CHAPTER 9

EVALUATING DATA

THE TESTABLE HYPOTHESIS AND MULTIPLE INDEPENDENT MEANS OF CONFIRMATION

Science deals with the mechanisms of how the world operates, and one of its basic tenets is that we determine how things operate by building hypotheses and then designing experiments to attempt to falsify or disprove the hypothesis. It is therefore extremely important to design the experiment and to describe it in such a manner that other scientists can repeat it to convince themselves of the validity of the results. Although it is not impossible, in some fields such as astronomy and evolution, for obvious reasons it is difficult to conduct experiments. In these cases we use other, less direct, means of validating our hypotheses. For instance, we might predict what we should find in a situation we have not yet investigated and then investigate it. Such an approach was used to test Einstein's Theory of Relativity by realizing that the theory predicted that light rays could be bent by gravity. The next full eclipse of the sun provided the opportunity to determine if the light coming from stars almost behind the sun was bent by the gravity of the sun (Fig. 9.1, discussed in more detail on page 114).

The experiment was conducted and proved to be one of the first convincing arguments in favor of the theory. This is called a thought experiment or, since this type of experiment was first elaborated in Germany, a Gedanken experiment. Most of us do such experiments on a regular basis. For instance, as a child and left-hander, in baseball I routinely hit balls to right field. Fielders knew this and positioned themselves to catch my hits. I hypothesized that I started my swing too early. If I could resist starting my swing for a fraction of a second, I could get the ball away from right field. I tried this approach and achieved at least a partial success.

Another approach, especially useful when one has little option to modify the situation, is to accumulate many independent lines of evidence that point to the same conclusion. This is a difficult concept to understand, even though it too is fairly commonly used in everyday life. For instance, if you see a flash of light, hear a boom, and smell smoke, you are quickly convinced that an explosion has taken place. Your seeing the flash of light depended on light waves, which you detected with your eyes. What you heard depended on sound waves, or vibrations of air (and sometimes more solid material, in which case you would feel the shaking) and were

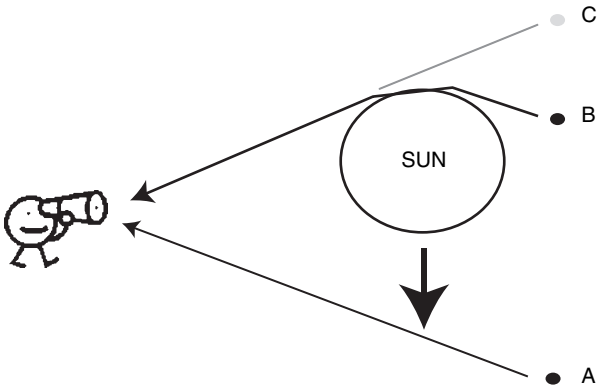

Figure 9.1. The Einstein Gravity Experiment. Since even if light is bent the observer interprets it to have moved in a straight line, if it is bent the star appears to shift position. As the sun moves across the heavens, during an eclipse the position of stars distant from the sun can be correctly assessed (A). When the star is behind the sun, the light is bent or deflected by the gravity of the sun (B) so to the observer the star appears to have jumped ahead (gray line, C). An object the size of the sun is large enough to produce a measurable shift, but the effect can only be seen when the light of the sun is blocked by the moon in a solar eclipse

detected with your ears. The odor is chemical and is detected by odor-sensing cells in your nose and sinus. You could see a flash, like a camera's flash, without hearing a boom; you could hear a boom, if an object fell, without light; and you could smell an odor independent of either. Not one of these sensations (inputs of data) depends on the other. Yet each one supports the theory that an explosion has taken place. Even more convincing, the direction from which each comes can be ascertained. If they all appear to come from the same place, the direction of each also supports the theory of an explosion. Each by itself could suggest an explosion or something else. An acrid odor alone might suggest a fire but not an explosion; a boom without a flash of light or an odor might suggest that something has fallen. But all three together indicate an explosion in a particular direction. A charred hole remaining in the region would be another bit of evidence, as would the information that a tank of acetylene gas (welding gas) had been stored in the area. This is what is meant by *multiple independent sources of evidence*. It is different, for instance, from the argument that you heard a boom and a sound-recording device indicated a loud noise. Both of these sources of data derive from the same source, the generation of sound waves, and are therefore not independent. They are independent evidence that there was a sound, but not independent evidence of an explosion. A worse example would be if, in an earthquake in a rainstorm, a building wall cracked and water got into the building, the entrance of water would not be further proof of the tremor, because it would depend on the first proof, the crack in the wall. It would be evidence dependent on other evidence. (Table 9.1).

Such documentation is routine in daily investigations, now so beloved of forensic police stories on television. For instance, one can deduce a car's speed from the

Table 9.1. Dependent evidence and Independent evidence

Independent evidence: The world is old				
Evidence 1	Evidence 2	Evidence 3	Evidence 4	Evidence 5
Isotope concentrations	Red shift	Thermal calculations	Uplift of mountains	Erosion
Theory of decay of radioisotopes	Wave theory (properties of waves in water as well as in sound and light)	Theory of energy, radiation, and specific heat of objects	Theory of plate tectonics (movement of continents)	Geophysics; friction and integrity of matter
Independent evidence: The world is round				
Measurement of angle of sun at noon	Shadow of earth on moon	Observation that other heavenly bodies are round	Circumnavigation of globe (Magellan)	View of earth from moon
Geometry (Greek)	Light propagation (light travels in straight lines)	Inference	Direct demonstration	Direct visualization
Dependent evidence: The earth was created 6000 years ago				
Calculations from the Old and New Testaments give this figure	The Catholic Church states this as belief	Jewish Fundamentalists and older philosophers state this as belief	Islamic Fundamentalists state this as belief	Fundamentalist (Protestant) Christians state this as belief
Written document	Ultimate source is same as first column	Ultimate source is Old Testament as in first column	Ultimate source is Old Testament as in first column	Ultimate source is same as first column

(*continued*)

Table 9.1. (*continued*)

Evidence 1	Evidence 2	Evidence 3	Evidence 4	Evidence 5
Dependent evidence: Little old ladies shouldn't drive				
I saw a little old lady driving dangerously	My wife saw her too	The driver coming the other way honked	My child says so too	
One observation is not a generalization	Confirms the fact that she was not driving well, but does not extend the generalization	Confirms the fact that she was not driving well, but does not extend the generalization	My child got the information about the incident from me	
Dependent evidence: News item is true (in several papers & book)				
A bomb was found on a plane Reported on CNN	A bomb was found on a plane NY Times reports that CNN reported it	A bomb was found on a plane Time magazine reports that CNN reported it	A bomb was found on a plane NY Post reports that CNN reported it	
Dependent evidence: Large cells are more likely to be cancerous				
There are more large cells in cancers than in normal tissues	If one separates large cells from small cells, one gets more large cells from cancers	If one isolates large cells, they are frequently cancerous	QUESTION: WHAT WOULD CONSTITUTE INDEPENDENT EXPERIMENTAL EVIDENCE TO TEST THIS HYPOTHESIS?	
Microscopic or other measuring observation	This is essentially the same observation as the first	This is essentially the same observation as the first		

length of skid marks, because the faster an object travels, the longer it takes to stop it. This is called momentum. One can also judge the speed by the amount of damage in an accident. This is also a matter of momentum, but is a measure of the amount of force needed to stop the vehicle, while the skid marks are a calculation of the amount of time and distance needed to stop the vehicle, making a reasonable assumption about the force applied by the brakes. Another means of judging speed would be the amount of time elapsed to cover a specific distance, for instance if a traffic surveillance camera recorded the car as it moved down the road. Or a bartender could remember that the driver, leaving the bar, walked in front of the television just as a home run was hit at 10:07, and the accident occurred 20 minutes later but 30 miles away, allowing a calculation of a minimum average speed of 90 miles/hour. Typically, if all three calculations gave approximately the same result, the conclusion would be reasonably certain. However, if one of the calculations gave results noticeably different from the other two, any good lawyer would be able to convince a jury that the speed was uncertain.

This type of argument holds even in much simpler cases brought to trial. I could claim that you insulted or threatened me and try to press charges against you on that basis. You of course would deny that you had, and a police officer or judge would simply say that it was a case of "he said, she said," and that there was no way to tell who was lying. In other words, there was no corroborating evidence. However, if other witnesses, not known to you, me, or the others, voluntarily came forward and each reported a version of the incident very similar to either your or my claim, we could describe this corroborating evidence as independent confirmation of the claim. If furthermore a video surveillance camera captured us gesticulating in a fashion that was consistent with that claim, and both of our subsequent behaviors were consistent with that claim, this would also be independent confirmation that a judge would accept as evidence.

This understanding of the concept of *multiple independent means of verification* is very important to the study of evolution because, in the narrowest sense, we cannot experiment to determine the age of the earth. We therefore must deduce it from any sources of information that we have. The two major types of source are creation legends and extrapolations of physical data. For instance, the Abrahamic creation legend, Genesis, as followed by Judaism, Christianity, and Islam, suggests an approximate age of the earth of 6000 years. Other legends, such as those of Asia, the natives of the Americas, and Africa, generally vary in the range of 6 to 10,000 years. On the other hand, several physical measurements suggest ages over one million times longer: 14.5 billion years for the universe, 4.5 billion years for the earth, 2 billion years since the origin of life, 400 million years since the rapid expansion of life, and 65 million years since the disappearance of the dinosaurs. For those accepting physical evidence, the 6–10,000 year figure more appropriately reflects the development of civilization, following the appearance of true modern humans approximately 50,000 years ago.

So which figures are we supposed to believe, and why? Scientists believe that in all cases, evidence and logic rule. But what is evidence? In many cultures and times,

holy writings were (are) considered to hold greater validity than what is perceived or calculated. Did the Red Sea divide? Did Joshua command the sun to stand still? Did Jesus walk on water? Most of us are familiar with the situation that different witnesses to the same event will later remember the event differently. Psychologists and sociologists can readily demonstrate that people's memories can be changed by social or psychological pressures, and that suggestion can alter memories. My children have vivid memories, from their formative years, of events that did not happen (I think) but which appear to be an amalgam of different incidents and suggestions.

This is the primary argument of Chapter 8, page 95. To a scientist, several branches of science produce data that has not been refuted, and each source of data points to the same conclusion of an age for the earth of approximately 4.5 billion years. The ratio of lead to uranium, and other ratios, in rocks are based on our understanding of the mechanisms of radioactive decay, with each calculation based on our measurement of the rate of decay of that element. Calculations based on the current temperature of the earth take into account various laws of physics relating to the dispersal of heat, radiational warming of the earth by the sun, generation of heat by friction and by radioactive decay. Calculations of the age and size of the universe rely on the properties of waves (of light) as one speeds toward or away form the source, as well as on laws of gravity and momentum. The age of the continents, and the age of fossils, are estimated from rates of sedimentation and erosion as well as the principles of how layering occurs and variations in magnetic fields. Today by direct measurement from satellites we can measure the movements of continents, previously inferred from magnetic fields and the types of fossils found on the continents. Thus to a scientist the hypothesis that the earth is billions of years old is supported by many independent lines of evidence, while the hypothesis that the earth is 6,000 years old is supported by one primary document. Whether or not one considers this document to be evidence or simply a hypothesis depends on one's faith, but to a scientist it does not constitute evidence in the sense that it can be tested and subjected to falsification. On the other hand, to refute the hypothesis that the earth is a few billion years old, one would have to deny the logic, experimentation, and evidence of several huge branches of scientific analysis and experimentation (geology, several branches of physics, astrophysics, biochemistry, molecular biology, among others) or at least explain the exceptions to theory that a much younger earth would pose. The data from these different lines of evidence converge in the same sense that a group of people in a circle, trying to locate an owl in the woods, each individually point in the direction that they hear the owl. Where the indicated lines intersect, the owl is likely to be (Fig. 9.2).

BIOLOGICAL DNA CLOCKS

Why biochemistry and molecular biology? We will return to this question in Chapter 15, page 243, but for the moment a simple explanation will suffice. Assume that DNA is one long string of chemicals that carries the information to construct an

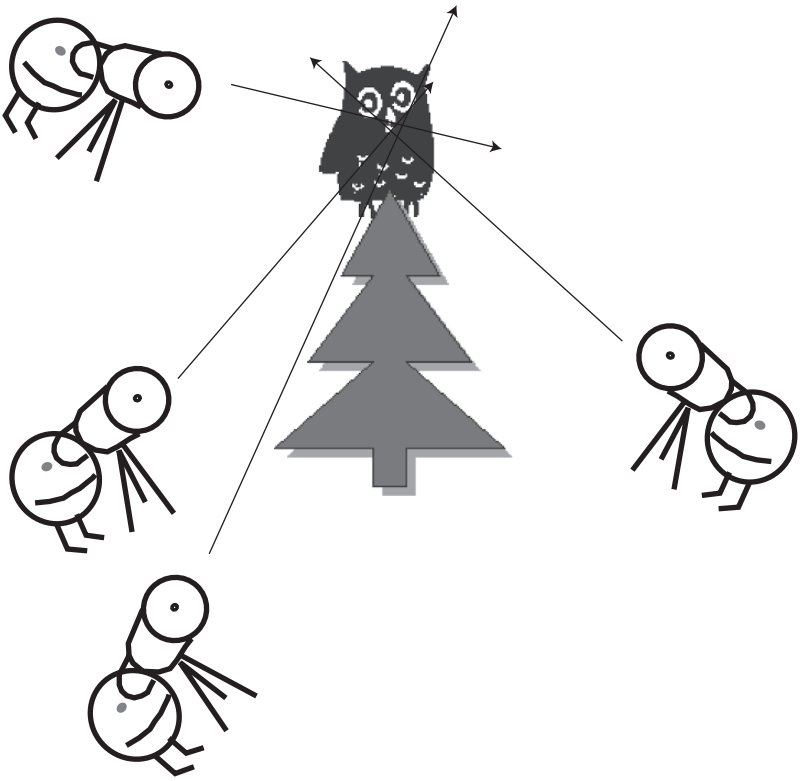

Figure 9.2. If several observers can each identify the direction of the sound of an owl, then the intersection (convergence) of the directions will give the position of the owl. If one considers all possible interpretations of a phenomenon as representing all possible directions in a circle, and a hypothesis as one of these directions, then the fact that many independent hypotheses converge on one central point adds considerable strength to the confidence that the hypothesis is correct

individual. You can consider it to be a very long string of natural pearls, each of which is distinct and identifiable. We will add one further consideration: rather than being loose on a string, each pearl has a snap-together end like children's toys.

Now let us make a further assumption: the string falls apart, or breaks, on a fairly regular basis. The actual date of breaking may vary, but if it breaks apart 36 times in a year, it averages 10 days between breaks. This is the same type of calculation that indicates the expected lifespan of light bulbs. We will also assume that the string is put back together each time it breaks, but the pieces are not necessarily rejoined where they broke. Thus the strand will look a little different each time it is repaired. If, for instance, the pearls were originally arranged in the order of size or color, the distribution of sizes and colors will become more random. For this argument, most importantly, the length remaining of the original ordered strand will become shorter with time. If we know how often it breaks, we can calculate

Figure 9.3. DNA breaks as a clock. Since each break is random, the probability of preserving the intact piece A…z decreases with time. This would be true for any arbitrary segment in the chain. Thus, the longer the time has passed since the chain was intact, the shorter the fragment will be

how long it has been since the first strand existed. We can also tell what that strand looked like by studying the different strands and determining the order of breaks.

All we have to do is substitute "DNA" for "pearls". DNA strands tend to break at a predictable average rate. We also know of short stretches of DNA that are found in almost all animals and plants. These must be very ancient and represent something similar to the original DNA. The shortness of these pieces gives us some estimate of the time that the DNA has been changing.

We can also use these comparisons to assemble an order of relationship. For instance, there are many more long pieces in common between apes and humans, or between lions and tigers, than between humans and tigers. We can conclude that there were stretches of DNA common to the ancestors of apes and humans long after the last time that the DNA was common to apes, humans, lions, and tigers. In other words, we can establish an evolutionary tree. See Fig. 9.3.

EVALUATING POPULATION MEASUREMENTS: BELL CURVES, STATISTICS, AND PROBABILITY

A major difference between a research laboratory and "the rest of us" is the level of control that can be exercised in a laboratory setting. In a laboratory if we wish to determine, for instance, that a given chemical can cause cancer, we might use mice that are so inbred that they are effectively identical twins or even clones. We

would have records of how long they normally lived and the causes of their deaths. We would give them all identical food, and maintain them with equal numbers of animals in each cage, at standard temperatures and lighting conditions. To test our drug, we would administer it in various doses at known times during the day, during a known point of the ovulatory cycle if we are using female mice, to mice of a specific age. We would treat the control group in exactly the same fashion, including giving them an injection of the solution in which our drug was dissolved if it were given by injection since the stress and pain of the injection might affect even a normal animal. We would be concerned that all environmental conditions should be the same. We would establish a standard protocol that would determine how long we would wait to examine the outcome and the criteria by which we would evaluate the outcome. We might carry the study further to examine the effect of the chemical on cells in culture, in this case controlling all parameters (measurable variants) to the point of artificiality. For instance, we might use cells that were altered so that they could not degrade the chemical, or cells that were rigged so that they were far more sensitive to potential carcinogens (cancer-causing agents) than most cells. Our goal would be to assure that every conceivable parameter was identical in the control and experimental situation, so that the only difference between the control and the experimental situation was the chemical.

Of course nobody is perfect, and there is always likely to be something that we cannot control, as scientists learned when they realized that animals responded differently to drugs depending on the time of day at which the drugs were administered. Other unanticipated factors have proved to be increasing levels of fear if, for instance, mice hear the squeals of the control group being injected before the experimental group is injected; the sound of a watchman doing his rounds at night being sufficient to synchronize certain aspects of physiological rhythms; the fact that females caged together will synchronize their ovulations; a slight difference in weight between muscles on the left and right sides of the body; or the fact that, when one opens a mouse cage, there is a difference between the mice that come up to see what the activity is and those that flee to the back of the cage. Even the most highly inbred animals do not all die at the same time or of the same causes.

Thus the issue of control is an extremely important one in science, and at meetings many arguments ensue concerning the adequacy of the control experiments. In fact, if one reviews Nobel Prize-winning research and asks how it differs from most other research, very frequently a major difference turns out to be that the future laureate was suspicious that the presumptive control for his or her experiment was not adequate and decided to verify that control, thus stumbling upon a surprising new interpretation of the data. This issue is discussed further under the heading of controls (page 135).

Try as we might, we never have perfect control over our experiment and, particularly in the life sciences, we can never guarantee that an exact copy of an experiment done yesterday will come out exactly the same today. This is why we must repeat experiments. Perhaps yesterday something distracted me, and I inadvertently incorrectly diluted the drug so that it was too concentrated. Perhaps one of the vials

or test tubes that I used was not clean, and it contributed color to the reaction. Perhaps the reagents had thawed and refrozen during shipping and were no longer good. Perhaps the mouse that I used for the experiment was already sick and I hadn't noticed. Perhaps the machine that I used to measure the results was not calibrated correctly. (The level of sensitivity that we rely on today is outstanding. We can easily measure a femtogram, or 0.000000000000001 (10^{-15}) g. In closer to real terms, if we dissolved a cube of sugar in a volume of water equivalent to 200 Olympic-size swimming pools, we would still be able to measure the sugar in one milliliter, or 1/16th of an ounce. As I tell my students, we can never actually see our results. We rely entirely on what our machines tell us. Therefore we have an urgent need to understand what the machine is measuring and to know that the machine is working properly.)

For these reasons a single repetition of an experiment is not acceptable and would not be accepted for publication. You are familiar with this logic. You know that a vaccination will not provide a 100% guarantee that you will not get the flu: you might have the flu already; the vaccination might not "take"; or you might get a different kind of flu. The same is true for our experiments. Not all of the mice exposed to the carcinogen will die of cancer. How then can we interpret our results?

EVALUATING DATA: CORRELATION AND STATISTICS

All of the above is an inferential or deductive logic, as is illustrated in Table 9.2. However, in most other instances, scientists attempt to test hypotheses. They try to devise experiments that establish evidence and can generate a logical hypothesis of

Table 9.2. Formal logic

Proposition (Hypothesis)	Prediction	Result	Comment
True	True	True	
The earth rotates toward the East	The sun will rise in the East	True	If the hypothesis is true then the prediction will also be true
False	True	True	
The earth is still; the sun traverses the Heavens from East to West	The sun will rise in the East	True	However, a result can be true even if the hypothesis is false
False	True	False	
The earth is still; the sun traverses the Heavens from East to West	If this is true, a point in the Heavens should be fixed relative to the earth, and should not move	False	With a suitable prediction, a hypothesis can be proved false

causality. Furthermore, this hypothesis is constructed in such a manner that it can be tested by trying to disprove or falsify it. These are the three tenets of the method presented in this book: **Evidence, Logic, Falsification** (ELF).

In many instances different phenomena can be correlated even if they are not related. A baby born in December gets larger as the weather grows warmer. We cannot conclude however that the temperature causes the baby to grow, and in fact the correlation will deteriorate as autumn comes. However, for a particular type of caterpillar that always hatches in early May and spins its cocoon in early July, we could not dismiss the correlation. A well-known example is the correlation between poplar trees and polio, seen in the United States in the 1950's. Poliomyelitis was a frightening disease in the US during the 1940's and 1950's. It killed 6,000 people and left 27,000 paralyzed during an expansion of the epidemic in 1916. Just before the appearance of the first polio vaccine, in 1952, there were 20,000 cases. Public parks, swimming pools, and cinemas were closed to attempt to limit exposure. During this period, some researchers noted that there was a pronounced correlation between the number of poplar trees in a neighborhood and the likelihood of a case of polio in the neighborhood (Fig. 9.4).

Do we conclude that poplar trees have something to do with the spread of polio? Well, it is certainly worth checking out, but in this case it proves to be a false correlation. What happened here was the intersection of two variables that were both correlated with income. First, in the Midwest and elsewhere, following the 2nd World War, there was considerable construction of housing for returning soldiers and to accommodate the growing economy. Of this housing, the cheapest housing tended to be apartment blocks with few trees. The most expensive housing was individually designed, with carefully selected landscaping and individual choice of trees or sparing of the original trees during construction. The middle-class housing tended to be in large tracts, where the developers used inexpensive fast-growing trees such as poplars. Similarly, in the face of the polio epidemics that expanded each summer, wealthier parents sent their children to isolated summer camps that the disease, spread from person to person, did not reach. In the poorest, crowded, neighborhoods, children were exposed to polio as infants. A peculiarity of the disease is that it can be a very mild disease in the youngest children. Some children die, but for many it seems to be a brief flu, not diagnosed as polio, and they recover with no nerve damage, immune to further attacks. So, many of the poorest children proved to be immune to polio. In the middle classes, however, attention to cleanliness and protection of infants was such that they were not exposed as infants. When these children were more sociable and went to movies or public swimming pools, they were exposed and came down with polio. Thus polio was more common in the middle classes, who tended to live in neighborhoods where poplar trees were planted. There was a reason for the association of poplar trees and polio, but the association did not establish that poplar trees caused polio (or for that matter, that polio caused poplar trees). There was some evidence for the argument, but there was no logic behind it, and indeed it was relatively easy to falsify it.

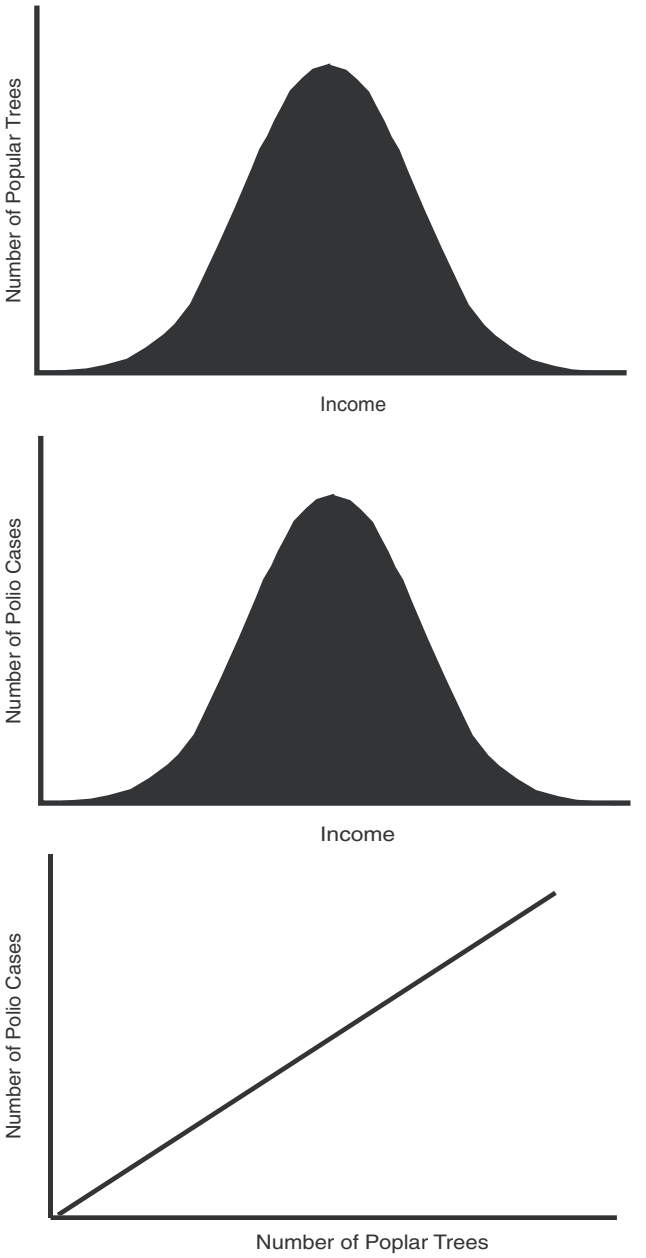

Figure 9.4. False correlations. Poplar trees were most commonly planted in developments intended for lower middle income families (A). For other reasons, frequency of polio was highest in this group as well (B). Thus there is a correlation between income and polio (C), meaning that the numbers are similar, but the correlation in no way implies that poplar trees cause or otherwise directly affect the chance of getting polio

This kind of false syllogism, that correlation proves causality, is extremely common and can be seen almost every day. In the 1990's there was considerable publicity over the possibility of transmission of AIDS by mosquitoes, based on a cluster of AIDS in Florida. In this instance, the ability of mosquitoes to transmit AIDS was examined from several angles and ruled out, and the cluster was later related to a truck stop near a swamp that was frequented by prostitutes. In another instance, there was a lawsuit that received considerable attention, in which a man claimed that his wife used a cell phone constantly and ultimately developed a rare brain tumor. He argued that the cell phone must have caused the tumor.

The point is that correlation may be used to support an argument, but it cannot be used to prove an argument. A cigarette smoker may get lung cancer. In terms of a large population, the correlation between smoking and lung cancer argues that it is extremely unlikely that they are not related, and many other potential explanations for the correlation have been disproved. However, consider it in terms of a lawsuit. Suppose the smoker changed brands of cigarette several times during his or her lifetime. Since cancers take several years to develop, the smoker cannot sue brand B claiming that brand B caused the cancer, since there is no way to determine if brand A, B, or C, or some combination of the brands, caused the cancer, or even if indeed the cigarettes caused the cancer. The problem in brief is this: you cannot determine which cigarette, on which day, caused the cancer, or even if a random cosmic ray caused the cancer. In other words, a single instance—as is often presented in commercials for non-prescription health supplements—is anecdotal; it is not proof. The fact that a single cell phone user gets cancer does not create a connection of cause and effect. The fact that a dog barked at my car on the same day that I found a $100 bill on the street does not prove that barking dogs cause money to fall in my path (or that money in my path causes dogs to bark). This is why we need the criterion of falsifiablility. In a correlation, you can never prove that one action caused another. However, you can prove that a hypothesis is false. If I hypothesized that eating lettuce caused lung cancer, you could set up an experiment in which lots of humans or guinea pigs ate lots of lettuce and did not develop lung cancer, proving my hypothesis false. Note the issue of the experiment. We can try to eliminate every variable except the eating of lettuce. We can have lots of guinea pigs of the same genetic background, control the quality of the air they breathe, and give them diets alike in every way except for the lettuce. In humans, we would have to consider all other types of variables: age, weight, sex, occupation (Do they work in environments in which the air is polluted in any way?) history of smoking, diet, genetic background, family history of lung cancers, etc. If we have a large number of extremely well-matched people who differ only in that one group smokes and the other does not, and it turns out that the smokers develop lung cancer, then we have very good correlation, and also a logic, since we can demonstrate in mice that cigarettes contain products that cause cancer. However, it always remains possible that we have missed a factor that produces the correlation. For instance, why do some people smoke and others not? Is it a difference in nervousness, and this difference can in some manner lead to greater susceptibility to cancer? By and large,

smokers drink more alcohol than non-smokers; is it the alcohol? Was there some difference in the childhood of the smoker compared to the non-smoker that led the former to smoke? Would this be associated with the difference in rate of cancer? In brief, correlation is evidence, but it does not provide the logic, and we need the falsifiability because we can disprove one hypothesis but it is never possible to rule out a universe of alternatives and therefore actually prove a second hypothesis.

EVALUATING DATA: STATISTICS

Since in many circumstances we cannot directly prove causality, we are often left with correlation. In the best of all possible worlds, we can construct an experiment that will create a situation so unlikely or restricted that it is nearly impossible to imagine any alternative to the primary hypothesis. This is discussed below. However, particularly in dealing with human data, we ethically cannot conduct an experiment such as injecting people with bacteria to demonstrate that the bacteria cause a disease, or the situation is simply too complicated to allow us to rule out other causes, such as diet, climate, ethnicity, etc. In this case we necessarily fall back onto the use of statistics. But what is statistics? What does it mean, for instance, that a "poll has a margin of error of 5 percentage points"? What is "statistically significant"? Let's start by considering a fairly common statistical issue, seen in every pediatrician's office, a growth chart (Fig. 9.5).

This chart gives heights and weights for growing boys. What does it mean? First, look at the heavy center line. This line marks the average, or mean for each age. It represents the number at which 50% of the boys are above the line and 50% are below the line. For instance, at age 10, 50% of boys are 54 ½" (4'6 ½") or taller, and 50% are shorter than that. Other lines could be drawn to mark the 25th and 75th percentile–the heights at which 25% of boys would be taller and 25% would be shorter, and 50% in between; or between the 10th percentile and 90th percentile, which would include the 80% of boys who are neither in the top 10% nor the lower 10% in height. At the 5% level (tallest 5% and shortest 5%) we consider the heights to be statistically significant. This does not mean that there is a problem, but that we might consider investigating further.

SIGNIFICANCE

The interest of this curve is to answer the question, when should we worry if a child is too small or too tall? The answer is that each decision will be individual, but we use statistics in a specific way to give us a hint. What we do is to define 'significant' in a highly technical fashion. To a scientist, 'significant' does NOT carry the popular sense of "having meaning" as in "a significant glance"; "having influence or effect" as in "a significant piece of legislation"; or "a substantial amount" as in "a significant number of votes" or even "much more than casual" as in "a significant other". To a scientist, the word "significant" has one meaning only: "unlikely to occur by chance alone more than five times out of 100". For instance,

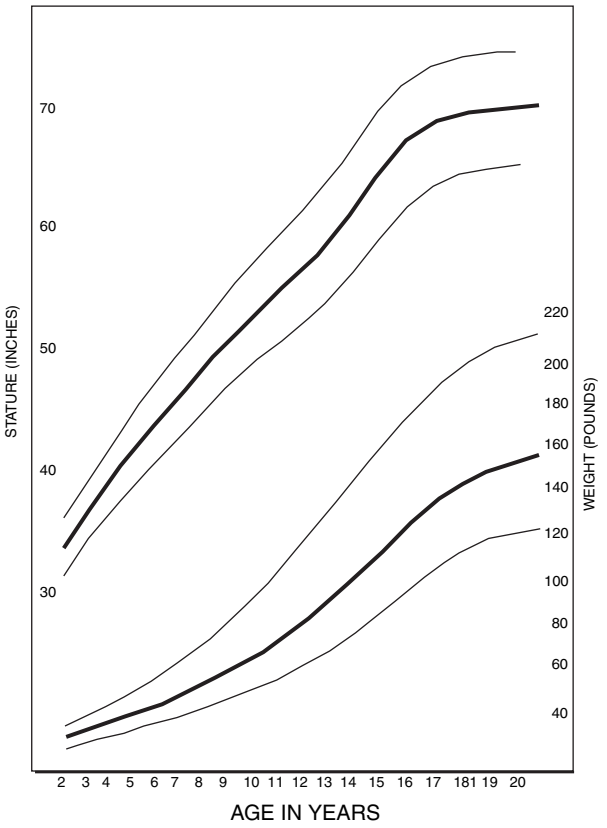

Figure 9.5. A growth chart. The middle line indicates the height at which, in a very large population, the boy with exactly average height for a specific age would fall. The upper line indicates the height above which 5% of the boys would be, and the lower line the height below which 5% of the boys would be. Since there will always be 5% in the upper or lower 5th percentile, no abnormality is implied. However, if a boy's height is at either extreme, a pediatrician might choose to verify that there is not a problem

in the example of the growth curves, we would perk up our ears if a boy was above the 95th percentile or below the 5th percentile—this would be significantly far from the mean—but we would not conclude that the boy had a problem. After all, 5% of the boys will be above the 95th percentile and 5% below the 5th percentile. We would simply say, "This boy is sufficiently far from the mean that we should investigate whether or not there is a problem." Is his height consistent with that of his parents? Is it clear that he is eating properly and digesting and absorbing his food properly? Are his hormone levels normal? Is his bone development normal?

Statistics comes from the property of all data to cluster around a mean. When the causes are random or numerous, the clustering takes on a particular shape, called a bell curve. If we were simply to measure the heights of a large number of

boys (for instance, in a city) of a particular age, the numbers would distribute as is illustrated in Fig. 9.6.

This shape of curve is so predictable that it has been analyzed mathematically. It is used to handle numbers in a particular sense. We state that something is statistically significant if it falls in the upper or lower 5% of this curve. Again, this is proof of nothing; it merely means that the difference from the mean is worth investigating to see if there is a logical or other explanation that would cause us concern. It suggests that we can place bets, but does not answer our questions.

A bell curve is used in two distinctly different ways. First, this information can tell us whether an individual is in the outlying regions of the normal distribution, as in the case of the growth curves. In this case we use the term *standard deviation* to describe the variability. A little less than 2/3 of any population falls within the range of one standard deviation from the mean. (The source of the number comes from the mathematics and is not important here.) The size of the standard deviation can vary. For instance, if the color of one variety of flower ranges from white to deep

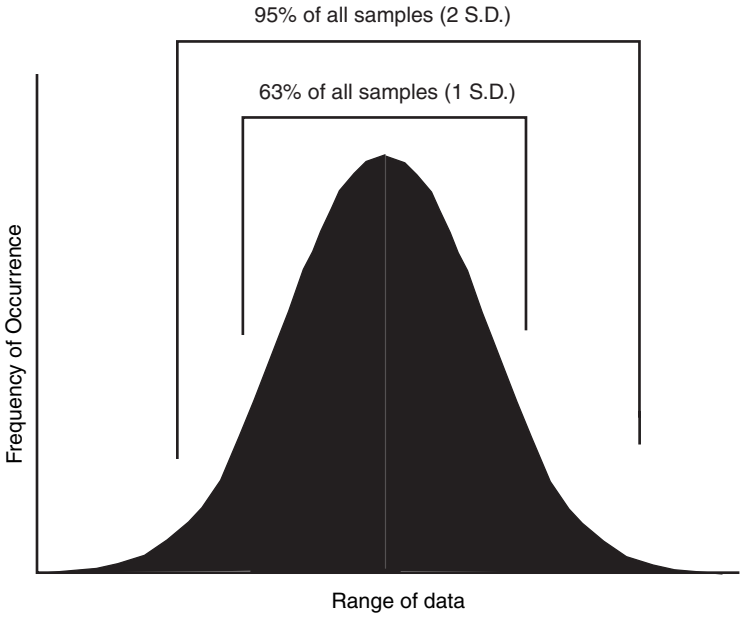

Figure 9.6. The bell curve or Poisson distribution. Fig. 9.5 is based on this distribution. All that it says is that, for any normal variable (height, weight, amount eaten at a meal, temperature in Chicago on June 20, number of points a given basketball player scores in a game) the most frequent numbers are those closest to the mean, with the probability of the more extreme numbers being much less. Thus, if you flipped 1000 coins, you would not be surprised if you got 500 heads and 500 tails, 499 heads and 501 tails, etc; but you would be very surprised if you got 300 heads and 700 tails. The Poisson distribution predicts how often this might occur. For a statistician, if the number falls in the upper or lower 5%, the result is considered significant and the situation should be explored for an explanation. As in Fig. 9.5, this result can occur simply by chance

red, with all possible shades in between, the standard deviation of the color intensity will be broader than for a variety of flower that ranges from pink to middling red (Fig. 9.7).

Also, because of the way that the number is calculated, the higher the number of samples, the more accurately the curve can be calculated. For instance, if you took a middle-school class that had 17 boys in it, the mean height (the total of all

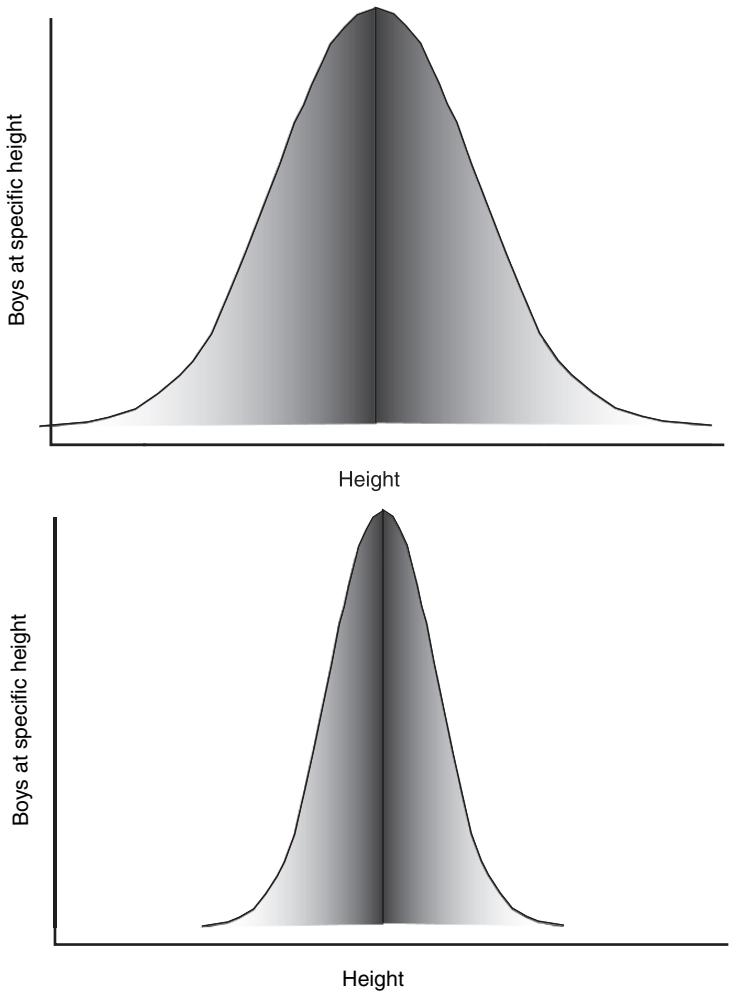

Figure 9.7. The range of variation can change from situation to situation. For instance, the range of weights of new pencils is much smaller than the range of weights of strawberries. However, in both situations a Poisson distribution exists, and it is possible to find the abnormally large and small individuals. Here, the range of heights of boys in the upper figure, perhaps from a highly heterogeneous community like a big city, is much greater than in the lower figure

heights divided by the number of boys) might be different from the median height (the height of the 9th boy), and the range between the tallest and the shortest might be very broad or very narrow. This is illustrated in Table 9.3 Let's compare those figures:

In the first school, the heights are very evenly distributed. The mean height of the boys is 54 inches (4'6") and the standard deviation is 5, meaning that about 2/3 of the boys (11 boys) will be between 49 and 59 inches. In the second school, the mean height is the same, but there is a greater range and variation in heights. From this school, we would expect that about 2/3 of the boys would be between 46 and 62 inches. In the third school, the mean height is 3" more than in the first school. Should we worry about the 46" child or the 62" child in school 1? These are falling out of the range of expected. On the other hand, they are within the range of the children in school 2, while the 49" child is abnormally short in school 3 but not in schools 1 or 2, and the 65" child is much more out of the range of school 1 than for schools 2 and 3.

Obviously, if we counted only three boys (choose randomly any three from the table) we would get a wide variation in numbers. Perhaps not so obviously but just as true, if we counted all the boys in one city, with the intent of comparing them to the boys of another city, we would expect the means and the standard deviations to be much more reliable (predictable) and useful for us. In brief: if I compared the height of my children to the height of my brother's children, it is very likely that there would be a difference, and we would not make much of an issue of it.

Table 9.3. Variation of heights of boys in different schools

Boy number	School 1	School 2	School 3
1	46	42	49
2	47	43.5	50
3	48	45	51
4	49	46.5	52
5	50	48	53
6	51	49.5	54
7	52	51	55
8	53	52.5	56
9	54	54	57
10	55	55.5	58
11	56	57	59
12	57	58.5	60
13	58	60	61
14	59	61.5	62
15	60	63	63
16	61	64.5	64
17	62	66	65
MEAN	54	54	57
STANDARD DEVIATION	5.0	7.6	5.0

If, however, my child's height was way out of the range of the whole city, I might well have reason to be concerned.

The second use of the bell curve is to compare populations. For instance, one might compare the heights of 11-year-old boys in Sweden, the United States, Thailand, and Liberia. In this case we would get individual population distributions as above, but the curves for each population might be shifted to the left or the right, and we could state that the mean height of Swedish boys is greater than the mean height of Thai boys. We can also plot the means of each population. If we have enough populations to compare, these means of populations would also distribute on a bell curve. We could use that information to ask if a population is on an outlier from the normal range. For instance, if the bell curve illustrated in Fig. 9.7 represented not individuals but different populations, and we found that the mean for a group of boys on a small island was in the lower 5%, we could ask if nutritional, genetic, or other factors were responsible for their short stature. This is the standard arrangement for an experimental set-up; we try to structure the experiment so that everything that we can think of is equivalent, except the one factor that we hypothesize is important, such as cigarette smoke. We then compare two populations, one in the presence of smoke and the other without smoke. This procedure is described in the next section.

Pollsters use the population statistics (the second group). When they say, "candidate A is ahead 54% to 46%, with a margin of error of 7 percentage points," what they mean is, "If we repeated this same survey, with equivalent numbers of respondents, in 90% of the surveys, candidate A would win by 8% or less. However, in 5% of the surveys candidate A would lose, and in 5% she would win by a larger margin."

EVALUATING DATA: EXPERIMENTATION

Since it is so difficult to interpret causality from correlation, we attempt to do experiments. Experiments have several properties. They are founded on a hypothesis that itself derives from certain observations. Then the hypothesis is stated in such a way that a means of testing the hypothesis is suggested. This test is structured in such a way that other interpretations are ruled out, and one remains as both logical and supported by the results. The hypothesis may at some point be proven wrong, when new evidence comes to light, but at present it is the best interpretation available. An example of the latter point would be Lord Kelvin's calculation of the age of the earth. Knowing the amount of heat absorbed by the earth from the sun, and the physics of cooling of a sphere, he made the assumption that the earth had split from the sun and since that time had been cooling. Based on the temperature at the surface of the earth and deep within it, he came up with a figure of between 20 and 400 million years. He later settled on the lower figure. However, he did not know about decay of radioactive materials, which are naturally common within the earth. This decay, like an atomic bomb or a nuclear reactor, generates substantial amounts of heat. When radioactive decay

is factored into the calculation, we come up with a much more reasonable few billion years.

Two relatively striking experiments can illustrate the meaning of testing a hypothesis: Einstein's theory of relativity, and the dietary origin of pellagra. In the first instance, it was well understood since Newton's time that light traveled in straight lines. Mirrors, prisms, and lenses depend on this property, as does the "peek-a-boo" game so beloved by infants of all cultures. Einstein's theory, however, predicted that light could be bent by gravity. How could one test the idea, since the force and the bending would be miniscule in any laboratory experiment? It became evident that the force and the bending would be measurable if the gravity was produced by a relatively large heavenly body. The sun would be adequate, but it is so bright that one cannot see the stars behind it. However, an eclipse provides a different situation. Since the trajectories of the stars were well known, one could calculate very precisely their positions at any moment. At the height of a solar eclipse, if gravity could bend light, a beam of light that should have shot past the earth might be bent to hit the earth, and we ought to be able to see a star that should be behind the sun but that seemed to "pop ahead of itself" in its trajectory (Fig. 9.1). After its angle moved past the sun, it would resume its normal trajectory. Thus the solar eclipse of May 29, 1919 was eagerly awaited. If a star "speeded up" then slowed its trajectory, there was no logical explanation of such a bizarre event other than Einstein's hypothesis. Thus the confirmation of the hypothesis was heralded as a major event.

There are many other examples, including some of the most famous of the Renaissance, but in the 19th and 20th Centuries, two of the most dramatic were those of Pasteur demonstrating the bacterial origin of fermentation and Goldberger's demonstration that pellagra was a nutritional disease rather than a contagious or genetic disease. Pellagra is an ugly disease caused by the lack of the vitamin niacin. This was not known when a physician from the Surgeon General's office, Joseph Goldberger, went to the U.S. South in 1912 to investigate the spread of the disease. At the time, two theories were popular: first, that it was an infectious disease, as Pasteur and Koch had demonstrated during the previous century for other diseases; and, second, propounded by Eugenicists (see Chapter 31, page 405) that it resulted from inferior, less resistant, genetic background.

What he first did was to travel through the South, taking notes on everything that he could. (See the discussion on control experiments directly below, page 134). The first common factor that he noted was that poor people, including also prisoners and orphans, had a very monotonous diet, consisting of cornbread, molasses, and pork fat. The poorer the people were, the more likely they were to get pellagra. There was another peculiarity: in institutions such as prisons, the prisoners had pellagra, while the guards did not. To Goldberger, this meant that pellagra was not infectious but rather was caused by something such as a problem with the diet. (Remember that at this time there was no concept of such things as vitamins, and the idea that food could vary in quality was generally scoffed at.) Finally, in 1915, he went to one prison that had its own farm and fed prisoners a varied

diet. He arranged that prisoners would be pardoned if they tried the diet common to other prisons. During the experiment, these prisoners lived with the others inmates and no effort was made to prevent infection. Nevertheless, after a couple of months, they began to show signs of pellagra. Meanwhile, Goldberger and his team tried to infect themselves with pellagra. As Walter Gratzer describes, conducting "filth parties," the eight researchers "injected themselves with blood from severely affected victims … rubbed secretions from their mucous sores into their nose and mouth, and after three days swallowed pellets consisting of the urine, feces and skin scabs from several diseased subjects." They did not contract pellagra. He finally convinced prison wardens and heads of orphanages to try to feed their wards more varied diets and, where he succeeded in convincing them, the pellagra disappeared.

Goldberger had first made a critical observation: though pellagra was common in prisons, it affected only the prisoners, not the guards. Thus he had a means of falsifying the first hypothesis. Since prisoners and guards were in daily contact, though they ate separately, contagion by personal contact seemed unlikely. Since the guards ate balanced meals while the prisoners were given only fatback (dried and salted fat from the back of a hog) and cornbread, he hypothesized that the limited diet was the problem. He therefore conducted his first experiment by feeding a balanced diet to children in two orphanages, who were on a similar diet and likewise suffered from pellagra. Their pellagra disappeared within weeks. Since the curative material in the diet was not known and in any case it would have been completely unethical to conduct the converse and more definitive experiment, to produce pellagra by restricting the diet, he based his argument primarily on the first half of the experimental protocol. However, his results did not convince his opponents. The Eugenicists persisted in their belief that weaker constitutions were involved, and the proponents of infection refused to relinquish their preferred hypothesis. Therefore Goldberger undertook his colorful and unusual efforts to falsify the hypothesis that infectious agents caused pellagra. This left standing the alternative hypothesis that it was the result of poor nutrition. Nevertheless, it took 20 years before the mood in the country came round to accepting the idea and actively promoting better nutrition. Part of the issue was a weakness in the perceived logic: the idea that foods contained materials of magical properties in vanishingly small amounts was not readily absorbed, but also most societies find it difficult to abandon old ideas when new evidence contradicts the ideas. As the sociologist Leon Festinger noted, when the world fails to end on the day predicted by a cult, the response of the cult members is not to abandon their prediction but to proclaim that the failure of the world to end is proof that their prayers were heard and that their belief is valid.

Other factors played a role, as we have seen in other circumstances. Here, for instance, the world had finally accepted the idea of infectious disease, and all focus was on infectious diseases. Second, the general perception of the social order, abetted by growing awareness of the implications of evolutionary theory, made the idea that poverty could cause disease, as opposed to disease being a

natural aspect of poverty, relatively unpopular. (To a politician, the suggestion that poverty could cause disease would lead to the conclusion that the state should invest effort in overcoming poverty, which could be politically and financially costly.) Finally, Goldberger was a Yankee and perhaps even worse, an immigrant Jewish physician from New York, working in a South still hurting from the Civil War and deeply suspicious of ulterior motives behind Yankee science. Thus Evidence, Logic, and Falsifiability, all brilliantly achieved by Goldberger, do not necessarily and obviously win. To many, the logic had not truly been established.

THE CONTROL EXPERIMENT

The purpose of a laboratory experiment is to restrict the range of possible interfering factors. For instance, if you wished to interview 100 people to determine the potential outcome of an election, you would get very different results if you stood in front of a school in a prosperous neighborhood, a store that sold computer parts, a courthouse, a dress shop, an ethnic restaurant, or a fast food restaurant; and you would likewise get different results if you stood in front of a supermarket at 8 AM, 10 AM, noon, or 8 PM. You would have to determine how each variation of location or time affected the distribution—sex, age, income level, race or ethnicity, etc.—of people who came by and thereby biased the numbers that you collected. Are young, well-off mothers likely to vote the same way as computer jocks? It becomes very complicated. There are many means by which statisticians try to address such issues, but there is another, more subtle type of bias, deriving from the experimenter's interest in the results. For instance, Mendel (see Chapter 13, page 175) classified peas as yellow or green. What did he do about yellowish-green peas? When several researchers tried to replicate his exact experiments, they got results that were similar, but never as close as Mendel's results to the ideal 3:1 ratio predicted by Mendel's Law. We suspect that, once Mendel realized the principle that he later espoused, he unconsciously classified the yellow-green peas in such a way that it tended to improve his numbers. We know this problem well. Almost any educational reform seems to work, because the greater involvement of the teachers in the new project of itself improves the learning environment, no matter what the change is. A physician or patient who feels that a medicine should have an impact will notice details of the condition that signal improvement even though by more physical criteria there is no improvement. This is so common that it is called the "placebo effect," meaning that patients will claim improvement, or physicians will see improvement, even if the "pill" is only sugar.

One way to defeat this bias of self-interest is to do a double-blind experiment, in which neither the experimenter nor the patient—if it is a medical experiment—knows which subject has received the experimental treatment and which has not. The way that it works is as follows: The size and shape of the nucleus of a cell can be an indicator of its health. The nucleus might be large and elliptical or small and rounded. I know how I conducted my experiment and therefore know which preparations were exposed to which chemicals. If I expect the nuclei to shrink

and become rounded, I am very likely to classify intermediates as small in the samples exposed to the chemical. I therefore number my samples with jumbled or meaningless numbers and ask a graduate student or a technician, who is experienced enough to make the measurements but who has not participated in the experiment, to examine the preparations, do the classifications, and give me the results. I may ask more than one person to do this. The technician comes back to me with the results, "This is what I got for the percent of nuclei that were small and round: In sample A, 80%; in sample B, 30%; in sample C, 5%; and in sample D, 50%." I go back to my notes and come back to say, "That's great! Sample C did not receive any of the drug, sample B had a low dose, sample D, an intermediate dose, and sample A, the highest dose!" (And I may give the technician a hug.)

Note the description of sample C. We call it a control experiment. In this preparation, we try to do everything that we do to the other preparations, except for one crucial step or component. Expressed differently, we attempt to assure that, between our test situation and our control situation, every variable that we can think of is the same except for the variable that we wish to test. We need to do this to convince ourselves that the mere act of conducting the experiment has not produced the results. For instance, surgical experiments involving removal of organs always include "sham operations" as controls, since anesthesia and the wounding of an animal will always produce dramatic responses by themselves, whether or not the organ has been removed. In other situations (in most experiments in biology, chemistry, or physics today), we rely on the readings of instruments to tell us things that we cannot see. Thus, a common means of measuring whether something has been taken up by an animal is to give the animal a radioactive form of the material and then to look for the radioactivity in a process called scintillation counting. In scintillation counting, the radioactivity collides with a material that will fluoresce (glow) when hit by radioactivity, and we can measure by machines the light that is emitted. You have seen this kind of fluorescence in wrist watches that glow in the dark. The machine gives us a count of the number of flashes of light given off in a second or a minute and therefore, ideally, the amount of radioactivity taken up. However, it gets more complicated: if our solution is too acid, the fluorescent material will be destroyed, and we will get no glow. If the solution is too alkaline, it will glow spontaneously without radioactivity. If we have too much water present, the radioactivity will be absorbed by the water and will not hit the fluorescent material. If the solution is cloudy, the light may be emitted but blocked or reflected away before the sensors detect it. Sometimes, especially if we are working with extremely small levels of radioactivity, variations in the natural background radiation, from cosmic rays and other sources, may be a substantial portion of the total radioactivity. If the radioactivity is too high, two emissions may occur simultaneously and be counted by the machine as only one count. Or the sensors may not be functioning properly and may produce too few or too many counts. Therefore, since we cannot actually see the differences, we need to have a sample in which everything is the same—the same amount of sample and the same reagents, processed in exactly the same way, except that there is no radioactivity

present. This is one of our controls, and we call it the negative control, meaning that it should give us the lowest possible value, which is not necessarily zero. We will also prepare a positive control, in which we will add a known amount of radioactivity to the mixture, to be certain that our counter detects it as it is supposed to. This positive control will allow us to establish that the fluorescent material has not been degraded by acid or other means, that the solution is not cloudy when it is counted (in a light-tight chamber), and that the counting machine is functioning properly. All good experiments include appropriate controls.

In fact, editors of journals look closely at the reported controls to verify that they are the best controls for the experiment and that they have been done correctly. When one looks at the speeches of former Nobel laureates, it is striking how often their prize-winning research started by their wondering if the control to the experiment was really adequate and then turning to verification of the control. One example was the work of Joseph L. Goldstein and Michael S. Brown. They discovered the proteins that bind cholesterol in the blood, which are now commonly known as HDLs and LDLs. They received the Nobel Prize in 1985. They had been attempting to determine what effect cholesterol had on cells in culture, and they added the cholesterol to the culture medium the same way everyone else did. They found that the cholesterol had no effect. Unlike other researchers, however, they reasoned as follows: cholesterol is very insoluble in water. How much actually dissolved in the water and reached the cells, as opposed to, for instance, sticking on the walls of the pipettes and flasks? Therefore they attempted to locate the cholesterol, in essence asking if the control (no cholesterol) was an effective control for the presumed experiment (cholesterol added). They found that the cholesterol was neither on the flasks nor in the cells but rather on proteins in the solution in which the cells were held. These proteins were what are now known as HDL and LDL.[5] The next year, the Nobel Prize was awarded to Rita Levi-Montalcini and Stanley Cohen, who similarly doubted their controls. They had been attempting to isolate and identify vanishingly small amounts of a factor important for the development and maintenance of nerves, called nerve growth factor (NGF). As the name indicates, it causes nerves to grow and survive. Since they could not get enough to analyze, they attempted to determine if it was protein or nucleic acid by digesting it with enzymes that specifically attacked proteins or nucleic acids. If the enzyme worked, the biological activity of NGF would disappear. In the experiment in which they attempted to destroy the nucleic acids, the activity did not disappear.

---

[5] Incidentally, it is not difficult to remember which is the "good cholesterol" and which is the "bad cholesterol". Some proteins are designed to bind one and only one type of molecule, such as cholesterol. They typically bind the molecule tightly and cannot hold much of it. Other proteins sop up all sorts of fats rather loosely, like a paper towel. They can hold a lot of cholesterol, but it can come off very easily, again like a dripping paper towel. Also, fats are lighter than water, and float. LDLs (bad cholesterol) are "low density lipoproteins," since they contain a lot of loosely-bound fat (cholesterol) that can easily be released to cause problems. HDLs (good cholesterol) are "high density lipoproteins," in essence heavier, because they do not carry much fat. However, they really hang on to the cholesterol, and keep it out of harm's way. It is much easier and more reliable to understand than to memorize!.

However, as a control, they tested whether the enzymes themselves would affect the growth of nerves. To their astonishment, the nucleic acid-digesting enzyme also stimulated nerve growth! In fact, the enzyme preparation was contaminated by nerve growth factor, revealing a new, rich source of NGF. With the new source of NGF, they were able to purify and characterize NGF, leading to the prize.

CAUSALITY

The purpose of experimentation is to establish a relationship that, by logic, timing, or sequence indicates causality. The relationship is "when the sun shines, snow melts". What we mean by timing or sequence is that, everything else (such as temperature) being equal—that is, we have a type of control experiment—the snow begins to melt after the sun begins to shine. Therefore, the melting of the snow could not have caused the sun to shine, and it is more reasonable to hypothesize that the shining sun causes the snow to melt.

The essence of setting up the experiment is to phrase the question in an appropriate manner. A well-phrased question will suggest an experiment. Here, "Why does the snow melt?" does not suggest an obvious means of finding an answer, but "The sun provides heat. Is this heat enough to melt the snow?" suggests that one could compare the amount of heat produced by the sun in various circumstances, including different heights of the sun in the sky, different levels of cloudiness, and different starting temperatures, to the rate of snow melt. It also suggests an experiment: On a bright sunny day, if one blocked the sun's rays by shading a patch of snow, would this affect the rate of melting? A well-designed experiment will address both ends of the statement, "if and only if": the result (snow melting) will occur if the sun shines and will not occur if the sun does not shine (only if the sun shines). We will see this result, of course, only if all other variables are equal and accounted for (the control experiment)—that is, the temperature is constant and slightly below freezing, the humidity is the same for both preparations, the wind is the same or absent, and there are no other sources of heat or cold. Even the weight (or more exactly, pressure, which is weight per square inch) on the snow should be the same. Ice skates work because the weight of your body on the narrow blade creates sufficient pressure to melt the ice, and glaciers move because their weight causes the bottom of the ice to melt. But the important element of the well-designed experiment is that we can test the "if" portion ("If this condition occurs, then we will see this result,") and the "only if" portion ("If this condition does not occur, then we will not see this result.") The "if" constitutes the Evidence and the "only if" constitutes the Falsification of our "ELF" rule. We also need the Logic, an explanation of the mechanism that led us to hypothesize that the result would follow from the condition.

An example of why the principle of falsification is so important is an incident involving the closure of a hazardous waste site. The community where the site was located had access to a scientifically-trained consultant for advice on the closure, and the community members were concerned that the approximately three-year

closure process would create further hazards to their health. They wanted the city and state to provide cancer screening during the process, and asked the consultant for his opinion. The consultant recommended against the screening for the following reasons:

- The development of cancer is a very slow process, often taking twenty years. If the closure process caused any cancer, it would never be seen during the monitoring process.
- The criterion would not be that a specific person in the community developed cancer but that the frequency of cancer was unexpectedly high in the community.
- The neighborhood near the landfill differed by ethnicity, diet, smoking habits, recreation, age distribution, and drinking habits from nearby neighborhoods. All of these could affect the frequency of cancer. It would be very difficult to establish a baseline or comparison to this community.
- The neighborhood had a relatively high turnover of population. Even if it were possible to recognize an increased frequency of cancer, one would have to know at what age the people were exposed to the landfill, and where they lived before and after they were exposed to the landfill. Perhaps a specific age would prove more susceptible; perhaps there was a problem in the community in which many of them were born and spent their childhood before moving to this community.
- The community was bounded by a major highway, which produced a lot of car fumes, all of which could affect cancer rates.
- The community was downwind of a major airport, so that fumes from planes landing and taking off wafted into the community. This also could affect cancer rates.

As you can see, the problem that the community faced included both the fact that it would be exceedingly difficult to define a suitable control by which to evaluate the results, and the inability to eliminate alternative hypotheses, such as the hypotheses that increased cancer was caused by fumes from the highway, by favored foods of a specific ethnic community, by heavier smoking by community members, or many others. This again turns to the issue that one cannot prove a hypothesis true; one can only eliminate competing hypotheses. The community finally agreed that a combination of monitoring air quality (looking for cancer- and asthma-causing chemicals or particles) and local hospital admissions for asthma attacks, with the ability to stop operations if any value was too high) provided a more immediate and direct response to the hypothesis that the closure of the landfill would produce disease-causing conditions.

THE CHARACTERISTICS OF GOOD SCIENCE: ELF

Although the scientists that we cite today as having performed classical experiments, a careful description of what scientists do awaited the writings of Carl Popper in the early 20th C and many authors since then. To summarize the idea, science studies mechanisms that determine how the world functions, and they do so by collecting Evidence, using Logic to generate a hypothesis of how one element or process

affects another, and then designing experiments to attempt to Falsify the hypothesis that they have made. The essence of good science is the ability to structure a question so that a good, definitive experiment can be performed. The answer cannot be wishy-washy, but must definitively rule out an opposing hypothesis. This culture is embedded in every function that scientists perform. All the following quotes are taken from conversations with scientists: professor to student, grant reviewers, manuscript reviewers. "Never do an experiment unless you have a table or a figure in mind." [This comment refers to the fact that an experiment must test a hypothesis. The table or figure makes the comparison of the control to the experiment and demonstrates the falsification. The comment also emphasizes the importance that scientists give to figures and tables.] "Is this application hypothesis-driven?" "These are shotgun experiments." "This is just a fishing expedition." [These last two comments refer to a paper or a grant application that is not based on an underlying hypothesis but, rather, is simply trying chemicals or processes that are known in the hope that something finally falls out of the study. Contrary to the interpretation given by Francis Bacon, this random collection of data is highly disfavored by working scientists.] "Does this contribution have a sufficiently biochemical or mechanistic focus to justify publication?" [A presentation that merely presents new data but does not have experimental justification to argue mechanism will be dismissed as "purely descriptive" and not accepted for publication.]

Ultimately, as is discussed on page 123, the experiment and the control must satisfy what we might describe as a "truth table," for which it becomes apparent that there is no absolute proof for the truth of a proposition, but one can prove the falsity of another one: Table 9.2 is in essence a truth table.

Thus the criterion becomes the true meaning of the expression "if and only if". One can declare a relationship IF, when A occurs, B always occurs AND IF, when A does not occur, B never occurs. If the tide rises when the moon is aligned with the sun and falls when the moon is opposite the sun, we can declare a relationship between the position of the moon and the tide. There may be exceptions, but we should be able to explain them without violating our original proposition. For instance, a fierce storm may push water into a bay, creating what appears to be a high tide at a different time, but we can determine that the wind and decreased atmospheric pressure from the storm are sufficient to account for the extra water. When we correct our measurements for the weather, we find that the movement of the tides has not really changed.

CLASSICAL EXPERIMENTS

Antonie von Leeuwenhoek

We will describe below a few classical experiments for which, in addition to the ones above, you should identify the elements of evidence, logic, and falsification. Do not underestimate the importance of logic or the social and intellectual situation of the time! As is argued in Chapter 14, page 191, DNA was not considered a likely

repository for genetic information until experimental evidence had generated the logic that almost required it to be the genetic material. An even more spectacular example was Antonie van Leeuwenhoek's development of the first microscope and his observation of micro-organisms in materials such as water, the scum that he scraped off of teeth, and other media. He was very excited by his findings, and described them in terms that he knew:

(Plaque).... I then most always saw, with great wonder, that in the said matter there were many very little living animalcules, very prettily a-moving. The biggest sort had a very strong and swift motion, and shot through the water like a pike does through the water; mostly these were of small numbers."
"In structure these little animals were fashioned like a bell, and at the round opening they made such a stir, that the particles in the water thereabout were set in motion thereby…And though I must have seen quite 20 of these little animals on their long tails alongside one another very gently moving, with outstretched bodies and straightened-out tails; yet in an instant, as it were, they pulled their bodies and their tails together, and no sooner had they contracted their bodies and tails, than they began to stick their tails out again very leisurely, and stayed thus some time continuing their gentle motion: which sight I found mightily diverting."

Remember that, before this time, 1675, no one had ever seen an organism smaller than the eye could resolve. Although it was clear that diseases could propagate, bad air ("malaria") or vapors from water were considered likely causes. Also, Leeuwenhoek was an extremely skilled craftsman, and the lenses that he made were far superior to those of anyone else. Thus, when he attempted to publish his findings in the proceedings of the prestigious London-based Royal Academy of Sciences, he received what may have been the worst rejection letter ever written:

"When I observed for the first time in the year 1675 very tiny and numerous little animals in the water, and I announced this in a letter to the Royal Society in London, nor in England nor in France one could accept my discovery, and so one still does in Germany, as I have been informed."
In a letter, Hendrik Oldenburg, the Secretary of the Royal Society, London, wrote the following to Antoni Van Leeuwenhoek, Delft, Holland, 20th of October, 1676:
"Dear Mr. thony van Leeuwenhoek, Your letter of October 10th has been received here with amusement. Your account of myriad 'little animals' seen swimming in rainwater, with the aid of your so-called 'microscope,' caused the members of the society considerable merriment when read at our most recent meeting. Your novel descriptions of the sundry anatomies and occupations of these invisible creatures led one member to imagine that your 'rainwater' might have contained an ample portion of distilled spirits–imbibed by the investigator. Another member raised a glass of clear water and exclaimed, 'Behold, the Africk of Leeuwenhoek.' For myself, I withhold judgment as to the sobriety of your observations and the veracity of your instrument. However, a vote having been taken among the members–accompanied I regret to inform you, by considerable giggling–it has been decided not to publish your communication in the Proceedings of this esteemed society. However, all here wish your 'little animals' health, prodigality and good husbandry by their ingenious 'discoverer".

There was little concept of the ability of a lens to magnify, and no concept of microscopic life; a group of prestigious scientists simply could not accept the idea that an unseen world existed. Of course it did, and improvements in microscope manufacture and many confirmations of van Leeuwenhoek's findings finally won out. This rejection of not-obvious new findings has often been repeated. Similar disbelief greeted August Semelweiss' demonstration that sterilizing a delivery room with carbolic acid eliminated childbed fever, of which many women died after

giving birth. Of course his demonstration carried the baggage of suggesting that other obstetricians were responsible for contaminating their patients. In another example, the Nobel Laureate Rosalind Yalow is one of many scientists who has been known to open a speech with a slide showing the letter rejecting her first submission of her findings. That letter is particularly revealing. The editor noted, "The experts in this field have been particularly emphatic in rejecting your positive statement…" because it contradicted then current theory.

Here, as in the case of Vesalius (Chapter 30, page 406), it is always unacceptable to defer to the wisdom of sages, but we do. As the father of modern physiology, Claude Bernard, noted in the mid 19th C, "When the fact that one encounters opposes the reigning theory, one must accept the fact and abandon the theory, even though the latter, supported by impressive names, is generally adopted." To be totally fair, however, the rejection letter received by Yalow did emphasize the conviction of the reviewers that the conclusions were not sufficiently justified—meaning that the reviewers wanted to see more definitive and unequivocal experiments. Thus, they may have been pig-headed, but they still relied on the triumvirate Evidence, Logic, Falsification. The scientists of the Royal Academy, however, did not seriously consider the possibility that the evidence was real, and they certainly did not propose any attempt to falsify it by attempting, for instance, to document the distortions that a piece of glass might produce.

Let us therefore look at four classical experiments that are in one sense related, in that they form a sequence documenting that germs are living creatures and that they can cause disease. The experiments are as follows: Redi's demonstration that maggots do not generate spontaneously; Pasteur's demonstration that spoiling of broths was caused by bacteria; Koch's establishment of rules for identifying disease caused by bacteria; and Snow's tracing of cholera to a living, water-borne organism.

In the 17th C, about the time that van Leeuwenhoek was building his microscope, the world below the resolution of the human eye was still totally unknown. There was not even a magnifying glass. Humans can recognize two dots or lines as being separate down to a limit of approximately 0.2 mm (about 3/64"—look at a good tape measure). Fly eggs are approximately 0.1 mm in diameter. Thus when maggots appeared on meat that had been left hanging in open-air markets, the presumption was that the maggots spontaneously generated on the meat. This was in line with what seemed to be obvious at the time. Although trees and grain crops obviously grew from seeds, molds and many other plants seemed to spring up out of nothing, frogs would appear suddenly after a rain, and small worms would appear in standing water. Life must arise spontaneously from inanimate objects and materials.

Francisco Redi

Francisco Redi was unconvinced, and he knew that the maggots, while seeming to appear magically, eventually turned into flies, and flies were always flying around hanging meat. He therefore conducted the following experiment.

In his book “Experiences around the generation of the bugs”, Francisco Redi wrote: “ I put in four flasks with wide mouths one sneak [snake], some fish of river, four small eels of Arno river and a piece of calf and I locked very well the mouths of the flasks with paper and string. Afterward I placed in other four flasks the same things and left the mouths of flasks open. Short time later the meat and the fishes inside the open flasks became verminous, and after three weeks I saw many flies around these flasks, but in the locked ones I never seen a worm ”. http://utenti.quipo.it/colettisb/ipertesto-redi/redi/redi-exp.htm

Later, facing the argument that the air inside the flasks would go stale, he improved the experiment. He took many different kinds of meat and covered them with a cloth fine enough to allow air to circulate but too fine to allow flies to pass. He exposed the covered meat alongside meat that was uncovered at different times and temperatures, and waited to see what happened. As you might expect, no maggots appeared on the covered meat, whereas they did on the uncovered meat. He then uncovered the original covered meat and demonstrated that, once uncovered, it too would generate maggots. He therefore concluded that maggots did not generate spontaneously but instead were produced by the flies that landed on the meat.

There are many elements to this experiment that are worth noting. The most obvious is that he used several kinds of meat and several conditions. This setup permitted him to make a general statement rather than a specific one, such as, “Maggots do not generate when a dead eel is covered with cloth and hung outside on a rainy day in June.” He seeks a more general principle, that maggots do not spontaneously generate under any condition. Thus he varies the conditions so that he might argue, “Maggots do not generate under any of the fifteen conditions that I have tested. Therefore I can extrapolate my findings to any other situation that others might wish to test.” In other words, he has made a hypothesis that can be tested by others.

Second, though not obvious in the paragraph quoted, he repeated the experiment several times to confirm that he always got the same result. In other words, the result was not a quirk of something that had happened that day. For instance, the wind might have been too high for flies to land, or the flies might have liked to lay eggs on all meats other than that of a snake.

The third and most critical issue is that he has established a control experiment, a preparation of meat that was, as far as he could tell, identical to the experimental meat with the single exception that flies could reach it. Thus his conclusion was, “all other things being equal, the ability of flies to land on the meat makes the difference between maggots and no maggots”. The “all other things being equal” phrase is important, since very slight differences can change the outcome of an experiment. The very act of injecting a drug into an animal may frighten it enough to cause its behavior, growth, ovulatory pattern, or other feature to change, notwithstanding the effect of the drug. Thus it is necessary to inject a saline solution or other harmless solution into a control animal, so that the controls have experienced equivalent stress. You can find many more examples yourself. A good exercise would be to imagine what other variables could have influenced Redi’s experiment.

Pasteur and Pouchet

Redi's results were sufficiently convincing that by 1651 the English physician William Harvey, himself an elegant experimentalist who demonstrated by both logic and evidence the circulation of the blood, could declare "Ex ova omnia" ("Everything [comes] from eggs"). However, the connection between insects and bacteria, and between bacteria and disease, was not yet established. By the mid-19th C, this argument was still open, and there were several practical consequences. These included questions as to the origin of diseases such as cholera, to be discussed below, and problems in France concerning spoilage of foods and disease among the grape vines of the wine industry. The issue of what caused spoilage of food was of such practical and theoretical importance that the French Academy of Sciences offered the Alhumbert Prize for the best proof of whether or not bacteria, or putrefaction, would generate spontaneously. Pasteur had demonstrated by this time that boiling milk or food would delay putrefaction, but many believed that the process of boiling had damaged either the foodstuff so that bacteria would no longer thrive on it, or had damaged the air so that bacteria could not survive. According to this argument, bacteria could generate spontaneously but needed an appropriate environment to grow. Thus prizes were offered for the proof of whether or not life could spontaneously generate. The leading contenders were Louis Pasteur and Félix Archimède Pouchet. Although in retrospect there were some elements that Pasteur did not understand, such as the ability of some bacteria to sporulate (go into a sort of hibernation, during which they resist heat and other killing agents) and in fact he was very lucky in the choice of his preparation, the experiment that he designed was elegant. The hypotheses were the following:

1. Bacteria arose spontaneously but required undamaged (uncooked) food to grow.
2. Bacteria arose spontaneously but required something from the air to grow, and whatever was in the air could be destroyed by heat.
3. Bacteria arose from other bacteria that could easily contaminate even a clean preparation.

Pasteur extended the third hypothesis by assuming that bacteria could be airborne and could drift in on breezes. However, they were heavier than air and would settle out in still air. He had already established that if a meat broth was boiled and the flask sealed, then it would remain uncontaminated. This experiment was very similar to one Lazzaro Spallanzani had done in 1767, in which he had demonstrated that small animals could not generate in boiled flasks unless and until the flasks were opened to the air. Others, however, protested that either the broth or the air had been damaged by the boiling and could no longer support the generation of life (hypothesis 2). Pasteur therefore constructed an elaborate flask that had an S-shaped loop (Fig. 9.8). The flask was open to the air which, it was already known, could diffuse even without a breeze. However, the narrow neck of the flask blocked breezes, and the air would penetrate only by diffusion. At this slow pace, bacteria would settle into the lower part of the loop. He then took some boiled meat in its juice, basically a bouillon, and let some cool in an ordinary beaker and in his flask. The bouillon in the beaker quickly became infected, proving that

Figure 9.8. The flask that Pasteur used for his famous experiment. See text for explanation

the boiling had not destroyed its ability to support bacteria. However, the bouillon in the special flask did not become contaminated. One could still argue that there was some problem with the interaction of the boiled bouillon and the air above it. Pasteur therefore tipped one of his flasks so that the bouillon reached the low point in the neck, where he hypothesized that the bacteria had settled, and then sloshed it back into the main part of the flask. Within a couple of days, it was apparent that the tipped flask was contaminated, whereas the one that had not been tipped was still clean. Pasteur was awarded the prize, and one of the flasks that he prepared over 150 years ago is still on exhibit, still open to the air, and still uncontaminated. Today thousands of laboratories studying bacteria and cells in culture use a modification of this experiment, called a Petri dish after the designer, Julius Richard Petri, to grow their cells. The Petri dish works on the same principle as Pasteur's flask, allowing potential contaminant bacteria and fungi to settle out. Although air can freely circulate in the dish, it remains uncontaminated because potential contaminants settle out by gravity (Fig. 9.9). Note that what most people tend to assume is the correct position of the Petri dish is upside-down.)

The logic and structure of the experiment is best illustrated by Table 9.4 It is a matter of some curiosity that the subject was sufficiently interesting to scholars that Pasteur's experiments were carried out in the context of a contest to prove or disprove the existence of spontaneous generation.

Koch's postulates

There were many further demonstrations that Pasteur was correct, and scientists and physicians turned to seeking the causes of infectious disease. In 1890 Robert Koch published a list now called Koch's Postulates of what would be required to argue that a microorganism caused a disease (note what is Evidence, Logic, and Falsification, and what constitutes the "if and only if" criteria):

1. The organism must be found in all animals suffering from the disease, but not in healthy animals.
2. The organism must be isolated from a diseased animal and grown in pure culture.
3. The cultured organism should cause disease when introduced into a healthy animal.

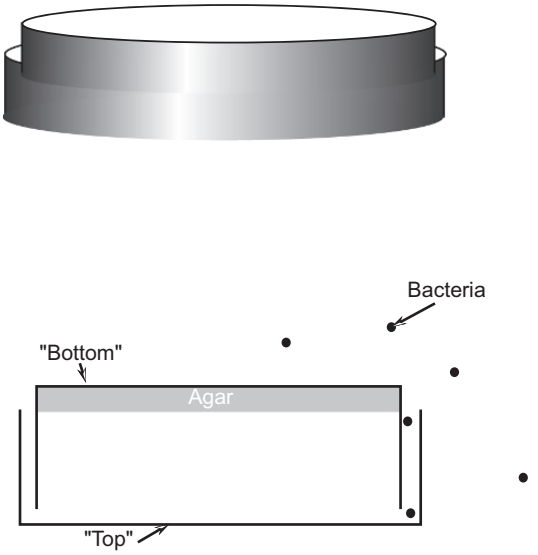

Figure 9.9. Upper A modern Petri dish. Note that its proper position is "upside down". (Lower) The mechanism by which a Petri dish functions. Note that this is the same mechanism as the Pasteur flask

Table 9.4. The logic of the Pasteur experiment

Aspect	Demonstration	Explanation
Hypothesis	Bacteria arise from other bacteria, as opposed to hypothesis of spontaneous generation	Broths in which bacteria can grow are contaminated only when bacteria have access to them; boiling kills them and sealing the flask prevents new ones from entering.
Evidence	Bacteria do not grow in flask that has been boiled and sealed.	According to hypothesis of spontaneous generation, bacteria should be able to generate (falsifies or nullifies this particular hypothesis, except in special case of damage to medium (broth))
Logic (if)	Bacteria do not grow in flask even if air can be exchanged, but without air currents	Addresses qualification of damage to air; nullifies hypothesis that quality of air is essential
Falsification (only if)	If flask is tipped to wash in dust that is fallen, bacteria will grow	Addresses qualification of damage to medium, since medium has not been changed; nullifies hypothesis that quality of medium has changed
Conclusion	Bacteria do not arise by spontaneous generation	Started with two hypotheses, of which one has been completely nullified; the hypothesis that bacteria arise only from other bacteria stands until a new hypothesis can be tested and shown to nullify it.

4. The organism must be reisolated from the experimentally infected animal.

Not all of these are fully attainable. Some organisms can be found in healthy animals and cause disease only rarely; other organisms cannot be easily grown in culture. Nevertheless the rules have validity.

Sir John Snow and Cholera

A final classic experiment using the logic of science was Sir John Snow's demonstration that cholera was an infectious disease. Cholera was a devastating disease. Essentially a severe diarrhea, but one that could drain so much fluid from a person that it could kill a person by dehydration in a few hours, it would break out in cities and spread rapidly, killing hundreds or even thousands in the space of a few weeks. There were two major hypotheses as to what caused it: "Effluvia," by which was meant odors or gases escaping from infected patients, who thus poisoned the air for healthy individuals; or biological or chemical agents in the bodies of the victims. Snow therefore looked at the logic of what the evidence was telling him:

1. Cholera traveled from city to city at the same rate that people traveled. Thus, if cholera broke out in Rome or Paris, it would not reach London faster than the time that it took stage coaches or boats to reach London.
2. If cholera came from another country, it would be seen first at a seaport. It would not appear suddenly in the Midlands of England.
3. It would break out on ships, but only if the ships came from cholera-infected countries. If cholera had broken out in Rome, a ship coming from Rome might develop cholera, but cholera would not appear on a ship coming from Stockholm.

All of this evidence suggested that cholera was transmitted from person to person, but it still did not resolve the two hypotheses. But then Snow encountered a new patient who had no personal contact with any cholera victim. However Snow, an astute observer, learned that the patient had received clothes from a recent cholera victim. This was not unusual; if someone died young, the clothes were often recycled. Snow then refined his hypotheses:

4. Hypothesis: If cholera is passed by effluvia, then all persons in contact with patient should get cholera, and those in contact with only the clothes should not.
5. Hypothesis: If cholera is passed by liquids, then those in contact with the liquids should get cholera, whether or not the patient is present. Since in a medical situation one does not usually have recourse to a lab experiment, Snow re-examined his evidence to see if the evidence supported one or the other of these hypotheses.
6. It might be disgusting, but it was an issue that physicians could note. In the later stages of cholera, patients have vomited and lost through diarrhea so much that their digestive tracts are empty, and anything further that is lost is clear and watery, and may not even be noticed. In such conditions, Snow considered the possibility that the last clothes the patients wore had not been washed after their death.

7. Those who washed more frequently, such as workers who handled mud and clay and other materials that they would want to get off their hands, did not get cholera.
8. Nurses and doctors, who washed frequently, did not get cholera even though they worked with cholera patients. This evidence suggested to Snow that the disease was spread not by the air but by liquid excretions from the body. Since these excretions normally went into the sewers, Snow then turned his attention to the distribution of disease and the distribution of water in the cities. The disease tended to be clustered, with some exceptions that caught Snow's attention:
9. In the city of Manchester, those getting water from a well near a leaky sewage pipe got cholera.
10. In Essex, there was an outbreak in one district served by a single well. A washerwoman living in that district was the only one who did not get cholera, but she used water from another well.
11. In Locksbrook, a landlord who lived elsewhere was accused, during an outbreak, of providing poor water to his tenants. To prove that it was safe, he drank water being delivered to those buildings. He subsequently died of cholera.

With this information in hand he looked at a new epidemic in London. Ultimately 300 people died during the outbreak. Snow plotted, on a map, the residences of all the victims, and saw that they all clustered around one source of water, known as the Broad Street pump. A brewery nearby was not involved in the epidemic, but the brewers had their own source of water for the beer and did not use the pump. Six cases were in a different neighborhood but, when Snow got a map of the water pipes, he realized that the people in that neighborhood also got their water from Broad Street.

From this evidence Snow argued that the source of the epidemic was the pump. Furthermore, he argued, it was not a chemical contamination, since a chemical would be expected to dilute out with time and thus cause less disease; but the severity of the epidemic was continuing, suggesting that the cause could reproduce. Therefore the cause was likely to be biological, in other words, a germ. With that information, he finally did his experiment. He removed the handle from the Broad Street pump, rendering it inoperable. People in the neighborhood had to go to the pumps in the surrounding neighborhoods to get their water. Within a few days the epidemic was over.

Note what he proved and what he did not prove. He demonstrated that it was likely that the contamination came from one pump, that it was carried by the water, and that it was biological in origin. He did not identify the organism. In fact, the germ that causes cholera is extremely difficult to grow in the laboratory and, though widespread, does not often cause cholera. But its existence is one of the reasons we chlorinate water. Snow did not prove anything, in the sense that he had no true experiment to falsify his hypothesis. He falsified competing hypotheses, leaving his hypothesis standing. It was enough to signal him to intervene, and the success of his intervention convinced everyone of at least the pragmatic value of sterilizing water.

In this digression from the subject of evolution, we have looked at the issues of what constitutes evidence, what we mean by multiple independent means of verification, what constitutes an adequate control, and the complexity of interpreting data that must be assessed by statistical comparisons. We have considered the relationship between evidence and the logic of the experiment, and have seen that the "if and only if" basis of experimental logic is the same as the ELF logic emphasized throughout this book. However, the necessity of the logic may not be apparent either to scientists or to the public until other information becomes available. Often, the interest in a question is driven either by new findings or by new social concerns.

REFERENCES

Best, Joel (2001). *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists.* University of California Press.

Desrosières, Alain (2004). *The Politics of Large Numbers: A History of Statistical Reasoning*, Trans. Camille Naish, Harvard University Press.

Huff, Darrell and Geis, Irving, (1993) How to Lie With Statistics (Paperback reissue) W. H. Norton, New York

Tijms, Henk (2004). *Understanding Probability: Chance Rules in Everyday life*. Cambridge University Press.

Dobell, Clifford, ed. 1960. Antony van Leeuwenhoek and his "Little Animals", Dover Publications, N.Y.

Festinger, Leon, 1964. When Prophecy Fails: A Social and Psychological Study of a Modern Group That Predicted the Destruction of the World, Harper Collins Publishers, New York.

http://www.ucmp.berkeley.edu/history/leeuwenhoek.html (biography of Leeuwenhoek from U. of CA, Berkeley)

http://www.euronet.nl/users/warnar/leeuwenhoek.html

STUDY QUESTIONS

1. Choose any claim, advertisement, or other propaganda that you find in the media. Present the arguments in the construct of "if and only if" or "Evidence, Logic, Falsification". After your presentation, do you still accept the claim?
2. Choose any news item describing a scientific advance reported in a newspaper, magazine, or on television. Trace the source of the story as far as you can, and analyze the presentation in terms of "if and only if" or "Evidence, Logic, Falsification". Can you identify the controls? What was falsified?
3. Which of the experiments described in this chapter do you consider to be the most convincing? Why?
4. People often say that obesity is a "metabolic problem". From a statistical standpoint, what would you say might be a reasonable indication that the problem was truly medical?

   Which of the arguments presented in the previous chapters as supporting the theory of evolution meet the criteria described in this chapter? Which do not? What would be required to complete the arguments?