about equal, and if the survey was representative of the whole population, then 49% of half the population (men) and 21% of the other half committed adultery, so 35% of the whole population committed adultery. The paper added the two percentages and got a number that was twice as high as it should have been.

Thinking carefully about numbers in context can improve your understanding of a complex situation. Very large numbers can leave one numb to the reality they describe. Get other numbers to compare with. Given a number like the amount of the Canadian national debt, turn it into a more-meaningful number by estimating how much each Canadian would owe if the debt were transferred to individuals, or how many BMWs one could buy with that much money.

A large (and growing) number of people dies of AIDS each year. To get some perspective, compare that number with the number of children who die of respiratory infections (eight times as many per year) or diarrhea (six times as many).

Numbers that get reported more frequently may seem more important. Deaths due to cocaine or heroin make news. About 8,000 people per year die due to cocaine, and about 6,000 due to heroin. However, tobacco is blamed for 400,000 deaths each year, and alcohol for 90,000. What is the *real* drug problem?

A commuter airplane crash makes news, but the number who die each year from smoking-related illness is the equivalent of three fully-loaded jumbo jets crashing every day of the year. The attention we give airplane crashes and heroin fatalities is disproportionate to their real significance.

Watch out for numbers that seem to be too "round" and numbers that seem to be too precise. Paulos reports a recipe which, after giving very approximate instructions and quantities of ingredients, states that the dish contains "761 calories, 428 milligrams of sodium, and 22.6 grams of fat per serving." Some parents worry when their children have a temperature of 99 Fahrenheit degrees, because "normal" is 98.6 degrees. They don't realize that "normal" body temperature can vary quite a lot, and the spuriously-precise 98.6ºF. figure was calculated from an approximate Celsius temperature of 37ºC. 37ºC. is somewhere between 36.5 and 37.5 degrees Celsius, or between 97.7ºF. and 99.5ºF. 22.6 grams of fat and 98.6 Fahrenheit degrees are meaninglessly (spuriously) precise.

12 Probability Theory

Theories of Probability

The **theory of probability** is the theory that aims to provide principles for estimating the likelihood that something will happen or has happened.

"Probability" is only meaningful where (1) it is impossible or impractically difficult to know or infer exactly what will happen (or has happened), and (2) we believe that the outcome is at least partly **random**. The concept of randomness is fundamental to probability and statistics. Deterministic phenomena[1] are not random. If one knows their causes, one can predict them with certainty, rather than merely probabilistically.

Some phenomena (e.g., the emission of a particle by a radioactive atom) are random. Other phenomena are parts of complex interacting systems where all of the interactions are the result of the operation of simple causal laws, but the huge number and complexity of the interactions make deterministic prediction impossible. We treat such systems as if their behaviour were random.

Randomness is more than just unpredictability. There must be a **long-term pattern or regularity** in the phenomena. If coins preferred coming up heads and occasionally (but irregularly) came up heads by choice, coin flipping would not be random. Although it would be unpredictable, there would be no long-term pattern.

When we use the word "**random**" in probability theory, we mean:
1. **The exact outcome is not predictable in advance of its occurrence.**
2. **A predictable long-term pattern of outcomes exists. We can predict the relative frequencies of the different outcomes in a series of many trials.**

Probability is conceived in several mutually-incompatible ways. Some of the notions of "probability" are (1) probability as believability, (2) the *a priori* notion of probability (mathematical probability), and (3) probability as relative frequency.

---

1 "Deterministic phenomena" are events where science has discovered the laws by which we can predict what will happen or retrodict what has happened when we know the conditions that prevail at some other time.

1. Believability

1(a) Personal or subjective probability

People estimate probabilities in an informal way that reflects their personal judgment about the likelihood of an event. Personal probabilities are more-or-less subjective. They may be based on some kind of reasoning and evidence, but another person who is aware of all the same reasons and evidence might appraise the probabilities quite differently. Even when they are expressed numerically, they are usually not the result of formal mathematical reasoning, and are not vulnerable to mathematical criticism. In response to such criticism, a believer can say, "That's just the way I see it."

1(b) Credibility of an hypothesis (based on experimental evidence)

When we prove a conjecture, as in the formal sciences, we show that there cannot be any counterexample. In empirical science we try to confirm theories by experimentation. After a conjecture is proposed, scientists deduce consequences of the conjecture and conduct tests to discover if the predicted consequences happen. Conjectures that have withstood more tests and more different kinds of tests are considered more probable. Yet some future experiment could disconfirm any theory, no matter how highly confirmed by previous observations. Theories in empirical science are never proved. Just one counterexample can falsify any hypothesis.

There are many stipulations and restrictions on statements about the probability of scientific theories or laws. Whether one conjecture is "more probable" than another depends on more than just the number and kinds of tests to which they have been exposed. The claim that one hypothesis is more probable than another is not usually quantifiable (i.e., scientists cannot put a numeric value on the respective probabilities), and not usually subject to any kind of formal mathematical or logical treatment.

2. *A Priori* Probability

In many situations, (1) any of a number of outcomes is possible and (2) there is no objective evidence that any of the possible outcomes will be favoured over any of the others. In such situations, take all of the outcomes as **equally probable** (equiprobable). If we assume that a die is symmetrical, we expect that any of the six possible outcomes is equally probable. If the two sides of a coin are symmetrical and if the person flipping the coin cannot make one side come up in preference to the other, we say that it is equally probable that either side will face up when the coin is flipped.

The ***a priori* interpretation** of probability specifies probabilities in terms of the behaviour of ideal coins, dice, decks of cards, etc., where none of the possible

outcomes is favoured. **Mathematical probability** (discussed below) is based on *a priori* ideal models.

3. Relative Frequency

The belief that heads is as likely as tails can be tested empirically. Just flip a coin over and over again, counting the number of heads and the number of tails. In *Statistics: Concepts and Controversies*[2], D.S. Moore reports that, around 1900, Karl Pearson (an English statistician) spun a coin 24,000 times, getting heads 12,012 times. The **relative-frequency interpretation** of probability uses such tests to determine the probability of an outcome. From Pearson's test, we could infer that it is slightly more probable that a flipped coin will come up heads than that it will show tails.

The relative-frequency interpretation **defines** "probability" as:

If, in a long sequence of repetitions of a random phenomenon, the relative frequency of an outcome approaches a fixed number, that number is the probability of the outcome. A probability is always a number between 0 (the outcome never occurs) and 1 (the outcome always occurs).

The **frequency** of some thing or event is the **number** of occurrences of that thing or event. It is a simple count. The frequency of red jellybeans in a jar of jellybeans is the number of red jellybeans. **Relative frequency** is a **proportion** or **ratio.** The relative frequency of red jellybeans is the percentage or fraction of the jellybeans that are red.

Suppose we want to know the probability that heads will come up when we flip a coin. We flip coins thousands of times. After a certain number of tosses, we write down the number of heads, and then we continue. We might make a table like this:

Number of tosses (n)	10	20	50	100	200	500	1000	2000	5000	10000
Number of heads (m)	7	11	24	43	86	228	491	977	2495	4978
Relative frequency (m/n)	0.7	0.55	0.48	0.43	0.430	0.456	0.491	0.4885	0.4990	0.4978

We see that seven out of the first 10 flips were heads. The **frequency** of heads was 7. The **relative frequency** was 7/10 (0.7 or 70%). After 20 tosses, the relative frequency was 0.55 (55%) heads. As we flipped the coin more and more times, the relative frequency got closer and closer to 50%. That is what the definition of "probability" means when it says that the relative frequency of an outcome "approaches a fixed number." In the long run, the relative frequency of heads seems to be approaching 0.500000….

The somewhat paradoxical nature of our definition of "random" is illustrated in this table. It is impossible to predict the particular outcome of one flip of a fair coin.

---

2 David S. Moore, *Statistics: Concepts and Controversies* (3d ed.) (New York: W.H. Freeman, 1991), p. 334.

Yet we can predict very accurately and with a high degree of certainty that heads will turn up almost exactly half the time in a long sequence of flips.

Another important point is illustrated in this table. While the relative frequency of heads is getting closer and closer to 50%, the frequency (number) of heads is not getting closer and closer to the number of tails. After 10 flips, the number of heads is 4 more than the number of tails. After 10,000 flips, there were 44 more tails than heads. This illustrates a famous fallacy called the "**Law of Averages**." Many people who fail to understand probability believe that the total number of heads should approach closer and closer to half the number of tosses. Based on this misunderstanding of probability, they believe that tails is more probable after a long run of heads (as it would have to be if the number of heads and tails is to get closer to even). This is fallacious. Another example may make this clearer.

Number of tosses (n)	10	20	50	100	200	500	1000	2000	5000	10000
Number of heads (m)	2	6	29	56	107	258	508	1032	2558	5092
Relative frequency (m/n)	0.2	0.30	0.58	0.56	0.535	0.516	0.508	0.5160	0.5116	0.5092
Number away from half	3	4	4	6	7	8	8	32	58	92

The idea of probability in random phenomena is that relative frequencies display regularities in the long run. The "**myth of short-run regularity**" is another fallacy, which proposes that the long-run regularity will reveal itself in the short run as well. People who fall for this myth believe that truly random coin flipping should not result in sequences like 10 heads in a row. While such a sequence is not very common (i.e., such sequences occur with a low relative frequency and therefore a low probability), they do occur, even when the coin toss is truly random. I looked for sequences of heads and tails in a simulation of flipping a coin 1,000 times. In 10 simulations, there was always at least one run of 8 or more heads or tails. There was a run of 15 heads in a row in one of the simulations, and 11-in-a-row came up in three of them.

The usefulness of the relative-frequency interpretation of probability can be seen in the classic "taking coloured balls out of a jar" situation.

Take a large jar containing 1,000 red and white balls. You randomly (i.e., without seeing, and without deliberately choosing balls from the top or bottom, etc.) select one ball. It's white. You throw the ball back in and shake up the jar, and randomly select another ball. It's white again. The relative frequency of white balls in the **sample** (2 balls) you've observed is 2/2 = 100%, suggesting that all the balls in the jar (the whole **population**) are white. You would estimate the probability that the next ball will be white to be 100% (certainty). To avoid writing "the probability that the next ball will be white," we write **P(W),** and **P(R)** for the probability of drawing a red ball. You draw another ball, and it's white again. Your estimate is unchanged. If the next ball is red, you revise your probability estimate. Since the relative frequency of white balls was 75%, you should now say that **P(W) = 75%**. One out of four balls was red, so you estimate **P(R) = 25%**.

You could believe that your fourth draw found the only red ball in the jar, and estimate that **P(W)** is still almost 100%. Or you could believe that there is only one white ball, and the first three draws were flukes. That's what's wrong with subjective probability. What you believe to be probable depends more on your psychological makeup than on the evidence.

One great weakness of the relative-frequency interpretation is that it is no use for estimating the probability of a unique event. It only works when there can be "a long sequence of repetitions of a random phenomenon."

Mathematical probability

Theorists are somewhat divided between the relative frequency and *a priori* interpretations. In 1931, the Russian mathematician A.N. Kolmogorov axiomatized probability theory. He created a postulational system for a "calculus of probabilities."

When we use a postulational system to describe the "real world," we **interpret** it – we find an isomorphism between the postulational system and the actual systems we observe. Reality is taken as a **model** of the postulational system. We can predict and control reality just to the extent that reality models (is isomorphic to) the system.

Mathematical (postulational or axiomatic) probability theory deals with ideal coins and dice, etc. To the extent that an actual coin behaves like the ideal coin, the mathematics of probability theory predicts its behaviour. If a coin behaves significantly differently[3] from our predictions, it is not a "fair coin." A fair coin is one that behaves as our theory says an ideal coin should. It should give the same results in the (very) long run as the ideal coin described by the *a priori* notion of probability.

So the postulational theory of probability is like the *a priori* notion. Its real-world interpretation will satisfy the relative-frequency interpretation, as long as the events we are dealing with are ideally random. Probabilities determined by the *a priori* model are **theoretical probabilities**. Relative frequencies give **empirical probabilities**.

Experiments

I have been using "event" and "outcome" more-or-less interchangeably. Probability theory defines these notions by reference to the concept of an "experiment."

---

3 What sort of behaviour is **significantly different** from random is itself a probabilistic concept. We use the postulational theory of probability to calculate how much difference is a significant difference.

An **experiment** is defined as any situation in which more than one possible outcome can occur. A **random experiment**[4] is one in which the particular outcome cannot be predicted in advance, and where we can:

1) repeat the experiment under essentially unchanged conditions;
2) describe the set of all possible outcomes of the experiment; and
3) see a definite pattern as the experiment is repeated many times.

Outcomes and Sample Space

The possible **outcomes** that we must be able to describe form a **set** of outcomes.

The set of all possible outcomes of an experiment is called the **sample space** for the experiment.

Each possible outcome is a member of the **sample space**. Each outcome is also said to be an **element** or a **sample point** in the sample space.

The **possible outcomes** of the experiment of rolling a die once are that the die comes to rest with the one on top, or the two, or the three, or the four, or the five, or the six. The sample space is the set **{1, 2, 3, 4, 5, 6}**. It contains six sample points. If the experiment is flipping a coin once, the sample space is **{H, T}** (heads or tails).

If we flip a coin three times the sample space is **{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}**. You might want to list the outcomes as **{**three heads, two heads, one head, three tails**}**. However, **HHT** is a different outcome from **HTH** and from **THH**. We may not care about the order of the heads' and tails' occurrence, but they are distinct.

If an experiment involves flipping a coin once and throwing a die once, then for each possible outcome of the coin flip there are six possible outcomes of the die roll. If we represent flipping heads and rolling a 4 as **H4**, and tails and a 6 as **T6**, the sample space is: **{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6}**. Notice that:

The **size (cardinality) of the sample space** can be found using the **fundamental counting principle**. If an experiment consists of several **procedures** (in this case, flipping a coin and tossing a die) and if the procedures are **independent** of each other (i.e., the possible outcomes from the second procedure are not affected by the outcome of the first procedure), then the number of sample points in the sample space is equal to the number of outcomes of the first procedure **times** the number of outcomes of the second **times** the number of outcomes of the third … (and so on).

---

4 Since probability theory concerns only random experiments, we can just refer to them as "experiments."

Example: In the case of flipping a coin three times (or flipping three coins), each flip is a procedure. There are two possible outcomes of the first procedure (flip): the coin can come up heads or tails. For each of these outcomes, there are two possible outcomes of the second procedure. This gives us four possible outcomes of the first two flips. For each of those four, there are two possible outcomes of the third procedure, giving a total of eight possible outcomes resulting from three flips. A tree diagram of possible outcomes (like the tree diagram in the discussion of the fundamental counting principle) may make this clearer. Draw one.

Exercise on the Size of the Sample Space

1. How big is the sample space of an experiment where a coin is flipped 10 times?
2. How big is the sample space of the experiment of throwing three dice?
3. How big is the sample space of the experiment of dealing five cards from a shuffled 52-card deck?[5]
4. A license plate uses three letters and three denary digits. If any combination of letters and digits is permitted, what is the cardinality of the sample space of making license plate registrations by random selection of letters and numbers?

Events

An **event** is **a non-empty subset of the sample space of an experiment**. Thus, an event is a **set** of possible outcomes.

In the experiment where we flip three coins, getting exactly two heads is an event. This event is the subset **{HHT, HTH, THH}** of the sample space. The event of getting **at least** two heads is the subset **{HHH, HHT, HTH, THH}**. The event of getting no heads is **{TTT}**. When rolling a die, throwing an odd number is the event **{1, 3, 5}**.

The empty set is a subset of the sample space, but it is not an event. The power set contains the empty set. So the cardinality of the set of all possible events in the sample space is one less than the cardinality of the power set of the sample space.

If the experiment is throwing one die once, the sample space, as we saw above, is the set $\mathbf{A = \{1, 2, 3, 4, 5, 6\}}$. How many events can one distinguish in this sample space? The number of subsets of **A** is the size of the power set of **A**, or $\mathbf{\#(\mathcal{P}(A))}$. The cardinality of the power set of a six-element set is $\mathbf{2^6 = 64}$. But one of these subsets is the empty set – the "non-event." So the number of subsets of **A** is $\mathbf{2^6 - 1 = 63}$.

---

5 This is tricky, as discussed below.

Exercise on Random Events

1. How many events can one distinguish in the experiment of flipping a coin three times?
2. How many events can one distinguish in the experiment of flipping a coin once and throwing a die once?
3. How many events can one distinguish in the experiment of flipping a coin ten times?

Permutations and Combinations

In the Exercise on the Size of the Sample Space, above, the third question asked the size of the sample space of the experiment of drawing five cards from a shuffled 52-card deck. You can consider this experiment as consisting of five procedures. The first procedure (drawing the first card) will give you any one of the 52 cards. There are 52 possible outcomes of this procedure. The second procedure can have any of 51 possible outcomes (there are only 51 cards left). The third has 50 possible outcomes, the fourth has 49, and the fifth procedure has 48. By the fundamental counting principle, there are $52 \cdot 51 \cdot 50 \cdot 49 \cdot 48$ possible outcomes of the experiment.

One possible outcome is a hand consisting of **K♠, 4♥, 7♣, 2♦,** and **9♣**. Another is **4♥, 2♦, K♠, 9♣,** and **7♣**. For a card-player, these are the same hand. They are just the same five cards in different orders.

If the deck contained only those five cards, what is the sample space for the experiment of drawing five cards? We could get any of the cards as our first card, so there are 5 possible outcomes of the first draw, 4 possible outcomes of the second draw, and so on. The sample space consists of $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ possible outcomes. But every one of them results in the same five cards. They're "all the same." We're counting the number of different **permutations** of five cards.

Conjecture: The number of permutations of **n** distinct things is **n!**.

Proof: The first element in a permutation of **n** distinct things can be selected in any of **n** ways (i.e., any of the things can be first in a permutation). For each of these **n** beginnings, one can select a second element in **(n−1)** ways.[6] So, (by the fundamental

---

6 **(n-1)** because one element has already been used as the first element. This is called drawing without replacement. When a particular card (say **4♦**) is drawn or dealt from a deck, there is no chance of getting another **4♦**. On the other hand, when we select numbers and letters for a license plate, we can get an **A** and then another **A**. The description of the distinction ("drawing with replacement" and "drawing without replacement") comes from experiments like drawing coloured balls from a jar. If you throw each ball back in (replace it) before you draw the next ball, the draws will be independent of each other. If you keep balls out after you've drawn them (drawing without replacement), then the proportions of different colours in the jar may be changed, so that the probabilities for the next drawing change.

counting principle) there are $n \cdot (n-1)$ ways to choose the first two elements. There will be $(n-2)$ possibilities for each of the $n \cdot (n-1)$ ways of choosing two elements, resulting in $n \cdot (n-1) \cdot (n-2)$ ways of choosing three, and so on. There will only be one way to choose the last component (all the other things being already used). There will be $n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$ possible permutations of n things. But this is just $n!$. So there are $n!$ ways to permute n things.■

$52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 = 311875200$ is the number of all the permutations of all the possible sets of five cards that could be drawn. Many of these permutations were just the same set of five cards drawn in different orders. Each set of five cards could be drawn in $5!$ (120) different ways. So the number of different **sets** (hands) of five cards, if we ignore the order in which they are drawn, is only $311875200/120$, or 2,598,960.

To develop the general mathematical principles for dealing with all this, start by looking at a case where we do care about order. Suppose that there are 12 boats in a regatta, and that any boat is equally likely to win. How many different possible ways can we get a first-, second-, and third-place finisher? Any boat might be first, so there are 12 possible first-place outcomes. For each of these possibilities, there are 11 boats that could come in second, so there are $12 \cdot 11$ first- and second-place possibilities or outcomes. For each of these, there are 10 boats that could place third. By the fundamental counting principle, the size of the sample space for this experiment is $12 \cdot 11 \cdot 10$ possibilities. Since we care about the difference between first, second, and third, all of these outcomes are distinct.

In mathematics, this problem is described (somewhat confusingly) as finding the number of permutations of 12 things taken 3 at a time. In general, we calculate the number of permutations of n things taken r at a time, where r is less than or equal to n.

> The number of permutations of n distinct objects taken r at a time, where $r \leqq n$, is denoted by ${}_nP_r$ or $P_{n,r}$ and is given by $P_{n,r} = \frac{n!}{(n-r)!}$.

Going back to the 3-finishers-out-of-12-boats experiment, you can see that $12 \cdot 11 \cdot 10 = \frac{12 \cdot 11 \cdot 10 \cdot (9!)}{9!}$, and that $12 \cdot 11 \cdot 10 \cdot (9!) = 12!$, so $12 \cdot 11 \cdot 10 = \frac{12!}{(12-3)!}$ $= \frac{n!}{(n-r)!}$ where $n = 12$ and $r = 3$, confirming the formula in the box, above.

We proved that the number of permutations of n things taken n at a time (the number of permutations of n things) is $n!$. According to the rule above it is $\frac{n!}{(n-n)!}$. These two claims are consistent only if $(n-n)!$ is equal to 1. That requires that we define $0! = 1$. So that is how we define it.

How many distinct permutations of the letters of the word "black" are there? There are five distinct letters in "black," so there are $5! = 120$ possible permutations.

How many distinct permutations of the letters in "daddy" are there? The difference is that three of the letters in "daddy" are not distinct. We can distinguish them by capitalizing the first "d" and italicizing the third, giving "Dad*d*y," and then there would again be **5!** permutations. "Dad*d*y," "Da*d*dy," "daD*d*y," "da*d*Dy," "*d*aDdy," and "*d*adDy" would be distinct permutations. But if we ignore the different typography, they are all just "**daddy**." There are six typographical variants because there are **3!** or **6** ways of permuting the three "d"s. Of the **5!** permutations of "daddy," there are only $\frac{5!}{3!} = \frac{120}{6} =$ **20** distinguishably different permutations. The rule can be stated generally as:

If **n** objects are made up of n_1 of one kind, n_2 of a second kind, … n_k of a k^{th} kind, such that $n_1+n_2+\ldots+n_k = n$, then the number of **distinguishable permutations** of these **n** objects is given by $\frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_k!}$.

Example: How many ways can you get exactly three heads in ten flips of a fair coin? The answer is the number of permutations of three (indistinguishable) heads and seven (indistinguishable) tails, or $\frac{10!}{3! \cdot 7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 120$ ways. The sample space for the experiment of ten flips contains $2^{10} = 1024$ outcomes. We could guess that the *a priori* probability of throwing exactly three heads in ten tosses is **120/1024** (about **0.117** or **11.7%**). We'd be right, as we shall see.

The selection of **r** objects out of **n** objects without regard to order is called a **combination** of **n** objects taken **r** at a time, and is denoted by ${}_nC_r$ or $C_{n,r}$ or, usually, by $\binom{n}{r}$ (pronounced "**n** choose **r**").

When we found the number of **permutations** of **n** objects taken **r** at a time, above, we noted that this number includes all the permutations of each set of **r** objects. There are **r!** permutations of **r** objects. To find the number of **combinations** of **n** objects taken **r** at a time, we had to divide the number of permutations by **r!**. That's what we did when we divided the number of ways of getting 5 cards from a 52-card deck by the number of permutations (120) of the five cards, above.

The number of combinations of **n** distinct objects taken **r** at a time, where $r \leqq n$, is denoted by $\binom{n}{r}$ and is given by $\binom{n}{r} = \frac{nPr}{r!} = \frac{\frac{n!}{(n-r)!}}{r!} = \frac{n!}{r! \cdot (n-r)!}$.

Example: How many different poker hands can be dealt from a 52-card deck? The answer is $\binom{52}{5}$, the number of combinations of 52 things taken 5 at a time, and we

calculate it as: $\frac{52!}{5!\cdot(52-5)!} = \frac{52!}{5!\cdot 47!}$. **52!** and **47!** are **huge** numbers.[7] It's easier to calculate if you remember how we figured out the formula for $P_{n,r}$. $52! = 52\cdot 51\cdot 50\cdot 49\cdot 48\cdot 47!$, so we can cancel **47!** from the top and bottom of the previous fraction, giving $\frac{52\cdot 51\cdot 50\cdot 49\cdot 48}{5!} = \frac{52\cdot 51\cdot 50\cdot 49\cdot 48}{5\cdot 4\cdot 3\cdot 2\cdot 1} = 2598960$ possible poker hands in a 52-card deck. This is the answer we got on page 174, above.

Example: How many heart flushes are possible? Here the answer is the number of combinations of the 13 hearts in the deck taken 5 at a time, or $\binom{13}{5} = \frac{13!}{5!\cdot(13-5)!} = \frac{13!}{5!\cdot 8!} = 1287$ possible heart flushes. By similar calculations, there are 1,287 spade, club, and diamond flushes. So the number of possible flushes that can be dealt is $4\cdot 1287$, or 5,148. The *a priori* probability of being dealt a flush is $5148/2598960 = 0.00198$, or about **0.2%**. A flush should be dealt about once in 500 hands of poker.

Example: How many ways can one get a full house? A full house is three cards of one rank and two of another (e.g., three Jacks and two sevens). First, calculate the number of ways of getting three of some particular rank (e.g., three Jacks). This is the number of combinations of four things (Jacks) taken three at a time, which is $\binom{4}{3} = \frac{4!}{3!\cdot(4-3)!} = \frac{4!}{3!\cdot 1!} = \frac{24}{6} = 4$. There are thirteen ranks, and we don't care whether the three cards are Jacks or some other rank. There are $13\cdot 4 = 52$ ways of getting three of the four cards of any rank. We can get two (of the four) sevens in $\binom{4}{2} = \frac{4!}{2!\cdot(4-2)!} = \frac{4!}{2!\cdot 2!} = \frac{24}{4} = 6$ ways. Since there are 12 ranks remaining after we have our three-of-a-kind, there are $12\cdot 6 = 72$ ways of getting a pair to go with it. By the fundamental counting principle, there are $52\cdot 72 = 3744$ ways of getting three cards of one rank and a pair of another rank. The *a priori* probability of being dealt a full house is $3744/2598960 = 0.00144$, or about **0.14%**.

Exercise on Permutations and Combinations

1. In the experiment of picking 4 letters from the English alphabet at random without replacement and lining them up in the order in which they were picked, how many different 4-letter "words" are possible? (This is the number of permutations of 26 things taken 4 at a time.)

---

7 **52!** is 80,658,175,170,943,878,571,660,636,856,403,766,975,289,505,440,883,277,824, 000,000,000,000, or about 8.0658×10^{67}.

2. If order doesn't matter in question 1, how many different 4-letter combinations are possible? What is the probability that the first 4 letters you pick will (after suitable rearrangement) spell either "LOVE" or "HATE" or "MATH"?
3. How many ways can 3 girls and 4 boys sit in a row if the girls and boys must alternate?
4. Calculate the number of distinguishable permutations of the letters in "MISSISSIPPI."
5. Three couples have six seats in a row at a ball game.
   (a) In how many ways can the six people be seated in the six seats?
   (b) In how many ways can they be seated if the couples sit together?
   (c) In how many ways can they be seated if all the men are to sit together?
6. A family of 5 has tickets to a show. (a) In how many different ways can they be seated in their 5 reserved seats? (b) In how many ways can they be seated if the father has to be in the middle seat?
7. A student has to write reviews of five books from a reading list of twelve books. Calculate the number of ways she can select five books.
8. A carton of a dozen eggs contains two that are bad. In how many ways can one choose three eggs so that
   (a) None of the bad eggs is selected?
   (b) Exactly one bad egg is selected?
9. The table of numbers below is the beginning of **Pascal's triangle**. The actual table is only the boldface numbers. The italic numbers down the left side just number the rows.

Row 0												**1**									
Row 1											**1**		**1**								
2										**1**		**2**		**1**							
3									**1**		**3**		**3**		**1**						
4								**1**		**4**		**6**		**4**		**1**					
5							**1**		**5**		**10**		**10**		**5**		**1**				
6						**1**		**6**		**15**		**20**		**15**		**6**		**1**			
7					**1**		**7**		**21**		**35**		**35**		**21**		**7**		**1**		
8				**1**		**8**		**28**		**56**		**70**		**56**		**28**		**8**		**1**	
9			**1**		**9**		**36**		**84**		**126**		**126**		**84**		**36**		**9**		**1**
Row 10	**1**		**10**		**45**		**120**		**210**		**252**		**210**		**45**		**45**		**10**		**1**

Each row begins with a **1** and ends with a **1**. Every other entry is the sum of the nearest two numbers in the previous row. Construct one of your own to see how this works. The second number in each row (except row 0) is the row number of that row. This triangle of numbers has many curious properties, and there are whole books written about it. The interesting thing for us now is this. If we call the first entry in a row entry **0** of that row and use the row numbers indicated

above, the value of $\binom{n}{r}$ can be found by looking at entry **r** of row **n**. Thus (looking at row 4), we see that $\binom{4}{2} = 6$ and $\binom{4}{3} = 4$. Going to row 8, we see that $\binom{8}{2} = 28$, $\binom{8}{3} = 56$, and $\binom{8}{4} = 70$. Notice that the triangle is symmetrical, so that $\binom{8}{3} = \binom{8}{5}$ and so on. That makes sense, since the number of ways to choose 3 things out of 8 is the same as the number of ways to choose the 5 that are left out.

Use Pascal's triangle (extending it if necessary) to solve the following problems:

(a) How many ways can you get 3 heads in 7 tosses of a fair coin?

(b) A student committee must have three first-year and four second-year students. Nine first-year and seven second-year students volunteer. In how many ways can the members of the committee be selected from among the volunteers?

Postulates of Probability Theory

P1: The probability of any event must be a number between **0** and **1**, where **0** indicates impossibility and **1** indicates certainty.

P2: If we assign a probability to every possible outcome of a random experiment, the sum of these probabilities must be **1**.

P3: **If all outcomes of an experiment have equal probability**, then the probability of an event **E**, symbolized by **P(E)**, can be calculated as:
P(E)=(number of outcomes favorable to **E)÷(**total number of possible outcomes**)**. Since the total number of possible outcomes is the size (cardinality) of the sample space **S**, and that the number of outcomes favourable to **E** is the size of a set of outcomes which is a subset of the sample space, this formula becomes: **P(E) = #(E)/#(S)**.

From **P1**, **P2** and **P3** we can deduce that the probability that an event **E** will **not** occur (symbolized by **P(E′)**) is given by **P(E′) = 1−P(E)**.

If the probability of throwing three heads in three flips of a coin is **0.125**, then the probability of **not** throwing three heads is **1− 0.125 = 0.875**. If **A** and **B** are events in (i.e., subsets of) a sample space **S**, then **A ∪ B** is the event that **A** or **B** or both occur, and **A ∩ B** is the event that both **A** and **B** occur. We saw in Chapter 7,
#(A ∪ B) = #(A)+#(B)−#(A ∩ B). It follows that:

$$P(A \cup B) = \frac{\#(A \cup B)}{\#(S)} = \frac{\#(A)+\#(B)-\#(A \cap B)}{\#(S)} = \frac{\#(A)}{\#(S)} + \frac{\#(B)}{\#(S)} - \frac{\#(A \cap B)}{\#(S)} =$$

P(A) + P(B) − P(A ∩ B). This leads us to the **addition rule for probabilities**.

P4: (**The addition rule for probabilities**). If **A** and **B** are events[8] in a sample space **S**, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Example: Find the probability of throwing at least one head in three flips of a fair coin. "Throwing at least one head" means throwing one head or two heads or three heads. There are three ways to throw one head (**HTT**, **THT**, or **TTH**), so the probability of one head $P(1) = 3/8 = 0.375$. There are three ways to throw two heads, so $P(2) = 3/8 = 0.375$. There is only one way to throw three heads, so $P(3) = 1/8 = 0.125$. There is no way to throw one **and** two heads or one **and** three heads, so $P(1 \cap 2 \cap 3) = 0$. So $P(1 \cup 2 \cup 3) = P(1)+P(2)+P(3)-0 = 0.375+0.375+0.125-0 = 0.875$.

Example: Another way to solve the previous example is to notice that the probability of **not** throwing at least one head is the probability of throwing three tails. The probability of three tails is $1/8 = 0.125$. The probability of throwing at least one head is $1 - 0.125 = 0.875$ (using $P(E') = 1 - P(E)$).

We didn't need to calculate the number of combinations in the previous examples. We just wrote out all the combinations. Consider an example where it would be too tedious to write out all the ways of getting **r** heads in **n** flips.

Example: Find the probability of throwing 4, 5, or 6 heads in 10 flips of a fair coin. There are $\binom{10}{4} = \frac{10!}{4! \cdot 6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 210$ ways to get 4 heads, and $\binom{10}{5} = \frac{10!}{5! \cdot 5!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 252$ ways to get 5 heads, and $\binom{10}{6} = \frac{10!}{6! \cdot 4!} = 210$ ways to get 6 heads. $P(4 \cap 5 \cap 6) = 0$ so, by the addition rule, the number of ways of getting 4 or 5 or 6 heads is $210+252+210-0 = 672$. The size of the sample space is $2^{10} = 1024$, so the probability of getting 4 or 5 or 6 heads is $672/1024 = 0.65625$ or about **65.6%**.

Example: Find the probability that we get an even number or a number greater than 2 in one throw of a fair die. The event of throwing an even number is the set $E = \{2, 4, 6\}$ and the event of throwing a number greater than 2 is $A = \{3, 4, 5, 6\}$. The probability $P(E)$ of throwing an even number is $3/6$, and the probability $P(A)$ of throwing a number greater than 2 is $4/6$. If we just added these probabilities, we'd get an answer of $7/6$. Probabilities must be between 0 and 1 (inclusive), so $7/6$ must be wrong. Notice that the set $E \cap A$ is not empty. $E \cap A = \{4, 6\}$, and its probability (the probability of throwing a number which is both even and greater than 2) is $2/6$. Correct use of the addition rule gives the probability that we get an even number or a number greater than 2 (or both) in one throw of a fair die as $3/6+4/6-2/6 = 5/6$. Of course, the probability of throwing a number which is not either even or greater than 2

---

8 Events, not outcomes!

is just the probability of throwing a 1, which is **1/6**, so the probability of an even number or a number greater than 2 is **1−1/6 = 5/6**.

P5: (**The multiplication rule for probabilities**) If **A** and **B** are independent events, then the probability that both **A** and **B** occur **P(A ∩ B) = P(A)·P(B)**.

This postulate of probability theory introduces the concept of **independent events**. This concept is fundamental to the notion of **conditional probability**, which causes a lot of confusion.

The idea of two events being **independent** is that the occurrence of one of them does not alter the probability of the other occurring. Flipping a coin and getting heads does not affect the probability of getting heads on the next flip of the same coin. So (by Postulate P5) the probability of flipping two heads in a row is the product of the probability of flipping a head times the probability of flipping a head, or **0.5·0.5 = 0.25**. This agrees with what we find if we use the cardinality of the event of getting two heads divided by the cardinality of the sample space of flipping a coin twice.

On the other hand, drawing cards from a deck[9] does change the probabilities for the next card drawn. Once you draw the King of clubs, for example, the probability that the next card will be the King of clubs is **0** (won't happen). The probability that the next card will be a King is only **3/51**, whereas the probability of getting a King as the first card is **4/52**. The probability that the next card will be a club is **12/51**, whereas if you had not already drawn a club, the probability that the next card will be a club is **13/51**. Successive draws (without replacement) are not independent.

The multiplication rule for probabilities is useful for two purposes. If we know (or have good reason to believe) that two events are independent, we can use it to calculate the probability that both events will happen. On the other hand, we can use the multiplication rule **to decide whether two events are independent**. This second use of the rule is important when we are dealing with empirical probabilities (i.e., when we use observed relative frequencies to decide on probabilities).

An example can be taken from the last Quebec sovereignty referendum. About 50% of the voting population voted "yes" and about 50% voted "no." (The relative frequency of "no" is about **0.5**.) If we randomly selected a Quebec voter, the probability that he or she is a supporter of the "no" would be about 50%. Let's say that about **1/7** of Quebec voters are non-francophone. Then the probability that a randomly selected voter is non-francophone is about **1/7** (i.e., non-francophones occur in the voting population with a relative frequency of about **0.14**). What is the probability that a randomly selected Quebec voter is non-francophone and supported the "no"? If these

---

9 We are assuming drawing without replacement in this example.

two events (being non-francophone and voting "no") are independent, the probability (by the multiplication rule) is **0.5** times **0.14**, or about **0.07**.

Now suppose that we conducted a survey of 1,000 Quebec voters selected by some random process (so that there was an equal chance of any 1,000 voters being chosen). Let's say that we found that 148 of the thousand were non-francophone (about what we'd expect from the probabilities), and that 503 of the thousand voted "no" (again, just about what we'd expect). But then we found that 132 of the 148 non-francophones voted "no." That would indicate that voting "no" was not independent of whether a voter is francophone. The calculated probabilities (above) would lead us to expect only about 70 of the non-francophones to vote "no." Based on relative frequencies in our survey, we'd estimate the probability of being non-francophone and voting "no" to be about **0.132**, almost twice as high as the calculated **0.07**. Here we used the multiplication rule to show that two events are not independent. The way a person voted is (at least partly) dependent on whether or not the person is francophone.

Exercise on Probabilities

1. The experiment is tossing 4 coins. Use the fundamental counting principle to find the number of points in the sample space. List the sample space (set). For the events: (a) exactly 3 heads come up; (b) 3 or more heads come up, define the event-set and calculate the probability of the event.
2. The last 20 birds that fed at my feeder were 15 sparrows, 3 blue jays, and 2 red-breasted nuthatches. Use this information to determine the empirical probability that the next bird to feed at the feeder will be (a) a sparrow; (b) a blue jay; (c) a nuthatch; (d) a falcon.
3. You find an irregularly shaped rock with 5 flat faces and label the faces with numbers from 1 through 5. You toss the rock 100 times and get the following frequencies: 1 - 32 times; 2 - 18 times; 3 - 15 times; 4 - 13 times; 5 - 22 times. Find the empirical probability that the next toss of the rock will be: (a) 4; (b) 2; (c) anything but a 1.
4. One card is selected at random from a normal deck of 52 playing cards. What is the *a priori* probability that the card will be: (a) a 3; (b) a spade; (c) not a 3; (d) a spade and a diamond; (e) a Jack or a Queen or a King; (f) a card greater than 5 and less than 10.
5. On a multiple-choice test with four possible answers for each question, what is the probability that a random guess will be the right answer to one particular question?
6. A traffic light is red for 30 sec., yellow for 5 sec., and green for 40 sec. What is the probability that the light will be green when you reach it?
7. There are seven coloured balls in a jar – 4 red, 2 white, and 1 blue. (a) Use the fundamental counting principle to determine the number of sample points in the

sample space for the experiment of drawing two balls at random from the jar with replacement (meaning that you put the ball back before drawing the next one). (b) List the outcomes in the sample space as a set. (c) Find the probability that 2 red balls are selected. (d) Find the probability that 1 red and 1 blue ball are selected. (e) Find the probability that 2 blue balls are selected.

8. Repeat Ex. 8 without replacement (i.e., you draw a ball and keep it out while you draw the next one).
9. (a) Use the fundamental counting principle to determine the number of sample points in the sample space for the experiment of throwing two dice. (b) Give the denotative definition of the sample space. For each of the events (c) a pair; (d) a 7, (e) a 12, give the denotative definition of the event-set and calculate the probability of the event.
10. A student who knows absolutely nothing about the subject being tested takes a true-false quiz of 10 questions. What is the probability that the student (a) answers every question correctly? (b) answers exactly 1 question correctly?
11. 3 girls and 4 boys are placed randomly in a row of 7 seats. What is the probability that (a) the girls and boys sit in alternate seats? (b) the 3 girls sit together?
12. Given that $\mathbf{P(A) = 2/5}$, $\mathbf{P(A \cup B) = 3/4}$, and $\mathbf{P(A \cap B) = 1/10}$, find $\mathbf{P(B)}$.
13. Of 10 balls in a jar, 3 are red. You select 3 balls at random, without replacement. What is the probability of getting (a) 3 red balls? (b) at most 1 red ball?
14. The experiment is flipping a coin and then drawing 1 card at random from a shuffled deck. What is the probability of (a) getting either a head or a red card? (b) getting both a head and a diamond?
15. What is the probability of getting 4 Queens in a randomly dealt 5-card hand?

Conditional Probability

When two events **A** and **B** are dependent we have to consider **conditional probabilities**. Suppose the probability that it will snow on any randomly chosen day of the year is 10%. Clearly, the probability that it will snow on a randomly selected day will be lower, given that the day is in August, than if the day happened to be February. The probability that it will snow is **conditional** on the season.

Here is an example involving empirical probabilities. Imagine that we conduct a survey of 1000 auto mechanics in our area. We check whether they are factory-trained or not, and give them a specially prepared test car to diagnose and repair. Those who spot the actual problem and fix it get a pass, and those who don't, fail. We describe our findings in a table, as:

	Pass	*Fail*	*Total*
Factory-trained	**322**	**107**	**429**
Not factory-trained	**161**	**410**	**571**
Total	**483**	**517**	**1000**

If you chose a mechanic at random, the empirical probability (rounded to two figures) that she would have passed the test is **P(G) = 483/1000 = 48%** or **0.48**. The probability that she was factory-trained is **P(F) = 429/1000 = 43% or 0.43**. Your chance of getting a good[10] mechanic purely at random is less than **50%**. But if you limit your choices to just factory-trained mechanics, you improve your chances. The probability that a factory-trained mechanic would pass the test is **322/429 = 75%**. That is, the probability of getting a good mechanic, given that it is a factory-trained mechanic, is **75%**. This is conditional probability. The probability of getting a good mechanic is **conditional** on whether the mechanic is factory-trained or not.

Going back to the card-dealing experiment, above, we can ask "What is the probability that the first two cards dealt will be Kings?" The probability that the first card dealt is a King is **1/13**. If the second card being a King were independent of whether or not the first one was, the probability of the second card being a King is, again, **1/13**. But it is not. If the first card dealt is a King, then the probability that the second one will be a King is only **3/51** or **1/17**. If the first card is not a King, the probability that the second one will be a King is **4/51**. So the probability of the second card being a King is dependent, or **conditional** on whether the first was a King.

We need a formula to calculate the probability that an event **B** occurs, **given that some other event A** occurs. We start with this definition:

> The probability of event **B** occurring, given that an event **A** has happened or will happen (the time relationship does not matter) is called the **conditional probability of B given A** and is written **P(B|A)**.[11]

The number of ways of getting two cards from a 52-card deck is $\binom{52}{2} = 1326$. The number of ways of getting 2 Kings is $\binom{4}{2} = 6$. So the probability of being dealt 2 Kings in 2 cards is **6/1326**. If **A** is the event where the first card dealt is a King and **B** is the event that the second card dealt is a King, then **P(A ∩ B) = 0.00452**. As we saw, above, **P(A)** is **1/13** and the probability that the next card dealt is a King, given that I already have a King is **P(B|A) = 1/17**. Do the math to see that:

> The probability of **B** given **A** is equal to the probability that both **A** and **B** occur **P(A ∩ B)** divided by the probability that **A** occurs **P(A)**, or $P(B|A) = \frac{P(A \cap B)}{P(A)}$.

---

10 If we define "good" as "likely to pass the test."

11 I know that we used "|" to represent "divides," earlier. In probability theory, it is common to use it to symbolize conditional probability, while in number theory it is the conventional symbol for "divides." Sorry about that.

Example : Use the rule above to calculate the probability that a mechanic is good given that she is factory-trained (i.e., calculate **P(G|F)**). **P(F ∩ G) = 322/1000** (that is the probability that the mechanic both passed the test and was factory-trained – the top-left entry in our table). **P(F) = 429/1000** as we saw. So **P(G|F) = 0.322/0.429 = 75%**. Test your understanding by calculating the probability that a mechanic is factory-trained given that she passed the test, and the probability that she passed the test given that she was not factory-trained.

Example: Calculate the probability that a single fair die comes up 2, given that the result is an even number. The probability that the die shows both an even number and 2 is just the probability that it shows a 2, which is **1/6**. The probability that the result is an even number is **1/2**. So the probability that the result is a 2 given that it is even is **(1/6)/(1/2)** or **1/3**.

Example: Assume that the probability that a baby will be a girl is **0.50**. In a family with two children, find the probability that the family has (a) two girls; (b) two girls, given that at least one of the children is a girl; (c) two girls, given that the older child is a girl. **Answers**: (a) The sample space of two children is **{BB, BG, GB, GG}** (where **B** is a boy and **G** is a girl). Since there are four equiprobable possibilities, only one of which is two girls, the probability that a two-child family has two girls is **1/4**, or **0.25**. (b) When one of the children is a girl, the sample space is just **{BG, GB, GG}**. One of the three possibilities is two girls, so the probability of two girls is **1/3** or **0.33**. (c) If the older child is a girl, the sample space reduces to **{GB, GG}**, and the probability that the family has two girls is **1/2** or **0.50**.

This example relates to one in Paulos' *Beyond Numeracy*[12]:

> It's known that in a certain curiously "normal" 1950s neighborhood every home houses a family of four—mother, father, and two children. One picks a house at random, rings the bell, and is greeted by a girl. (We assume that in the 1950s a girl, *if* there is at least one, will always answer the door.) Given these assumptions, what is the conditional probability that this family has both a son and a daughter? The perhaps surprising answer is not 1/2, but 2/3. There are three equally likely possibilities—older boy, younger girl; older girl, younger boy; older girl, younger girl—and in two of them the family has a son. The fourth possibility—older boy, younger boy—is ruled out by the fact that a girl answered the door.

Example: Suppose there is a disease that affects 0.1% of the population. A blood test is discovered which is 99% accurate. That means that the test will be wrong 1% of the time – either giving a false negative (a person tests negative even though she has the disease) or a false positive (a person tests positive even though he does not have the disease). Imagine you test positive for the disease. How much should you worry?

If 100,000 people take the test, about 100 of them will have the disease, and about 99 of those will test positive. Of the 99,900 people who don't have the disease, about

---

12 John Allen Paulos, *Beyond Numeracy: Confessions of a Numbers Man* (New York: Alfred A. Knopf, 1991), p. 190.

999 of them will test positive (false positives). So, out of a total of 1098 positive tests, most (999) are false positives. Thus the conditional probability of having the disease, given that one tests positive, is only **99/1098**, a bit over 9%! The fact that the test is 99% reliable does not mean that there is a 99% chance that you have the disease.

Exercise on Conditional Probability

1. Suppose **E** is the event that a job-applicant has experience, **C** is the event that she has a car, and **G** is the event that she is a college graduate. State in words what probabilities are expressed by:
   (a) $P(C|G)$ (b) $P(E|C')$ (c) $P(C'|E)$
   (d) $P(G'|C')$ (e) $P(C|(E \cup G))$ (f) $P((E \cap C')|G)$.
2. A single card is drawn from a deck. (a) What is the probability that it is a club, given that it is black? (b) What is the probability that it is black, given that it is a club?
3. One hundred people are surveyed to find which TV news channel they watch. The results are:

	CBC	CTV	CNN	Other	Total
Men	30	20	40	25	115
Women	50	10	20	15	95
Total	80	30	60	40	210

   If one of these people is selected at random, what is the probability he or she watches:
   (a) CBC or CTV
   (b) CBC given that the individual is a woman
   (c) CBC or CTV given that the individual is a man
   (d) A station other than CNN given the individual is a woman
   (e) CBC, CTV, or CNN, given that the individual is a man

13 Practical Probability

Empirical Probability

The relative frequency interpretation of probability defines probability in terms of relative frequencies.

> The **empirical probability of an outcome or event** is defined as the relative frequency of that outcome or event in many repetitions of a random experiment.

If the experiment is truly random, there will be a pattern in the outcomes (and thus in the events) over the long run. As the number of repetitions gets bigger and bigger, the relative frequency of any event will get closer and closer to some particular number. How many repetitions are "a long run"? Since it is always possible that some unusual sequence can occur the number will have to be very large.

Some of the factors involved in determining the sex of a baby have only been discovered in the last century. Yet people have known for millennia that the probability that a human child will be male is just slightly more than 50%.

It may be impossible to predict with accuracy that any particular 21-year-old man will die this year, but many years of mortality figures tell us that about 0.18% of all the 21-year-olds in America die each year.

We discover empirical probabilities by tallying frequencies of events over a long sequence of repetitions. We take the observed relative frequency of an event as an approximation to the probability that the event will happen. We can then use the mathematics of probability theory to arrive at interesting conclusions.

Simulation

The mathematics of probability becomes formidable when we deal with complex phenomena involving many possible outcomes and a lot of interlocking dependencies. We can avoid some of the difficult math by simulating the experiment we are interested in. A simulation is like a model of some part of the real world. Events in the model are easily repeatable, and the model omits many of the complexities of the real world. We try to design our model world to reflect the real phenomena we are trying

to simulate. The test that a model is correct is whether its behaviour adequately resembles the behaviour of the phenomenon in the real world.

Most simulations use random numbers as the basis of the simulation. There are books containing tables of random digits. Every individual digit should occur in the table about 10% of the time. Every pair (from 00 to 99) should occur about 1% of the time. Every possible triple of digits (from 000 to 999) should occur with a relative frequency of about 0.1%. And so on. Yet it should not be possible to predict what the next digit will be.

Probability Model

The first step in simulation is to build a probability model of the phenomenon we are interested in.

> A **probability model** is a list of all of the possible outcomes of an experiment, where every outcome is assigned a number, which represents its probability.

Any legitimate probability model must satisfy postulates **P1** and **P2** of probability theory. That is:

> **P1**: The probability of any event must be a number between **0** and **1**, where **0** indicates impossibility and **1** indicates certainty.
> **P2:** We assign a probability to every possible outcome of a random experiment, and the sum of these probabilities must be **1.**

Probability Distributions

The description of a probability model is called a **probability distribution**.

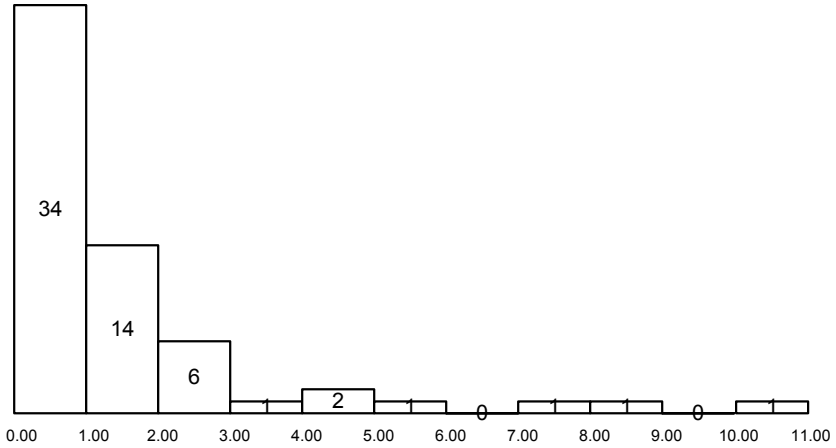
I asked 61 students how much change they were carrying. 8 students had that none. Two students each had \$0.02, \$0.04, \$0.05, \$0.10, \$0.40, \$0.75, \$1.00, \$1.50 or \$1.70. Three each had \$0.40, \$0.65, or \$2.50. The rest had different amounts. The most any student had was \$10.72. A bar chart where each bar was an amount between \$0.00 and \$10.72 would have 1073 bars. Most would have a height of 0 (no student was carrying that amount) and most of the others would show 1.

A better way to graph our results is to **consolidate** the data. We establish how many vertical bars we want to use to represent the data. We might choose, say, 11 bars. The first bar would represent an amount of money between \$0.00 and \$0.99. We describe this as "**0.00** $\leq$ amount $<$ **1.00,**" which means "amounts greater than or equal to 0.00 and less than 1.00." We are dividing the range of data (from 0.00 to 10.72) into **classes** of equal width. We then count how many students are in each class. We have to be sure that our classes are defined so that every amount fits into one and only one

class. There must be no ambiguity. Our classes and their frequencies (numbers of students) would be:

Class	Freq	Class	Freq
0.00 ≤ amount < **1.00**	34	**6.00** ≤ amount < **7.00**	0
1.00 ≤ amount < **2.00**	14	**7.00** ≤ amount < **8.00**	1
2.00 ≤ amount < **3.00**	6	**8.00** ≤ amount < **9.00**	1
3.00 ≤ amount < **4.00**	1	**9.00** ≤ amount < **10.00**	0
4.00 ≤ amount < **5.00**	2	**10.00** ≤ amount < **11.00**	1
5.00 ≤ amount < **6.00**	1		

We graph these data with a **frequency histogram**. In a frequency histogram (unlike a bar chart) the bars touch each other. The width of a bar represents the size of a range of values. The height of the bar represents the number (frequency) of instances of values in that range. The bars in a frequency histogram should have equal widths if the classes cover equal ranges of values. We label the edges of the bars (the limits of the classes), whereas in a bar chart we label the bar.



A **relative frequency histogram** looks just like a frequency histogram, except that the heights of the bars depend on the **relative** frequency, rather than on the **absolute** frequency. In a relative frequency histogram of the above data, the first bar would be 34/61 units high.

Since empirical probabilities are the same thing as relative frequencies,[1] a **probability distribution histogram** is identical to a relative frequency histogram.

---

1 Remember this!

Exercise on Probability Models

1. A package of 72 coloured candies contains 21 brown, 15 red, 17 yellow, 8 green, 7 orange, and the rest are blue. How many are blue? What is the relative frequency of each colour? Draw the probability distribution histogram.
2. If the manufacturer of the candies makes 30% brown candies, 20% red, 20% yellow, 10% green, 10% orange, and the rest blue, what percentage of the candies produced are blue?
3. In the long run the relative frequencies of candies in the boxes will approach the relative frequencies of the various colours manufactured. Using the proportions indicated in question 2, build a probability model for the experiment of taking one candy at random from any box.
4. Suppose that Canadian census data on all Canadian women between 20 and 29 years old indicates that 28.8% of them are single, 0.3% are widowed, 7.6% are divorced, and the rest are married. Construct a probability model for the experiment of selecting a Canadian woman at random where the outcome is her marital status. Draw the histogram showing the probability distribution.

Simulation in General

To find the probability of an event by simulation, you:

1. Build a probability model for the random experiment. Assign probabilities to individual outcomes (assuming independence where appropriate).
2. Decide how to simulate the basic outcomes of the experiment. You might decide to use dice, coins, random digits, etc. Decide how your simulation will represent each outcome of the real-world experiment.
3. Decide how to simulate an event in the experiment by combining simulations of basic outcomes from step 2.
4. Estimate the probability of an event by the relative frequency of the event in many repetitions of the simulation.

Imagine a community where women are highly valued. Every family decides to keep having babies until a girl child is born. What might this do to the population growth rate? How would you figure out what the likely family size will be if every family keeps having kids until a girl is born, and then stops? Is it possible that there may be too many women in the resulting population?

Birth records show that, of 1210 babies born in that community, 590 were boys. The ratio of boy and girl birth rates in that sample is 48.8% boys to 51.2% girls. Assume that the ratio in the long run will be about 50-50, so the probability of having a boy baby is 0.5, and the probability of a girl is 0.5. Assume that the sex of a baby is independent of the sex of the previous babies born to the same family.

Using this probability model, we have to simulate the experiment of one family having babies until they stop. The basic outcomes are (1) having a girl baby and (2) having a boy baby. An event will be a finished family of **b** boys and **g** girls (where **b** and **g** are natural numbers). We simulate an event by getting random digits until an even digit (which represents having a girl baby) comes up. We can decide that any even digit (0, 2, 4, 6, or 8) will represent a girl baby. Odd digits will represent a boy.

A table of random digits consists of many numbered lines. Before looking at the digits on any particular line, I decide that I will start at line number 110. Line 110 looks like this:

110 38448 48789 18338 24697 39364 42006 76688 08708

The first digit (3) represents a boy baby. Since it was not a girl, the family has another kid. The next digit (8) represents a girl, so the family stops having kids. The event is a boy and a girl, or **BG**. The proportion of boys to girls is 1-1, and the family has only two kids, so the population is not growing.

Probabilities are relative frequencies in a large number of repetitions of an experiment. We have to simulate many repetitions. Continuing with the next random digits, we find that a second family has a girl (4) right away. So do the third (4) and the fourth (8) and the fifth (4) and the sixth (8). At this point our "families" have had 1 boy and 6 girls. It looks bad for population balance! The seventh family has a boy (7) and then a girl (8). The eighth has two boys (9 and 1) and then a girl (8). The ninth has two boys and a girl. The next three (families 10, 11, and 12) have one girl each and stop. Then comes a family that has (9,7,3,9,3,6) five boys before stopping with a girl.

The events we observed in the simulation were **G** (8 times), **BG** (twice), **BBG** (twice) and **BBBBBG** (once). Out of 13 simulated "families" we have:

Number of kids	1	2	3	4	5	6
Frequency	8	2	2	0	0	1
Relative frequency	0.62	0.15	0.15	0	0	0.08

We might infer that there is about a 77% (62%+15%) probability that families will have two kids or less. There appears to be only about a 23% probability that a family will have more than two kids (causing population growth). If our probability estimates from this simulation were reliable, we should expect that out of 100 couples, 62 of them would have one kid, 15 would have two, 15 would have three, and 8 would have six. That means that we would expect 100 couples to have a total of $\mathbf{62 \cdot 1 + 15 \cdot 2 + 15 \cdot 3 + 8 \cdot 6 = 185}$ children. If 200 parents have only 185 kids, the population will decrease.

As to the sex ratio, our probability estimates suggest that 62% of families will have one girl and no boys, 15% will have one and one, 15% will have one girl and two boys, and 8% will have five boys and a girl. Out of 100 families, this would mean there were **100** girls and $\mathbf{15 + 30 + 40 = 85}$ boys. It looks like there will be too many women.

However, our simulation was very small. We only repeated the random phenomenon 13 times. This is hardly a "long run" from which to obtain relative frequencies that are reliable indications of probabilities. We should try the simulated experiment hundreds or even thousands of times.

In this (relatively simple) case, we can use *a priori* methods to calculate the result. We don't have to do a simulation. But the calculations are too difficult for this course.[2] In reality the probability that parents will try for another kid depends on the number of kids they already have. This would not be too difficult to simulate, but much more difficult to calculate *a priori*.

Computer Simulation

Picking random numbers out of a table and counting the number of times a simulated event occurs can get very tedious. Humans also miscount in ways that can seriously affect results. Computers reduce the tedium, improve accuracy, and permit us to make really large numbers of repetitions.

In a computer program, we specify the probabilities of a number of simple outcomes or events. We then have the program generate "events" whose probability depends on the probabilities of the simple outcomes. The program can repeat an operation thousands or even millions of times at high speed.

Our computer program uses a procedure called a "Random Number Generator" or "RNG." A RNG is a sub-program designed to generate "pseudo-random numbers." The program is deterministic; the numbers are predictable if we know the computer's starting number and know the algorithm by which it makes each new number from the previous one. However, a well-designed RNG generates numbers that can pass almost any test of randomness.

The following table gives relative frequencies from two runs of a simulation of the family-planning probability model, above. Two runs of 100 families were simulated.

Kids	1	2	3	4	5	6	7	8	9	>9
Run A	0.48	0.29	0.13	0.06	0.02	0.02	0	0	0	0
Run B	0.43	0.25	0.14	0.11	0.03	0.03	0	0	0	0.01
Theory	0.50	0.25	0.13	0.06	0.03	0.02	0.01	0	0	0

Many of the estimates were not too bad, but some (e.g., 0.43 for the probability of one child and 0.11 for the probability of four children in Run B) were pretty far off. When I ran the same program to simulate 10,000 families (a much longer run), I got better results, as:

---

2 The analytical calculation shows that there will be an average of two kids per family and exactly half of them will be girls if our probability model is correct.

Kids	1	2	3	4	5	6	7	8	9	>9
Run A	0.4946	0.2471	0.1291	0.0608	0.0294	0.0141	0.0070	0.0041	0.0020	0.0018
Run B	0.4958	0.2550	0.1265	0.0637	0.0293	0.0145	0.0086	0.0032	0.0012	0.0022
Theory	0.5000	0.2500	0.1250	0.0625	0.0313	0.0156	0.0078	0.0039	0.0020	0.0019

When I tried a run involving 100,000 simulated families, the estimates improved slightly. There's a kind of "law of diminishing returns" in longer runs.

Kids	1	2	3	4	5	6	7	8	9	>9
Run A	0.5011	0.2491	0.1249	0.0634	0.0303	0.0158	0.0082	0.0037	0.0018	0.0017
Run B	0.4986	0.2496	0.1260	0.0639	0.0303	0.0156	0.0082	0.0039	0.0020	0.0019
Theory	0.5000	0.2500	0.1250	0.0625	0.0313	0.0156	0.0078	0.0039	0.0020	0.0019

Suppose the computer program's RNG gives numbers between 0.0000… and 0.9999…. Any number less than 0.5000… can indicate a girl baby and any number between 0.5000… and 0.9999… (inclusive) can indicate a boy baby. A good RNG will provide about equal numbers of boys and girls in the (very) long run.

If we want to simulate something whose probability is other than 50%, we just pick a different subrange of the numbers that the RNG can develop. Suppose a basketball player regularly makes 70% of her free throws. What is the probability that she will miss three out of five free throws? That she will hit nine out of ten?

Let a random number less than 0.7000… represent a hit and any other number stand for a miss. We try a few thousand runs consisting of five "shots" each and count how many times there were three or more misses. I tried 10,000 sequences of five free throws. The computer missed three out of five shots in 1598 sequences, or about 16% of the time. There is about one chance in six that a steady 70% shooter will shoot as poorly as 40%. I tried 10,000 sequences of ten free throws. The computer got nine or ten hits (out of ten tries) about 15% of the time. One would expect a 70% shooter to get nine or ten out of ten about 1/7 of the time.

Exercise on Simulation

1. Sociologists have studied how children do or do not move out of their parents' occupational class. The overall result can be expressed in terms of probabilities based on relative frequencies. The relative frequencies of the occupational classes of adult sons of white-collar fathers are: 20% professional; 50% white collar; 20% blue collar; 10% no steady job.
   (a) Explain in detail how to simulate the occupational class of a randomly selected son of a white-collar father.
   (b) How would you use your simulation to answer the question "Given five randomly-selected sons of white-collar fathers, what is the probability that at least two of them will wind up in the professional class?"
2. Explain in detail how to simulate the birthday probability. The birthday probability refers to the fact that out of any randomly selected group of 23 or more people,

there is a better-than-even chance that at least two of them celebrate their birthday on the same day.

Probability and Odds

The chance of an event is often stated in terms of odds, rather than probability. We can translate a statement about the odds of some event happening to a statement of the probability of the event, and *vice versa*.

The **odds** that an event will occur are given by the probability that the event will occur divided by the probability that the event will not occur. Therefore the odds in favour of an event **E** are calculated as **P(E)/P(E′)**. Since **P(E′) = 1−P(E)**, the odds in favour of **E** are **P(E)/(1−P(E))**.

We express the odds in terms of integers, saying things like "the odds are 2 to 1 against" or "the odds are 1 to 6 in favour." In the expression "the odds are **A** to **B** in favour of event **E**" we calculate **A** as **P(E)** and **B** as **1−P(E)** and multiply both probabilities by a number that will turn them both into integers. For example, in throwing a die, the probability of getting a 6 is **1/6**, so the odds against getting a 6 are **5/6** to **1/6**, or **5** to **1** against. If the relative frequency of an event **F** is **2/3**, then we would say **P(F) = 2/3** and **P(F′) = 1/3**, so the odds that **F** will happen are **2** to **1**.

To say that the odds against some event **A** are **a** to **b** is to say that the probability of that event is **P(A) = b/(a+b)**.

Thus, if the odds against a horse in a race are 3 to 2, the probability that the horse will win is **2/(3+2)** or 2/5 (40%).

To convert from the probability of an event **A** to odds against the event, you first convert the probability **P(A)** to a fraction (rational number), say **a/b**. Then the odds against the event are **(b−a)** to **a**.

Exercise on Odds

1. In the example of the change in students' pockets, what are the odds that a student will have more than $4.00 in change?
2. If the odds are **5** to **3** that an event **H** will not occur, and **2** to **1** that event **J** will occur, and **3** to **1** that they will not both occur, are these two events **independent**?

Expected Value

Expected value is the term used to describe the probable winnings from a contest or a game or the payout on insurance policies. More generally:

> The **expected value** of a variable in a random experiment is **the average value of that variable in a long run of repetitions of the experiment**.

One can imagine a (boring) game one could play with a single die. If the die comes up 1, 2, or 3, there is no payoff. If the die comes up 4 or 5, the payoff is \$2. If it comes up 6, the payoff is \$5. In this game, how much should a player expect to win? Another way to phrase this question is "What is the expected value of this game?"

There are three events that matter. The first (E_1) is the event where you throw a 1, 2, or 3. The probability is $P(E_1) = 1/2$. The second event (E_2) is where you throw a 4 or a 5. $P(E_2) = 1/3$. The third event (E_3) is the event of rolling a 6. $P(E_3) = 1/6$. If you played this game 10,000 times, you would expect that you would probably win nothing in 5,000 games, and \$2 each on 3,333 games, and \$5 each on 1,667 games. In 10,000 games, you'd probably win about $5000 \cdot 0 + 3333 \cdot 2 + 1667 \cdot 5 = 0 + 6666 + 8335 = 15001$ or \$15,001,[3] an average of \$1.50 per game.

We don't actually calculate it that way, however. Instead, we multiply the probability of each event by the value of that event, and add the results. In this case, the expected value of a game is $0 \cdot 1/2 + 2 \cdot 1/3 + 5 \cdot 1/6 = 2/3 + 5/6 = 9/6 = 1.50$, or \$1.50 per game. If it cost you \$1.50 per game to play, you could expect to break even in the long run.

> In an experiment where there are **n** events whose probabilities are P_1, P_2, P_3, … , P_n, (where P_1 is the probability of event E_1, etc.) and where the payoffs are A_1, A_2, A_3, … , A_n, the **expected value E** is calculated by multiplying the probability of each event by the **net** amount that will be gained or lost if the event occurs, and summing the results. That is, $E = P_1A_1 + P_2A_2 + P_3A_3 + \ldots + P_nA_n$.

The payoff does not have to be money. On a fair multiple-choice test, a student who knows nothing should get a grade of 0. To discourage guessing, many teachers subtract marks for wrong answers. One scheme gives one point for a correct answer and subtracts 1/4 point for a wrong answer. If a question has four possible choices and you haven't the slightest clue which one is correct, should you guess? The chance that a random guess will be correct is 0.25. The probability that it will be wrong is 0.75. The expected value of a random guess is $0.25 \cdot 1 + 0.75 \cdot (-0.25)$ which works out to be 0.0625 marks. This scheme rewards guessing. If there are five possible answers, the expected value of a guess is $0.20 \cdot 1 + 0.80 \cdot (-0.25)$ or 0 marks. It is rare that a student

---

3 Actually, the answer is \$15,000 exactly. The extra dollar was a "round-off error."

is so ignorant that every choice seems equally good. Usually even people who haven't studied the course can eliminate one or two of the choices. To discourage guessing, a more realistic grading scheme would be to subtract 1/2 mark for each wrong guess. If there are four possible answers and all seem equally likely to a totally ignorant student, then the expected value of a guess would be $0.25 \cdot 1 + 0.75 \cdot (-0.5) = -0.125$ marks. Blind guessing will probably lead to a loss of 1/8 mark per question. If you are not entirely ignorant, you can improve your odds by eliminating one of the four answers. Then you would have a 33% chance of guessing correctly. The expected value of a guess would be $0.33 \cdot 1 + 0.67 \cdot (-0.5) = 0$. If you can eliminate two possible answers, the probability of a correct guess goes up to 50%, and the expected value goes up to $0.5 \cdot 1 + 0.5 \cdot (-0.5) = +0.25$.

If a lottery sells 500 tickets at $2.00, and the prize is $1000.00, what is the expected value of a ticket on the lottery? Only one ticket will win, so the probability of a win is 1/500, or 0.2% or 0.002. The expected value of a ticket is $0.002 \cdot 998 + 0.998 \cdot (-2)$, or 0. The reason the positive value is only $998 is because even the winner pays for the ticket.

Exercise on Expected Value

Here are a couple more lines of a table of random digits to use for simulation. When you use the last digit on one row, you proceed to the next row.

110	38448	48789	18338	24697	39364	42006	76688	08708
111	81486	69487	60513	09297	00412	71238	27649	39950
112	59636	88804	04634	71197	19352	73089	84898	45785
113	62568	70206	40325	03699	71080	22553	11486	11776
114	45149	32992	75730	66280	03819	56202	02938	70915
115	61041	77684	94322	24709	73698	14526	31893	32592
116	14459	26056	31424	80371	65103	62253	50490	61181
117	38167	98532	62183	70632	23417	26185	41448	75532

1. In the gambling game of Keno, 20 numbers between 1 and 80 are chosen at random. Gamblers bet on what numbers will be chosen. There are many kinds of bets. Some of the simpler bets are:
   (a) A $1 bet on "Mark 1 number" pays $3 if the single number you mark is one of the 20 chosen; otherwise you lose your bet.
   (b) A $1 bet on "Mark 2 numbers" pays $12 if both your numbers are among the 20 chosen. The probability of this is about 0.06.
   Using expected-value calculations, decide if Mark 2 is a better bet than Mark 1.
2. Suppose the price of a stock gains or loses $1 per share per day, with the probability that it gains **P(G)** equal to the probability that it loses **P(L)**. An investor buys shares for $10 per share. He plans to sell as soon as it goes up to $11 per share. After five days, he will sell anyway. He can gain $1 per share or lose $5 per share, depending on the stock's price changes over the five-day period.

(a) Describe in detail how to simulate his gain or loss.
(b) Use the random digits given above or computer-generated random numbers or coin flipping or dice-rolling (as described in your answer to part (a)) to simulate his gain or loss over at least 50 repetitions.
(c) Based on your simulation, estimate the probability that the investor will finish with a gain. Estimate the expected value of his gain (take losses to be negative gains, so that a $2 loss is a gain of (-2)).

3. The size of a household is the number of people sharing one dwelling, regardless of whether they are related to each other. Suppose the census gives relative frequencies of different household sizes as

Household size	1	2	3	4	5	6	⩾7
Rel. Frequency	0.24	0.31	0.19	0.16	0.07	0.02	0.01

(a) Check whether this is a legitimate assignment of relative frequencies.
(b) How would you figure out the probabilities of the different household sizes?
(c) Pretend that no household has more than 7 members. Find the expected size of a randomly chosen household.

4. A couple plans to have children until they have at least one boy and at least one girl. What is the expected number of children they will have, assuming that **P(B) = P(G) = 0.5**?

14 Statistics

Objectives of this Chapter

To distinguish between data and information. To illustrate how conclusions about large collections of people or things can be reliably based on samples. To show how a few quantitative descriptions can characterize large collections. To illustrate the importance of the normal distribution for interpretation of appropriate kinds of data.

Key concepts

Population: the whole collection or set of people or things that we are studying.
Unit: any individual member of the population.
Variable: a generic property of units. When the units are crayons colour and length are possible variables.
Value (of a variable): the particular instance of a variable that is characteristic of a particular unit. For example, a particular crayon may have red colour and a length of 2 inches.
Census: measuring the value of a variable for every unit in a population.
Sample: a subset of the population chosen to represent the whole population.
Sampling: measuring the value of a variable for every unit in a sample.
Sampling method: an algorithm for selecting units for a sample.
Representative sample: a sample chosen according to a sampling method that ensures that the values of the variable in the sample resemble those values in the whole population.
Bias (of a sampling method): A sampling method is biased when that method produces samples that consistently and repeatedly differ from the population in the same direction. The values of the variable in the sample differ in predictable ways from the values in the population.
(of a measuring technique): A measuring technique is biased when repeated measurements of the same unit using that technique give results that systematically overstate or understate the value being measured.
Sampling frame: the list of units from which a sample will be taken. A subset of the population.
Simple random sampling: a sampling method whereby units are chosen from the sampling frame according to a randomizing procedure that ensures that every collection of **n** units (where **n** is the size of our sample) is equally likely to be chosen. A sample chosen by simple random sampling is called a simple random sample (**SRS**).

Parameter: the value of a variable that is characteristic of a whole population. It might be an average or a relative frequency or some other way of describing the value of a variable. Usually symbolized as **p**.

Statistic (no "s"): The value of a variable that is characteristic of a sample. Usually symbolized as $\hat{\mathbf{p}}$ ("p-hat").

Sampling variability (of a sampling method): the range or spread of values of the statistics obtained from repeated sampling of a population using that sampling method.

Sampling distribution (of a statistic): the pattern of values one would get by repeatedly sampling a population using a particular sampling method and sample size.

Measurement: any method that gives the value of some variable for one unit.

Precision (of a sampling method): the inverse of the sampling variability; that is, a sampling method that has high precision has low variability.
(of a measuring technique): repeated measurements on the same unit using that technique give similar results from one measurement to the next.

Margin of error: the range of values one would get when measuring the same value repeatedly using the same measuring technique or sampling method.

Centre (of a distribution): the mean (arithmetic average) or median or mode of the values of a variable in a sample or a population.

Outlier: a value of a variable in a sample that is so different from other values as to raise doubt about its correctness or meaningfulness.

Mean: the arithmetic average of a set of values, symbolized as $\overline{\mathbf{x}}$.

Median: the value such that half of a set of values is less than or equal to it, and the other half is greater than or equal to it.

Mode: the most frequently occurring value in a set of values.

Spread: the amount of variability in a set of values.

Percentile: The $\mathbf{c}^{\text{th}}$ percentile of a set of data is a value such that **c** percent of the numbers are less than it and the rest are greater.

Quartile: The 1^{st} quartile is the value such that 25% of all values are equal to or less than it; the 3^{rd} quartile is the value such that 25% of all values are equal to or greater than it. The 2^{nd} quartile is the median.

Five-number summary: a summary of a distribution which states the minimum, maximum, and median values and the first and third quartiles.

Variance: the mean of the squares of the deviations of the data from the mean, symbolized as $\mathbf{s}^2$.

Standard deviation: the square root of the variance, represented by **s** (for a sample) or σ (for a population).

Coefficient of variation: the standard deviation expressed as a percentage or fraction of the mean.

Standard score or **Z-score**: the number of standard deviations above (when positive) or below (when negative) the mean.

Statistics in Liberal Arts

Data only become **information** when they are interpreted – when someone makes sense of the data. **Statistics** is the systematic study of empirical facts. It includes techniques to quantify the data (express the data numerically) and to provide insight into the meaning of the data using mathematical tools.

Descriptive statistics consists of concepts and methods used to collect, organize, analyze, and present data. The goal of descriptive statistics is to collect empirical data in numerical form and assemble and summarize and depict those data so as to reveal patterns. Significant information can be extracted when we discover the patterns and relationships among the data. **Inferential statistics** consists of concepts and methods for making generalizations or predictions from data collected and organized using the methods of descriptive statistics.

In the formal sciences, we work with precise and knowable data. In "the real world" where most data are empirical, we have to cope with variability and uncertainty. Individual people and things are similar to each other in many ways, but there are important differences. To know reality, we need methods to deal with variability and uncertainty.

Collecting Data

Studies in the social sciences have shown that the concepts with which you think depend on the vocabulary you have to express those concepts. You can think more clearly about anything if you master the vocabulary. A lot of technical vocabulary is used in statistical discussion and reasoning.

Statistical information is usually about some collection of people or things. The whole collection of people or things that we're interested in is called the **population**. In statistics the word "population" does not just mean the bunch of people living in some geographic space (city, country, etc.). It is any collection that a researcher wants to find out about. It does not have to be people. A population can be all the ball bearings that a particular company manufactures or all the pigeons on this campus.

Each individual thing or person in the population is called a **unit**.

Usually we don't want to know everything about the units in our population. We want to study just one or a few of the characteristics of the units in the population. A property that we are studying (e.g., diameter of ball bearings, make of shoes worn by tennis-players) is called a **variable**. If the variable is make-of-shoes, then the particular make worn by a particular tennis-player is a **value** of that variable. The diameter of a particular ball bearing is a value of the diameter variable for that unit.

When we want to find out something about a whole population, we can get data on every unit. When we find the values of some variable(s) for every unit in the population, statisticians call it a **census**. In statistics, a census is not a national or provincial questionnaire. Measuring the diameter of every ball bearing produced by one company is a census of those ball bearings.

A census is often difficult or expensive. The data may not be important enough to justify the effort or expense of measuring every unit in the population. We can investigate a **sample** from the population and generalize our findings to the whole. In **sampling**, we select a **sample** as representative of the whole population. We find the value of the variable(s) for every unit in the sample. In a well-designed **sampling method**, data from the sample will allow us to draw reasonable conclusions about the whole population. Good methods allow us to calculate the probability that our sample results are representative of the values we would obtain if we conducted a census.

Using statistical sampling techniques and probability theory, we can sample a fairly small number of units (say, 2,000 ball bearings out of a total production of millions) and make reliable statements about the whole population of ball bearings. Measuring 1,500 people can tell us something about 27 million Canadians. We can perform an experiment[1] where we test a drug or food or medical procedure or production technique on a few units, and learn something about its usefulness or harmfulness in general, for large numbers of untested units.

Sampling

The trick in sampling is to get a sample of units that is smaller than the whole population, but which is **representative** of the whole population. If you are interested in diameter of ball bearings, you want a sample of ball bearings whose average diameter is a reliable indicator of the average diameter of all the ball bearings in the population. You might also want a sample that shows as much variation from the desired diameter as there is in the whole population. If you are interested in the buying habits of Canadian shoppers, you want a sample consisting of a bunch of people (units) whose preferences are most like those of the whole population of Canadian shoppers.[2]

The best way to ensure that samples are representative is by **random sampling**. In random sampling, we get an **unbiased sample**. We select a subset of the units in the population by using randomizing methods. In the (conceptually) simplest kind of

---

1 In the more-normal sense of "experiment," where we do something to some people or things, observe the results, and compare the results to untreated people or things (i.e., people or things to whom or to which we did not "do something").

2 The statistical sense of the word "population" is not the ordinary use of the word. The population in this case just includes Canadian shoppers, not all Canadians. Infants and people in institutions might be excluded, for example.

random sampling (obtaining a **simple random sample** or **SRS**), we make sure that every unit has an equal probability of being included in the sample.

A sampling method is called **biased** when that method produces results that consistently and repeatedly differ from the truth about the population in the same direction. It does not have to be because some person has a subjective or emotional bias. **Voluntary response** techniques (where people are asked to call a special number to record their preferences, or to send in a card from a magazine, or to write in) are notoriously biased. People who volunteer (who pay to make the call or bother to write) are probably peculiar in a number of ways (they listen to that show, or read that paper or magazine, they care enough to pay or make efforts to have their opinions counted, etc.). The difference between the results of such a sample and the opinions of the population can usually be predicted – both that it will be unrepresentative and in what direction the results will differ from the population. Sampling techniques that select people by phone number will also be systematically biased. Unlisted numbers may not be selected, and people without telephones will not be selected. People who are not at home or who choose not to answer the phone will also be left out. These omissions will produce results that differ in predictable ways from the results of a census.

To select a simple random sample we first decide the number of units we want to investigate. This is our **sample size**. Then we decide on our **sampling frame**. The sampling frame is the list or set of units from which the sample will be chosen. To sample voters, we might get the voters-list for our population. We give each unit in our sampling frame a number. With a voters list we might number the names in the order in which they appear in the list.

Then we let a random-number-generator or table of random digits select our sample for us. If there were a million units in the sampling frame, we would select random numbers between 0 and 999,999, inclusive. No number should be any more likely than any other of being chosen. If we want a sample of 1,000 units, we'd pick 1,000 distinct[3] random numbers. The units that were assigned those numbers are selected as our sample.

Other sampling designs are often used. **Systematic random sampling** selects only the first unit randomly. Then every $\mathbf{n}^{th}$ unit after that first one would be selected. There are risks in this procedure. Every $\mathbf{n}^{th}$ unit might differ systematically from the other units in the population. For example, every 100th ball bearing may have been produced by a particular machine that makes ball bearings that differ from those made by all the other machines.

Cluster sampling divides up the whole sampling frame into blocks or clusters. We might divide people up into groups of 1000 depending on where they live. We

---

3 It would be wrong to use the same number twice, because this would mean including the same unit twice in the sample.

might choose city blocks as clusters, when our sampling frame is the people in some urban area. English classes might be clusters for sampling students at this college. We select some clusters at random. We can either investigate all the units in the selected clusters, or we can take random samples from each cluster. This technique would be useful for ball bearings, where we could randomly select boxcars full of bearings, randomly select boxes of ball bearings from those boxcars, and then select bearings at random from those boxes.

Stratified sampling divides the population or sampling frame into parts, called **strata**, based on differences in some characteristics called **stratifying factors**. Stratifying factors can be things like sex, religion, or income. Among ball bearings, we might stratify our population according to the size the bearings were supposed to be (i.e., separate 1/8" bearings from 1/2" bearings).

Suppose one wanted to find out how students at an engineering college rate the college's sexual harassment policy. The opinions of female students might differ from the opinions of male students. If the college has 100 female students and 900 males, a SRS of 100 students would probably contain about 10 females. There might be none. Even if we got 10 females in our sample, such a small sample is probably not representative of all the females. We'd get more representative results by taking an SRS of 30 females from the 100 and 70 males from the 900. If the average of the ratings given by the 30 females was 2 on a scale from 1 to 10, and the average of the ratings given by the 70 males was 6, we should probably report the two numbers separately. If we really wanted to state an overall rating as how all the 1,000 students rate the policy, we would say that we expect the 100 females to rate the policy at 2, and the 900 males to rate it at 6. The overall average rating is given by $\frac{(100 \bullet 2)+(900 \bullet 6)}{(100+900)} = 5.6$.

A variable is usually a numerical characteristic (average diameter, relative frequency of blue eyes, relative frequency of people in particular age groups, etc.) of a population. When we're discussing the whole population, the value of that variable is called a **parameter**. The parameter of a population is usually symbolized as **p**.[4] If the population is students at this college, a parameter might be the average number of hours of TV per week students watch. The parameter is some particular number. To know that number we might study the viewing habits of all the students in the college (a census) over several weeks and take the average number per student per week. But that would be expensive. A parameter is a characteristic of a population.

We might take a sample – select 100 students and study their TV watching over a single week.[5] We'd calculate the average number of hours for students in the sample

---

4 That's a small "p," not a capital "P."

5 How to randomize what week? A week in the midst of mid-terms would not be representative. A week when a particularly attractive program-mix was offered would also be risky.

and take that as an estimate of the parameter. The estimate based on a sample is called a **statistic**. A statistic is a characteristic of a sample. A different sample might give a different value of the statistic. The statistic is symbolized as $\hat{\mathbf{p}}$ (called p-hat).

Exercise on Sampling

1. Some studies attempt to decide disputes about authorship by comparing mathematical properties of the texts. To compare Hemingway with Faulkner, you could use random numbers to pick a book from each author's corpus, and more random numbers to pick a page in each book. You record the lengths of each of the first 250 words on the selected page. What is the population? What is the sampling frame? What is the variable? What is the parameter? What is the statistic?
2. Imagine a sheet of paper with many non-intersecting circles of different diameters drawn on it. You want to know the average diameter of the circles, but you don't want to measure them all. Someone proposes that you select circles at random by closing your eyes and putting your pencil on the sheet. If the pencil lands in a circle, mark it as being in your sample. Repeat until you have 6 different circles marked. Those 6 are your sample. Is this likely to be a SRS? Why or why not? Is there bias in this sampling technique? What bias? What is the parameter? What is the statistic? Describe a better way to obtain a SRS of 6 circles, and say how you would estimate the average diameter of all the circles from your sample of 6.

Sampling Variability and Sampling Distributions

Suppose we want to know the average reading-ability level of students at this college. Our parameter $\mathbf{p}$ will be the average score of all students on a reading-ability test. Testing all 5,000 students would be expensive. We can get a list of student numbers from the registrar to use as our sampling frame and select a SRS of 200 students. Testing these 200 students gives a statistic $\hat{\mathbf{p}}$.

Our sample might accidentally get the 200 best readers in the college. Or the worst 200. Or we might have 100 of the best and a fairly representative bunch for the other 100. This is called **sampling variability**. There is less variability among larger samples than smaller. Sampling variability depends on the sampling method and on the size of the sample. In simple random sampling, sampling variability depends almost entirely on the size of the sample, and very little on the size of the population. Since the methods of selecting a SRS are random, we can use the mathematics of probability to study sampling variability in a SRS. There will be patterns in the long run of getting statistics from many random samples. The pattern of values one would get by repeatedly sampling a population is called the **sampling distribution** of the statistic obtained by that sampling method and sample size.

Imagine a large box containing many beads. The beads are exactly alike except that some proportion of them are red and the others are white. The whole box of beads

is the population. The parameter is the percentage of red beads in the box. We take a random sample of beads from the box and calculate the percentage of red beads in the sample. The percentage in our sample is the statistic.

To select a random sample, we mix the beads really well and plunge a "paddle" in. The paddle has 25 bead-shaped indentations on it, so when we withdraw it exactly 25 beads will be caught in the indentations. The statistic is the relative frequency of red beads in the paddle. If there are 4 red beads among the 25, the statistic $\hat{p} = 4/25 = 16\%$. We estimate that the parameter p is 16%.

In a second sample, the paddle contains 7 red beads. $\hat{p} = 7/25 = 28\%$. Another try gives 4 again. And so on. Over many tries, we get quite a few samples containing 4, 5, or 6 red beads, some with none, some with more than 10, and so on. This variation from sample to sample is sampling variability.

This is a random experiment. The outcome is the percentage (relative frequency) of red beads in the paddle. We can repeat the experiment many times, or we can simulate many repetitions with a computer program. After many repetitions of the experiment, we can summarize our findings in a table or histogram.

A simulation of 1,000 samples gave the results shown in the following table. The upper row gives the values of $\hat{p}$ that we got, and the lower row is the number of times (frequency) that each value of $\hat{p}$ came up.

value of $\hat{p}$	0.0	0.04	0.08	0.12	0.16	0.20	0.24	0.28	0.32	0.36	0.40	0.44
frequency	6	16	60	124	172	221	196	106	55	29	11	4

The value of p in this experiment was actually 0.20 (i.e., 20% of the beads in the box were red). We had quite a lot of sampling variability because 25 is a small sample. Our samples gave results ranging from 0 (no red beads in the paddle) to 0.44 (11 out of 25 beads were red). The sampling distribution table shows that 221 out of our 1000 samples (22.1%) were excellent samples, giving a statistic that was exactly equal to the parameter. Almost 60% (589) of our samples had between 4 and 6 red beads, so they were quite representative of the population of beads. They gave statistics that were between 16% and 24% (within 4% of the parameter). Although it was possible to get a random sample (paddle) that contained no white beads, none of our samples had more than 11 (44%) red beads.

We can simulate a situation where we take larger samples, even though a larger real paddle might be unworkable. As long as the total number of beads in the box is much larger than the number in a sample, the samples are still random. Simulating a 100-bead paddle, the worst $\hat{p}$ was 31% (in three of 1000 samples). In 985 samples (i.e., 98.5% of the time), the results were between 11% and 29%. 94.4% of the samples gave a $\hat{p}$ between 13% and 27%.

Measurement

The words "measure" and "measurement" are used in a somewhat unusual sense in statistics. Any method that finds the value of a variable is called "**measurement**." For example, if the variable is eye-colour, one could "measure" it by asking a subject (unit) what colour her eyes are, or by looking at her eyes, or whatever gets the value of that variable for that subject.

Some variables have values that are not numeric. If the variable we are interested in is non-numeric, we can assign a numerical value by using numbers as labels. For example, we can say that owning a Ford is a 7, or that having blue eyes is a 3, etc.

If the values of our variable are numeric, then the measurement of the values is probabilistic. We can only measure to within the **limits of precision** of our instruments. Some variables (like height) vary from one measurement to the next. Measurements only make sense within the limits of normal variability of the variable itself. For example, saying that someone is exactly 183.2 cm. tall is silly; a person's height varies by a cm. or more in the course of a normal day. Measuring a person's height to the nearest half-centimeter is **spurious precision**.

Most numerical measurements measure a continuous quantity. That is, lengths, weights and so on are real numbers, with infinite decimal representations. Most instruments can only measure to within a certain **margin of error**. The margin of error of a measuring technique is the range of values one would get when measuring the same value repeatedly. There is always a margin of error in the measurement of continuous or near-continuous quantities. When we are being very precise, we should specify values of such variables as "7.016±0.002," which says that the measurement could be as much as 0.002 off in repeated re-measurements of the same value.

Measuring such values is like taking a sample, and the same kinds of consideration of variability apply.

A measuring technique is **biased** if repeated measurements of the same value give results that systematically overstate or understate the value. By "systematically" we mean that measurements made by that technique are typically too high, for example. A technique is **precise** if repeated measurements of the same value give results that are close to each other. Notice that a measuring technique can be both precise and biased (repeated measurements are close to each other but all too high, say), or imprecise and unbiased (repeated measurements show large variability, but they are not systematically too high or too low), or any combination of precision and bias. Obviously the goal is to get unbiased measurements that are as precise as possible.

Descriptive Statistics

Numerical Descriptions

Once we obtain values for the variables we are interested in, we have to describe what we have discovered. We need to transform the data into information by summarizing and extracting the most important features from collections of numbers. Graphs and charts appeal to our visual pattern-recognition skills. The goal of graphs and charts should be to clarify the information. The danger in such presentations is that they can accidentally or deliberately misrepresent the meaning of the data. Numerical summaries can also distort the meaning of data. This is why some people are so suspicious of statistical reasoning. They cannot recognize distortion, but they know that it can occur.

One way to clarify the meaning of data is to sort and list the numbers. We can summarize the data by creating ranges of values and say how many or what percentage of the values in our census or sample were in each of the ranges (as in making a frequency or relative frequency histogram or table).

Before they look for information in a set of values, statisticians take a quick overview of the data to see if there are any values that don't seem to "belong. Sometimes a value is so hugely different from all the other values that it is reasonable to doubt its reality. Perhaps the number was entered incorrectly, or the value came from a unit that should never have been included in the sample. Such out-of-pattern values are called **outliers**, and are usually excluded from the data-set. When exclusions are made, careful researchers will note the exclusion(s) so that others who read their conclusions can be aware that some values were not included when conclusions were drawn.

It is useful to specify the centre and spread of the data.

Centre of a Distribution

Centre is the average value or the most common value or the middle value of the variable in our population or sample.

"Average" is ambiguous. It can mean the **mean** or the **median** or the **mode** of our values. The **mean** of a set of values is their arithmetic average. We add up all the values and divide by the number of values to get the mean. Statisticians usually use the symbol $\overline{\mathbf{x}}$ (pronounced "x-bar") to represent the mean of a bunch of values of a variable called **x**. The mathematical formula for the mean is: $\overline{\mathbf{x}} = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{\mathbf{n}}$. The numerator

of the fraction $\sum_{i=1}^{n} x_i$ means the sum of all the **n** values of **x** (called x_1, x_2, x_3, …, x_n). The whole formula says to calculate that sum and divide by **n**. The mean of the values 3, 3, 7, 9, 10 is (3+3+7+9+10)/5, or 6.4. The mean value is usually not one of the values among the data. The data can be something like counts, which can only have whole-number values, while the mean can be a rational-number value.

The **median** of a set of values is the middle value. Half of the values in the set are smaller and half are larger than the median. To find the median (unless we have a computer or calculator that can do it) we list all the values in order, from the smallest to the largest. If there is an odd number of values **n**, the middle number is the **(n+1)/2**th value in the list. We count down the list to that value, and that's the median. The median of the values 3, 3, 7, 9, 10 is the third (**(5+1)/2 = 3**) value, which is 7. If there is an even number of values, then **(n+1)/2** will be a number with a fractional part (e.g., 11.5). The median will be the 11.5th value, the mean of the 11th and 12th values.

The **mode** is the most common value among the data. The mode of the values 3, 3, 7, 9, 10 is 3, because there are more 3's than any other value. Sometimes a **distribution** (a collection of values of a variable) has no mode, because no one value predominates. Sometimes we speak of a bi-modal distribution, where there are two clear peaks in the distribution histogram.

Averages are not always meaningful. If the data are ordinal numbers (like birth dates or house-numbers), then taking the mean of the values makes no sense. To summarize values of an ordinal variable, it might be meaningful to find the frequency distribution of ranges of the value (how many people have birthdays in January, etc.), but the "average birthday" is nonsense. If the data are "nominal numbers" (i.e., numbers used merely to name values, as saying that owning a Ford or having blue eyes is a 3), then neither mean nor median is meaningful.

Exercise on Measures of Centre

1. Students often want to know the class average on a project or final grade. In what circumstances would the mean grade seem the most informative average? For what purposes would the median grade be the most revealing? When would the mode be most significant?
2. When would each of the three measures of centre be most appropriate for discussing the average income of some group?

Spread of a Distribution

Knowing the centre (mean, median, or mode) of a bunch of data tells only part of the story. We also need to describe the **spread** or **dispersion** of the data. The spread of the data is the amount of variability in the data. If everybody in one company

makes $1,000, then the mean income in that company (no matter how many people work there) is $1,000. If exactly half the employees earn $10 and the other half earn $1,990, mean income is also $1,000. The means (and medians) are the same, but it is important to know that the values of the variable "income" have a much larger spread in the second company. If 1,000 employees make $10 and one employee makes $991,000, the mean is still $1,000 (the median is now $10), but the spread is even larger.

A table listing ranges and the number of units whose value falls in each range describes the data's spread very clearly. So does a frequency or relative frequency histogram. Other numerical representations of spread are **percentiles** and **standard deviation**.

When the most appropriate measure of central tendency (centre) is the median, we use percentiles to indicate the spread of the data. **The c^{th} percentile of a set of data is a value such that c percent of the numbers are less than it and the rest are greater**. The median is the 50^{th} percentile.

To find the percentiles (without a computer program), we arrange the data in order from smallest to largest. To find the 8th percentile, we would count up the list until we had counted 8% of the values in the list. If there is no value that is exactly 8% of the way up the list, we'd have to interpolate between values. The process is too complicated for this course.

We can describe the spread of a distribution by specifying the **extreme** values (the smallest and largest values) and the **median** (the 50^{th} percentile) and the other **quartiles** (the 25^{th} and 75^{th} percentiles). The first quartile (the value such that a quarter of all the values are less than or equal to it) is found by finding the median and then finding the median of the values below the median. The second quartile is the median. The third quartile (the value that is greater than or equal to three-quarters of the values and less than or equal to the top one-quarter) is the median of all the values greater than the median.

Specifying the range of a distribution using extremes and three quartiles is called the "**five-number summary**" of the data. It is the most useful description of many distributions. Half of the data values are higher than the median and half are smaller. One quarter of the values fall between the low extreme and the first quartile, a quarter between the first quartile and the median, a quarter between the median and the third quartile, and a quarter between the third quartile and the high value. Half of the data values are between the first and third quartiles.

Example of the five-number summary: Mrs. Fredkin's class of 22 students raised money for a charity. The amounts (in dollars) collected by each student are listed in the following table.

29	**16**	**47**	**196**	**31**	**42**	**56**	**231**	**38**	**24**	**26**
22	**27**	**31**	**29**	**33**	**41**	**23**	**32**	**19**	**28**	**40**

The total is \$1061. The mean is **1061/22 = 48.23**. Because of the two unusually high values, it is misleading to call \$48.23 the "centre" of the data. It is higher than all but three of the values. A better description of these data is the five-number summary. We arrange the data in order from smallest to largest, as:

16 19 22 23 24 26 27 28 29 29 31
31 32 33 38 40 41 42 47 56 196 231

The median value is the **(22+1)/2 = 11.5**th value. The mean of the eleventh and twelfth values is **(31+31)/2 = 31**. The median amount is \$31. The first quartile is the median of the values below the 11.5th value, or the median of the first 11 values. **(11+1)/2 = 6**, so the first quartile (the median of the first 11 values) is the sixth value – \$26. The third quartile is the median of the values above the median, which is the sixth value above the median – \$41. The low extreme is \$16, and the high extreme is \$231. The five-number summary is 16, 26, 31, 41, 231.

Standard Deviation and Coefficient of Variation

Often we describe collections of data using the mean as the measure of centre, and **standard deviation** to describe the spread of a distribution. Standard deviation measures how much the data differ from the mean. Mean and standard deviation are most informative if the data are approximately normally distributed (if the distribution is a normal distribution). A normal distribution is one whose graph looks like the familiar "bell-curve." **If the distribution is very asymmetrical or otherwise does not resemble the bell-curve, you should not use the mean as a measure of centre, and you should not use the standard deviation as a measure of spread**. If you have good reason to believe that the distribution is approximately normal and approximately symmetric (either because the data is the kind that is usually normally-distributed (like people's heights or I.Q. scores) or because you drew the histogram, for example), then you can (and should) use mean and standard deviation to summarize the data.

Fancy calculators and "spreadsheets" or other mathematical computer programs will calculate mean and standard deviation. Still, you should understand the calculations to know what "standard deviation" means.

If all of the values are the same, then there is no spread in the data. Usually each value deviates more or less from the mean value. In the simple distribution given above (3, 3, 7, 9, 10), the mean $\bar{x}$ was 6.4. To find the deviation of a value, subtract the mean from the value. The first two values deviate from the mean by −3.4. The deviation of the next value (7) is 0.6. The other values show a deviation of 2.6 and 3.6. If we add up these five deviations, as **(−3.4)+(−3.4)+0.6+2.6+3.6**, we get **0**. The mean of the deviations is **0/5 = 0**. This tells us nothing about the spread of the values. If we calculate the mean of the absolute values of the deviations, we get **(3.4+3.4+0.6+2.6+ 3.6)/5 = 2.72**. The average distance of any unit from the mean is 2.72. This is more meaningful, but it is not the measure of deviation that is most

frequently used. Instead we calculate a number called the **variance**, and from it we calculate the **standard deviation**.

The **variance** is the mean of the squares of the deviations of the data from the mean, and is symbolized as s^2. In symbols, $s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$. The **standard deviation** is the positive square root of the variance and is represented by **s** (for a sample) or the Greek letter σ (for the population).

In our simple example, we calculate the deviation $x - \overline{x}$ for each value, as we did above (getting positive and negative answers). We then square each deviation. We get $(-3.4)^2 = 11.56$ (twice), $0.6^2 = 0.36$, $2.6^2 = 6.76$, and $3.6^2 = 12.96$. We then add these five numbers and divide by 5, getting $\frac{11.56+11.56+0.36+6.76+12.96}{5} = 8.64$. This is the mean of the squares of the deviations from the mean – the **variance**. Standard deviation is the square root of the variance, or $\sqrt{8.64} = 2.94$.

Standard deviation is always zero or positive. If it is zero, it means that every value x_i is exactly the same as the mean $\overline{x}$. There is no spread. Larger standard deviation means a larger spread of values.

To calculate the **standard deviation for a whole population**, or the **sample standard deviation** for a large sample, we get the variance by adding the deviations and dividing by **n**, as above. If the **n** is small enough[6] that it makes a difference, we divide by $(n-1)$ rather than by **n**. Many calculators have two keys, one for computing the **sample standard deviation** and the other for calculating the **population standard deviation**. The reason for using $(n-1)$ for smallish samples is because the sample's spread might accidentally miss some of the more extreme values that could occur in the population). Dividing by $(n-1)$ (which is a smaller number than **n**) gives a larger standard deviation.

In the example above the sample variance is 10.8 instead of 8.64 (because we divide by $n-1 = 4$ instead of by **5**), and the sample standard deviation is $\sqrt{10.8} = 3.29$ instead of **2.94**.

The standard deviation has the same kind of units as the variable. If the variable is measured in dollars, then the standard deviation is in dollars. If the variable is in grams, the standard deviation is in grams.

It is often more informative to get a kind of "relative" standard deviation. If our variable is in dollars, and the mean value of the variable is, say 200 billion dollars, a

---

6 That is, the sample is not a sufficiently large sample, or the population is too small.

standard deviation of a million dollars would be small. But if the variable is the amount of change a unit (person) is carrying, the mean might be only $3.00, and a standard deviation of $1.00 would be quite large.

To allow more meaningful comparisons of the variability (spread) of different distributions, we use the **coefficient of variation**. This is just the standard deviation expressed as a percentage or fraction of the mean. In symbols, $\mathbf{CV = \frac{s}{\bar{x}}}$.

A standard deviation of a million dollars in a distribution with a mean of 200 billion dollars would give a coefficient of variation of 0.00005 (very small). A standard deviation of $1.00 in a distribution with a mean of only $3.00 gives a coefficient of variation of 0.33 (rather large). The relative spread is more informative than the spread.

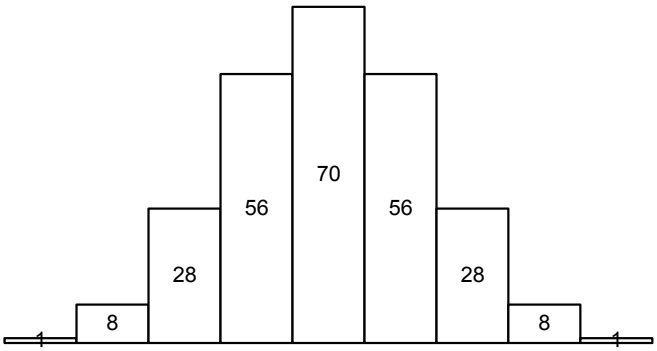
Exercise on Spread

1. Calculate the mean, the variance, the standard deviation, and the coefficient of variation of each of the following sets of numbers:
   (a) 24, 0, 6, 24, 18, 36
   (b) 25, 15, 5, 15, 20, 10
   Which set has greater spread?
2. Wechsler IQ test scores have mean of 110 and standard deviation of 25. What is the coefficient of variation of the IQ scores?
3. Suppose scores on the Scholastic Aptitude Test in a particular year have a mean of 500. If the coefficient of variation is 0.20, what is the standard deviation? Which are more variable – SAT scores or Wechsler scores?
4. Two teachers teach the same English course to Liberal Arts students at the same college. Grades in teacher A's last five classes have $\bar{x} = 72$, $s = 12$. Teacher B's last five classes have $\bar{x} = 77$, $s = 6$. If students in all classes were similar, what can you infer about the teachers' grading methods? Which English teacher's grading method do you think is more objective? Why?

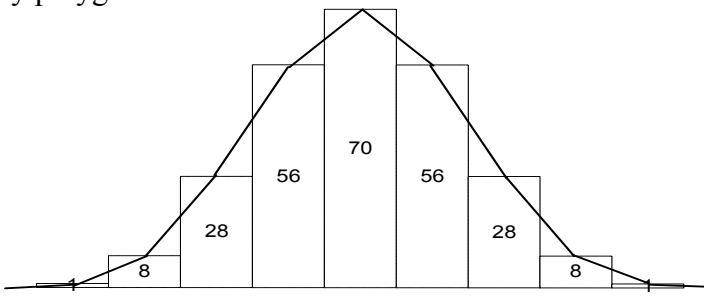
The Normal Distribution

If we take a frequency histogram and draw lines from the middle of the top of each bar to the middle of the top of the next bar, we get a **frequency polygon**.

For example, starting with the frequency histogram below,



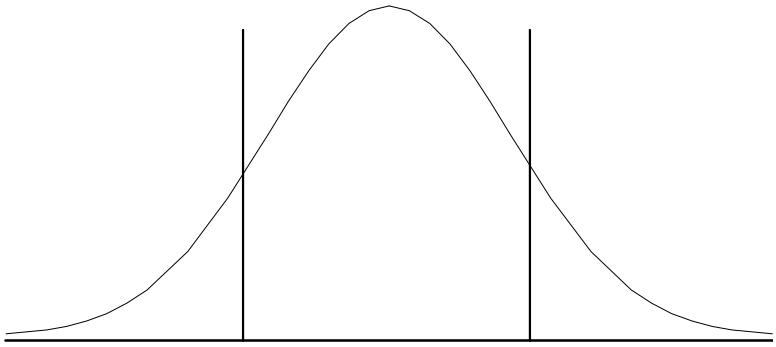
we get the frequency polygon



If you imagine a very large number of values divided into a very large number of ranges, you can see that the frequency histogram will have very narrow bars, and the resulting frequency polygon will look almost like a smooth curve. There are very powerful mathematical techniques that work with smooth curves, so we approximate distributions by using such curves. The distribution shown above can be approximated with the **normal curve**.

The normal curve is symmetric, so the mean and the median and the mode lie together in the centre of the curve. The centre of the curve is also the highest point on the curve – the mode.

One of the great virtues of the normal curve is that it gives a clear graphical meaning to the standard deviation. If we look at a normal curve, as



we see that the vertical lines cut the curve at the points where the curve changes direction. That is, as we go up the curve from the left end, the curve is getting steeper and steeper until it crosses the vertical line. After the vertical line, the curve slopes less. As we pass the peak of the curve, the slope becomes negative. It becomes increasingly negative until we get to the second vertical line, where the slope stops increasing and the curve becomes less and less steep. At the extremes, the curve is almost horizontal.

The points where the curve changes direction (marked by the vertical lines) are exactly one standard deviation away from the centre of the curve. Thus, the distance between the two vertical lines is two standard deviations. If the curve shows I.Q. scores on the Stanford-Binet test (mean of 100, standard deviation of 15), the left vertical line would be at a score of 85, the peak at 100, and the right vertical at 115.

More importantly, the areas under the curve are proportional to the number of values in the data. 68% of all the values will occur between the two vertical lines. That is, if our data are normally distributed, then 34% of the units will have values between one standard deviation below the mean and the mean, and 34% will be between the mean and one standard deviation above the mean. If Stanford-Binet I.Q. scores have a normal distribution (and they do), then 68% of all the units (people) we test will have an I.Q. between 85 and 115. The mathematics of the normal curve also tell us that 95% of all values will be within two standard deviations of the mean. Using the I.Q. example again, we can predict that 95% of people will score between 70 and 130 on the Stanford-Binet test. Finally, 99.7% of all normally distributed values will be within three standard deviations of the mean. Only 0.3% (three in a thousand) will score lower than 55 or higher than 145 on that test. Since the normal curve is symmetrical, only 3 in two thousand (0.15%) will score above 145. These special facts about the normal distribution are called **the 68-95-99.7 rule**.

All of this says that if our data are normally distributed, the standard deviation tells us a great deal about the distribution of the data.

People's heights are approximately normally distributed. If the mean height of Canadian women is 163 cm. and the standard deviation is 6.5 cm., we can easily estimate what percentage of Canadian women is over 176 cm. tall. 176 cm. is two standard deviations greater than the mean. 95% of all Canadian women will be within two standard deviations of the mean, so the other 5% will be taller or shorter. By symmetry, about 2.5% will be taller than 176 cm.

The number of standard deviations above or below the mean is often more meaningful than the actual value of the variable. Knowing how many standard deviations away from the mean a value is tells us how "normal" or how unusual it is. For this reason, we have a special name for the number that represents how many standard deviations above or below the mean a value is. We call it a **standard score**

or **Z-score**. The standard score is calculated as $\mathbf{Z = \frac{x - \bar{x}}{s}}$. Subtract the mean from the value of the variable for a unit and divide the result by the standard deviation. This gives the number (positive for values above the mean, negative for values below the mean) of standard deviations between that value and the mean. A woman whose standard score for height is 1 is pretty tall (one standard deviation taller than the average, or taller than 84% of women. A woman whose standard score for height is 2 is very tall – taller than 97.5% of women. Only 2.5% of women are taller than she. If she has a Z-score of 3, only 0.15% of women are taller. This tells us much more than just noting that she is 183 cm. tall.

We saw that the coefficient of variation was a more meaningful measure of the variability of data than just the standard deviation. Here we see that the variation of an individual from the norm is more meaningfully expressed in relative terms, i.e., relative to the standard deviation from the norm.

What about fractional Z-scores (e.g., 1.35 standard deviations above the mean)? There are tables that give the relationship between percentiles and standard scores. Using the tables or a computer (or some *very* complicated math), we can find what percentile some value is in by calculating its standard score and looking it up.

Exercise on the Normal Distribution

1. Heights of 18- to 24-year-old men are approximately normally distributed with mean 180 cm. and standard deviation 8 cm. What percentage of men in this age group are taller than 188 cm.? What percentage are shorter than 164 cm.? What percentage are taller than 196 cm.? What percentage are taller than 204 cm.?
2. Wechsler Adult Intelligence test scores for 20- to 34-year-olds are approximately normally distributed with mean 110 and standard deviation 25. Wechsler scores for 60- to 64-year-olds are approximately normally distributed with mean 90 and standard deviation 25. You and your 62-year-old grandma take the tests. You score 148, and Granny gets 135. Whose score is higher relative to his/her age group? Defend your answer.
3. In a simulation using a 100-bead paddle to draw beads from a box, the beads in the box (the population) had a parameter value of 0.20. It has been calculated that very large numbers of such samples will be approximately normally distributed with mean = 0.20 and standard deviation 0.04. What is the probability that a sample could have 32 or more red beads?
4. A large company has a clerical staff of which 20% are males. Of the last 100 clerical workers chosen for promotion, 32 were males. What is the probability that 32 or more males would be chosen if the selection were random?