CHAPTER IV
THE HYPOTHETICO-DEDUCTIVE METHOD --
CONJECTURES AND REFUTATIONS

It is certainly not the least charm of a theory
that it is refutable.

--Nietzsche

A fall in the pit, a gain in the wit.

--Chinese proverb
(quoted by Mao)

The central features of the hypothetico-deductive theory of scientific method may be diagrammed as follows:

1. Start with a scientific problem
2. Propose a testable theory as solution to the problem
3. Derive an observable consequence from the theory
4. Do an experiment to check the truth of the consequence
   - If the prediction is correct, test another consequence
   - If the prediction is wrong, propose another theory
5. If the theory passes stringent tests, start with a new scientific problem
   - etc.
There have been many accounts of the hypothetico-deductive method, dating at least from Whewell's *Novum Organon Renovatum* in the mid-nineteenth century. But perhaps the most sophisticated account of it to date is Popper's and we will present a version similar to his here. Let us now discuss each step of the process in turn.

a. What Is a Scientific Problem?

According to Popper, the scientist can never begin with a completely empty mind. Regardless of what the topic may be, the scientist, like all of us, begins with a motley collection of ideas, some clear, some confused, some true, some false. In the course of living in the world or thinking about it, the scientist encounters various types of problems. Here are some typical ones:

(i) **Problems arising from violated expectations.** A common sort of scientific problem arises when something surprising or unexpected occurs and we wonder how or why it happened.

An important problem for early astronomers was the following: In general, celestial bodies, such as the sun, moon and stars, move across the sky in smooth arcs. However, it was discovered that the planets wander around the sky irregularly. Can one describe precisely how the planets move and explain why they move differently from the other heavenly bodies? Plato called this the problem of the planets. Ptolemy, Copernicus, and Kepler each offered a different solution to it.

Here is another example of a scientific problem caused by violated expectations: In 1896 Becquerel found that a batch of photographic plates which had been carefully stored in black paper were fogged. According to the best scientific knowledge available at the time, only visible light or x-rays could expose photographic plates. What could have happened? Becquerel finally began to suspect that the fogging was caused by an unusual rock he had used as a paper weight. And it was thus that he discovered radioactivity. Later Madame Curie showed that the rock contained radium.
(ii) Problems arising from a quest for deep explanations. Even if the scientist is lucky enough to discover a generalization which seems to have no exceptions, he or she is still faced with a problem: What causes the regularity? Why do things happen just that way?

For example, early astronomers asked why the sun rose every day in the east. Some said it was because the sun moved in a circle around the earth. Later this geocentric theory was replaced with a heliocentric theory. In either case, a further question arose: What caused the sun (or earth) to move? According to Aristotle, there was a Prime Mover. Later people suggested a law of circular inertia, saying a wheel would move forever if there were no friction. Newton explained the regular motion in terms of linear inertia and the force of gravity.

There are many other cases in which the problem is to explain a regularity. Bohr wondered why the wavelengths of the spectral lines of hydrogen should fit the simple mathematical formula discovered by Balmer. Mendeleev and other chemists of the late 19th century wondered why the elements should arrange themselves so nicely into a Periodic Table.

By the end of the 18th century, after the work of Boyle and Charles, everyone knew that gases expanded on heating. But why? Caloric theorists said that heat was a fluid which flowed into gases and as a result they took up more room. Kinetic theorists said heat was kinetic energy and hot gases expanded because their molecules moved faster. Both sides agreed on the regularity to be explained, but they offered competing explanations of it.

(iii) Problems arising from a quest for unity. As a science develops, a new sort of problem often arises: Can one find a unified theory which covers two or more domains which have previously been treated separately?

For example, for a long time organic chemistry (which deals primarily with covalent compounds) and inorganic chemistry (which is mainly concerned with ionic compounds) were considered to be quite distinct fields. At this time people believed that naturally occurring organic compounds, such as urea, could not be synthesized in the laboratory because they
contained a vital life force. However, today's theories of chemical bonding apply equally well to inorganic and organic materials.

Before Galileo, it was held that terrestrial bodies and celestial bodies obeyed different laws. Galileo (and latter Newton) gave a unified account of the motions of all bodies.

A pressing problem in physics today is the search for a unified field theory--a theory which would successfully combine relativity theory and quantum mechanics. Psychologists are looking for a unified theory of learning. Behaviorists can account for some kinds of learning; cognitive psychology provides explanations for other types of learning. But one would like to find a single theory which covers all instances of learning.

In each of the three types of scientific problem-situation discussed above, the problem arises out of a rich background of information and expectations. New scientific theories are invented when scientists are faced with a problem: Why did my old theory or set of unconscious expectations fail? What causes this regularity which I have observed? Can I unify these two branches of science?

Science begins from puzzlement about existing bodies of knowledge. It does not arise out of a vacuum.

b. Where Do Hypotheses Come From?

We have described the various sorts of problems which trigger scientific inquiry. Later on we will describe how scientists criticize and test the hypotheses which are offered as tentative solutions to these problems. But where do the hypotheses come from? How do scientists discover them?

Early philosophers were optimistic about the prospects of describing a method for discovering true theories. As we have seen, Bacon and other inductivists thought that through careful observation and systematic use of his tables one could easily arrive at the solution to scientific problems. Descartes and other rationalists thought that a systematic analysis of our clear and distinct ideas would provide the answers.
Most modern philosophers of science would say that there is no recipe for discovery. All the scientist can do is guess at the answer. Some conjectures will be "happy guesses" as Whewell described them; others will turn out to be dead wrong. It's all a matter of trial and error.

Popper has compared the growth of science to biological evolution. Mutations occur by chance—we can't predict what new variations will occur. But natural selection will filter out those who are not adapted to the environment. Likewise for science. People make up all sorts of crazy hypotheses. But tests will weed out those which do not match reality. Quality control is insured by careful testing procedures, not by censorship of new ideas.

Or to propose another analogy: Science is like a University with an open door admissions policy. Anybody can enter (because who is to say ahead of time who will succeed), but examinations will quickly screen out those who are not doing the work.

Still, you may be wondering, how do scientists ever dream up the hypotheses which will be put to the test? Much more research needs to be done on this question, but I can provide a few suggestions.

First of all, scientists often use analogical reasoning to generate hypotheses. I will not try to give a formal characterization of reasoning from analogy but simply provide an example.

The organic chemist Kekulé was trying to figure out the structure of benzene. He knew its formula was $C_6H_6$ (i.e., it consisted of six carbon atoms and six hydrogen atoms) and he knew the valence of carbon was four and that of hydrogen was one. (The valence of an atom describes how many bonds it must form.) But he was unable to think of a structure which both satisfied these constraints and corresponded to the known chemical properties of benzene. He tried linear models, branched chains, multiple bonds, but all in vain.
Then one evening in 1865 Kekulé was dozing in front of the fire. As he later described it, "... atoms were gambolling before my eyes ... [in] long rows, sometimes more closely fitted together; all twining and twisting in snake-like motion. But look! What was that? One of the snakes had seized hold of its own tail, and the form whirled mockingly before my eyes. As if by a flash of lightining, I awoke."^1

And thus it was that Kekulé literally dreamed up the ring structure for benzene which we still accept today:

```
  H   \  C   \  C   \  H
  \   C   C   C   H
  \   H   C   C   H
  H   C   C   C   H
```

The argument from analogy goes as follows:

Chains of atoms are like snakes.
Snakes are normally open-ended curves, but they can form a circle.
Chains of atoms are normally open-ended but maybe they too can form a circle.

Such an argument carries very little weight. There is no reason for expecting atoms to behave like snakes. And of course the story does not really explain why Kekulé made the comparison to snakes in the first place --presumably the darting flames of the fire suggested the dancing serpents/atoms.

The story of Kekulé's discovery gives us no reason whatsoever to believe his hypothesis is true. But it does give us some idea of why he first thought of the ring structure.

---

Kekulé recorded his moment of discovery. There are many other fables about discovery episodes within the history of science, many of them not very well-documented. Some historians have argued that Harvey's discovery of the circulation of the blood was influenced by the Copernican theory. The analogy in this case would go as follows:

The sun is the source of heat and life in the solar system.
The heart is the source of heat and life in the body.
The planets revolve around the sun.
Maybe the blood moves around the heart.

This is a very rough argument indeed. Even if we grant the analogy between the heart and the sun, why should the planets and the blood be similar?

There is an important moral to be learned from such stories: The pattern of reasoning which leads to a new hypothesis is not important—it may be based on dreams, mystical experiences, weak analogies or what have you. The origins of the idea are irrelevant; what is crucial is how well the scientist's hunch stands up to testing. But not all speculative theories are even capable of being tested. Let us now turn to the problem of testability.

c. Which Theories are Testable?

As our account above makes clear, the solutions to problems which scientists propose start out being mere hypotheses or conjectures. When they are first proposed, we have no particular reason to believe them true. Furthermore, these hypotheses tend to be rather bold and far-reaching. This is because typical scientific problems all require as solutions theories of high content. Consider Problem Type 1: To explain why our expectations are violated, we need a theory which accounts both for the exceptions and the normal states of affairs we had expected. For example, a good answer to the problem of the planets' irregular motions would also explain the sun's regular motion.
To turn to Problem Type 2: Trying to give a deep explanation of a regularity (such as the Balmer formula for hydrogen spectral lines) generally results in a conjecture which has many other consequences as well (such as a formula for the spectral lines of sodium).

As for Problem Type 3, it is clear that a unified theory will have more content than either of the separate fields. And generally such a theory will have lots of new consequences as well. (For example, the unified theory of chemical bonding covered not only traditional organic and inorganic compounds, but a whole new domain of organo-metallic compounds, such as hemoglobin.)

Although they are bold conjectures, from the very beginning scientific conjectures do have one very important property in their favor: they can be tested by means of experiments. If one of our conjectures is false, it is realistic to hope that we will eventually discover its erroneous nature.

The claim that the characteristic aspect of scientific theories is their vulnerability to possible refutation has been particularly emphasized by Karl Popper. Early in this century Popper began to work on the problem of demarcating science from what he called pseudoscience because of his dissatisfactions with the psychanalytic theories of Freud and Adler, and Marxist theories of society. He was particularly concerned to articulate the important differences he felt separated the theories and methods of physicists, such as Einstein, from what went on in the fields of psychology and sociology at that time.

Now both the physicists, on the one hand, and the Freudians and Marxists, on the other, were trying to give naturalistic explanations of phenomena. Both depended on data collection and abstract theoretical constructs. What was it that made physics so much more intellectually satisfying?

Popper was quite sure that it wasn't a question of the truth of the two sorts of theory. After all, Einstein's theory of special relativity was a highly speculative conjecture. When the Eddington eclipse expedition
set out in 1919 to see if light actually was bent in the gravitational field of the sun as the theory predicted, no one was at all sure of how the experiment would turn out. And if Einstein's theory did pass this and other tests, then this meant that Newton's theory was wrong! So it hardly seemed to be a question of truth.

No, Popper decided, it was not the reliability of Einstein's theory which impressed him. Quite the contrary, it was the boldness of the theory -- the way it made precise claims which a well-executed experiment might refute. (And considering the surprising nature of its claims, we might well be inclined to guess that the theory would be refuted!) The theory "stuck its neck out", as it were, and practically invited an experiment to shoot it down.

Popper contrasted this situation with Freudian or Marxist theory. These systems could come up with an explanation of practically anything. If an industrial strike received front-page billing, it was a sign that the contradictions in capitalist society were reaching a crisis point. If the wage dispute did not get much coverage, it was due to a conspiracy by the management-dominated press. If someone dreamt of cigars or other long-shaped objects it was a sign of interest in male genitals. If someone did not dream about cigars, etc., it was a sign of intense, but repressed, interest in male genitals. No matter what happened, these theories could give an account of it.

Popper then realized that the very feature which Freudians and Marxists thought gave their systems explanatory power (namely, their ability to assimilate any state of affairs) was actually the source of their scientific inadequacy. Theories can only explain certain events by ruling out others! If no conceivable state of affairs could ever discredit a theory, then none of its so-called successes are of any significance.

Popper concluded that science was different from pseudoscience in two important ways:

(1) Theories in science are highly testable ones; those in pseudoscience are not.
(2) The methods which scientists adopt (especially severe testing) are designed to eliminate false theories as quickly as possible; pseudoscientists try to protect their theories from refutation.

Let us now discuss the precise requirements that a theory must satisfy in order to be testable in Popper's sense.

(i) The Logical Requirement. As we have seen in Chapter II, statements of the form "Some A's are B's" cannot be refuted by any report involving a finite number of instances, but universal generalizations, be they affirmative or negative, can be.

A necessary condition for a theory to be refutable is that it be logically possible to contradict it by a finite conjunction of sentences which describe particular instances.

Thus, "Everybody loves his/her mother" satisfies the logical requirement. It could be contradicted by the following finite conjunction: "β is α's mother and α does not love β." The sentence "Everybody loves someone" does not satisfy the logical requirement. It could be undermined by the following report: "α does not love β, nor γ, nor δ, nor ...." but it is impossible in principle to prove in this way that there is no one whom α loves.

(ii) The Empirical Requirement. Having the proper logical form is not sufficient to insure that a hypothesis is scientifically testable. "All repressions are seated in the libido" satisfies the logical requirement but, as it stands, it is not subject to experimental test. How exactly are we to recognize a repression? And even if we could, how could we tell whether or not it is seated in the libido?

Contrast the following sentence which has the same logical form:

"All samples of iron have a melting point less than 2000° C."

This universal generalization is subject to test. We can easily determine whether a sample is iron or not through chemical analysis. (We might use the potassium thiocyanate test, for example.) And there are also a
variety of reliable procedures for measuring melting points.

The contrast in the above two cases suggests the following requirement:

A testable theory is one which is inconsistent with at least one finite conjunction of observation sentences.

An observation sentence is one the truth or falsity of which can easily and reliably be agreed upon by any observers in the vicinity.

Whether a theory is testable or not depends on the technology and state of scientific development available at the time. Before the invention of the mass spectrograph, "All atoms of an element have the same weight" would not have been considered testable because as yet there was no way to determine the weights of individual atoms. What counts as an observation sentence also changes with the development of instrumentation and with new theoretical developments. For modern scientists, "This sample is oxygen" and "This is an electron track" are considered to be observation statements. In an earlier era they would not have been. "This sample is a gas which supports combustion" and "This track in a cloud chamber curves towards the positive plate" might have been used instead, if the identity of the gas or of the particle was still in question. The truth of observation statements cannot be decided with certainty; even so, members of the scientific community can tentatively agree in their judgments about the truth of observation statements.

In judging whether a particular sentence is testable or not, one must apply the above requirements with a certain amount of tact and common sense.

For example, "There is a golden mountain somewhere" is clearly not open to refutation by a finite number of observations. However, the qualified existential claim, "There is a golden mountain between Toronto and Buffalo" can be tested and is in fact false. Since there are a finite number of places the mountain could be, it is possible to perform an exhaustive search.
Likewise, the claim "Some copper is brittle" looks like it is not open to refutation by a finite observation report. However, if the claim is accompanied by a recipe, "To make copper brittle, place a thin sheet of it for three days in a nuclear reactor where the neutron flux is . . ." it becomes testable.

Sometimes it is not completely clear which sentences should count as observation sentences. For example, in most cases we can easily agree on whether a particular person is bald or not. Yet there are also borderline cases of baldness. Generally, scientists adopt the following policy: If any observation term such as bald seems controversial or vague, they replace it with a more precise, less troublesome expression, such as "... has fewer than three hairs per square inch in the area delimited by the ears, eyebrows..."; in this way, they make it easier to test their theories.

A caveat to the reader concerning terminology: Following Popper's practice, we have used the terms testable, refutable, and falsifiable interchangeably.

The words falsifiable and refutable remind us that on the basis of finite experimentation it is logically possible to prove a universal generalization false, but we can never prove it to be true. However, to some people they have a misleading connotation which I now wish to dispel.

When we say a certain glass is breakable we mean that it would in fact be physically possible to break it should we choose. However, when we say that a claim is falsifiable or refutable we do not mean that it is in fact false. We do not mean that a counter-example actually exists and that if we are lucky we will find it.

Quite the contrary. Many true sentences are falsifiable. (For example, "All sodium salts burn with a yellow flame" is true as far as I know, yet it is falsifiable.) And many false sentences are unfalsifiable! (For example, I believe the sentence "There exists a Philosophers' Stone which turns base metals into gold" to be false. However, one cannot refute it with any finite set of experiments.)
For these reasons, the word *testable* may be preferable. In any case the central point is this: To say that a sentence is falsifiable/refutable/testable is simply to assert that if it were false, there would exist at least one finite conjunction of observation sentences which would be inconsistent with it.

Note that according to the Popperian definition of testability, the sentence "Some women are bald" is not testable because no finite conjunction of observation sentences could ever disprove it even if it were in fact false.

Also note that according to Popper's demarcation criterion all scientific systems are testable in his sense and no pseudoscientific ones are.

**Exercises A:**

1. Which of the following are testable in Popper's sense? (Do they satisfy both requirements?) If they are testable, give an example of an experimental report which if true would falsify the hypothesis. You may put these statements within a modern scientific context and make any reasonable assumptions about the available stock of observation statements.

(a) Nitrates are water-soluble.
(b) The class of organic compounds called esters have a strong fruity smell.
(c) Sometime, somewhere there has been or will be a stone which will turn manure to gold.
(d) Light has a maximum velocity.
(e) Every metal has a melting point.
(f) Of the various mutants available in the genetic pool, only the fittest survive in the long run.
(g) In a capitalist economy, violent revolution is inevitable.
(h) Any car with an eight-cylinder motor has a higher rate of gasoline consumption than any car with a six-cylinder motor.
(i) All solids vaporize to some extent in a vacuum.
(j) Only the good die young.
(k) If the sky gets completely clouded over, then within five minutes, either it will rain or the sun will shine.

(l) The probability that a new born baby is a girl is 0.51.

2. The class of observation statements depends on the technology and instrumentation available at the time. Thus, which theories are testable may change over time.

For each of the following advances, give a statement which seems likely to have been untestable before the discovery, but testable afterwards.

(a) Microscope
(b) Rockets capable of landing on the moon
(c) Thermometer
(d) X-ray machine
(e) I.Q. test
(f) Rohrschach inkblot test

3. (a) Give an example of a testable theory which has been refuted.
(b) Give an example of a testable theory which has not been refuted.

4. Criticize the definition of observation sentence given above. (Recall that observation sentences are the way we check theories against reality.) For example, which of the following sentences might have satisfied the definition at the time specified? Do we want them to serve as the basis for testing scientific theories?

(a) This woman (who is having what we would call an epileptic fit) is possessed by devils. (1690 in Salem, Massachusetts)
(b) This immigrant child is dumber than this native child. (1978 in the East end of London)
(c) The sun moves and the earth stands still. (1600)

5. Test the adequacy of Popper's testability criterion as a means of demarcating science from non-science by applying it to one or more actual examples. Pick examples you are familiar with. (For instance, consider
cognitive dissonance theory in psychology, the theory of reaction intermediates in organic chemistry, or the biological theory of evolution.) According to the refutation criterion, are these part of science? Also try to find refutable elements in the theory of Tarot cards, handwriting analysis, or the Ouija board, etc.

6. Generate some other important differences between science and pseudo-science. Should any of these be used to replace the refutability criterion? Or to supplement it?

7. A slightly mad philosopher claims the following: "The world only exists when someone is looking at it, smelling it, touching it, or perceiving it in some way. By perception we create the world. We can make it disappear by not looking at it." Could you use tests to argue that human perception is not the cause of the existence of external objects?

d. Stringent Testing

Once we have succeeded in inventing a testable theory which seems to provide a suitable solution to our scientific problem, the next step is to try to make sure that our proposed solution is not false. To do this we need to test our conjecture (and to test it severely.) If a theory is false, we would like to eliminate it as quickly as possible. On the other hand if a theory passes the most stringent tests we can devise, this gives us good reason to hope that it is true -- or at least near the truth.

In the above section we described the minimum logical and empirical requirements a test should satisfy: it should be designed so as to make it possible to turn up an observable counterexample to our theory, should one exist.

But in addition, we want our tests to be stringent, i.e., to provide efficient means of finding counterexamples. Here are some of the features which make a test stringent.
(i) Extensiveness. Consider the hypothesis, "All schizophrenics lack Vitamin V." One way to test this conjecture would be to sample the class of people who have sufficient Vitamin V in their diet. If any turn out to be schizophrenics, we will have refuted the universal generalization. Other things being equal, looking at a sample of 200 such people is a better test than looking at only 20. More extensive tests are more likely to uncover any exceptions to the claim.

The stringency of a test increases with increasing sample size; i.e., extensive tests are more stringent.

(ii) Precision. Suppose the conjecture to be tested says, "The half-life of any radioactive isotope is unchanged by external conditions."

Let us suppose we have methods of measuring the half-life of an isotope (i.e., the time required for one half of any sample to decay) and we investigate the effect of strong magnetic fields on the rate of decay.

Suppose our half-life measurements are only 10% accurate. Then we might collect data, such as the following:

<table>
<thead>
<tr>
<th>Half-life of sample of ordinary radium:</th>
<th>Half-life of sample of radium in magnetic field:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1600 years (± 160 years)</td>
<td>1700 (± 170 years)</td>
</tr>
</tbody>
</table>

This data does not refute the conjecture. Furthermore, using this method of measurement we could never refute the conjecture even if it were false, unless the true difference in decay times were more than 10%. (If the discrepancy were 20%, we could hope to detect it; if it were 2%, this measuring device would never reveal that the conjecture was false.)

The stringency of test increases with increasing precision of determination of the property-in-question.

Of course, if the conjecture does not make very precise claims, there is no need to make our tests very precise.
(iii) **Severity.** In designing tests of scientific hypotheses, we should always remember that our goal is to find a refutation of the conjecture if one does in fact exist. (A _true_ hypothesis will, of course, have no refuting instances.) Often we can use our general background knowledge to predict where the theory is likely to be vulnerable.

For example, when Kohlberg put forward a theory about the development of moral reasoning in children, he was well advised to test it on children from Turkey and Taiwan. We might expect a theory developed on the basis of experience with kids in Boston to fail when applied to children from quite different cultures and religions. (As it turned out, the Kohlberg theory passed this severe test.)

Similarly, theories about the universality of the Oedipal complex should be tested on aborigines, and theories about language learning on deaf and blind children. Theories about geological change and biological evolution should be tested, where possible, by data from other planets. Physicists know that theories often fail under conditions of high energy or high velocity; and often processes at the micro-level violate generalizations which work well with medium-sized objects. For this reason physicists want to build ever bigger accelerators for smaller and smaller particles.

The general procedure for designing a severe test is as follows: The hypothesis under test always makes a series of claims. For example, the claim "All arsenic compounds are poisonous" says that both soluble and insoluble arsenic compounds are poisonous. It also says that both yellow and green non-poisonous substances are free of arsenic. (Don't forget the contrapositive!)

According to our background information, some of these claims sound less plausible than others. For example, since we know that many poisons have to be digested in order to act, we may decide that insoluble arsenic compounds are less likely to be poisonous than soluble ones.
A severe test is one which tests the least plausible claims of a theory. In our example, given our background theories about the relationship between solubility and poisonous character, we should start testing by looking at insoluble arsenic compounds. If the conjecture passes this severe test, we will then look at the class of soluble arsenic compounds.

Other things being equal, severe tests, i.e., tests of the least plausible claims of a conjecture, are more stringent than less severe ones.

Note that our appraisal of the severity of tests depends on the background information available at the time. Consider the two claims: (a) "All yellow non-poisonous substances are free of arsenic" and (b) "All green non-poisonous substances are free of arsenic". Which domain should be investigated first if one wishes to perform a severe test of the original conjecture? Recall that a counter-example to the original conjecture would be a non-poisonous arsenic compound. So if we think green substances are more likely to contain arsenic than yellow ones, we should sample the domain of non-poisonous green substances. If we know nothing about the typical color of arsenic compounds, however, or if we have reason to believe that color is not correlated to chemical composition, we would judge the tests to be equally severe. (As a matter of fact, many arsenic materials are yellow or black, so there may be a slight preference for a test of yellow non-poisonous substances.)

(iv) Variety. In a severe test we deliberately stack the deck against the conjecture. We use our background information to choose a test sample we think more likely to yield a refutation.

But what if our background information is faulty? To pursue the arsenic example, what if we were wrong in thinking that many arsenic compounds are yellow or that soluble materials are more apt to be poisonous? What if it turns out that the counter-example to our conjecture is a green soluble non-poisonous arsenic compound? If we had performed only the tests thought to be the most severe, we would never have found it!
Due to the fallibility of our background belief system, it is often wise to make the sample domain as varied as possible. To be sure, what counts as variety also depends on our background system. Should we test arsenic compounds stored in square bottles as well as those stored in round ones? Very probably not. The effort required to vary our sample in every conceivable way would be prohibitive.

Within reason and other things being equal, the more varied the sample domain, the more stringent the test. (The background information system determines what counts as relevant variety.)

(v) Crucial Experiments. Another way to probe the vulnerability of a hypothesis is to compare its predictions with those of a plausible rival conjecture. If hypothesis A predicts $P$ and rival hypothesis B predicts $\neg P$, checking on whether $P$ or $\neg P$ is the case will allow us immediately to eliminate one alternative. Contrary to what is sometimes stated a crucial experiment does not prove the truth of the undefeated hypothesis because there may exist more alternatives which we have not yet thought of.

For example, according to the Copernican theory, Venus should wax and wane like the moon. The Ptolemaic system, on the other hand, predicted that Venus should not exhibit extremely different phases at different times. This conflict between the two rival cosmological systems was noted by Copernicus in 1543. However, it was not possible to conduct a crucial experiment without a good telescope. In 1610, Galileo observed that Venus did have phases and so the Ptolemaic system was refuted.

This crucial experiment in no way established the truth of the Copernican heliocentric theory for in 1588 Tycho Brahe had proposed a geocentric system which also gave the correct predictions concerning Venus. The next order of business was to design a crucial experiment between the Tychonic and Copernican system.

Crucial tests are only stringent when the rival hypothesis is a fairly plausible one (as judged against background knowledge).

The more plausible the rival conjecture to the hypothesis in question, the more stringent is a crucial test between them.
For example, no one would have thought it necessary to design a crucial test if the only rival were an ad hoc hypothesis to the effect that Venus shone by its own light but periodically varied its luminous area from crescent shaped to circular!

One common method of collecting data which is erroneously thought to provide support for a hypothesis involves no risk to the hypothesis at all. We might describe this practice as performing tests of zero stringency. Strictly speaking, such data collecting constitutes no test of the conjecture at all for its design makes it logically impossible for a refutation to appear.

Here is an artificial example of no-risk data collecting. Suppose a little kid brags, "I can toss a coin so that it lands heads up every time." You smile and say, "Show me." The kid tosses up a whole handful of coins, carefully selects a few for your inspection and triumphantly announces, "See, they're all heads!" "But what about the rest?" you say. "Well, I didn't want to bore you," says the kid, "so I just took a sample. Isn't that the way scientists operate?"

The methodological error here is blatant. If one only looks at heads, one can never refute the claim that all tosses turn up heads. But in other contexts the same error is less obvious.

For example, after teaching scientific method for a number of years, I caught myself reasoning as follows: I observed that all of my close friends who blinked a lot and tipped their heads back when looking at me wore contact lens. I then started investigating other people who behaved similarly and sure enough I nearly always found independent evidence that they were wearing contacts. Sometimes I asked them. Other times I would see a lens holder in their purse or bathroom, etc. I soon jumped to the following conclusion:

"All people who wear contact lens blink a lot and peer down their noses when they look at you."
This conclusion was obviously too strong, given that I had done only an informal study on a very small sample. But I did think that my experience justified a more modest statement:

"All contact lens wearers whom I have met blink a lot, etc."

What was not clear to me for quite some time is that none of my observations had served as a test for either conjecture. For I had always begun my observations with people who blinked! Given this choice of sample domain, I could have investigated all the blinkers and peerers in the world and never have found a counter-example to my conjecture -- not because there weren't any, but simply because it was logically impossible for my method of testing ever to uncover them.

The child in the story above may have been deliberately cheating when it failed to report any tails thrown. I was not consciously cheating, but no-risk data collecting is always a methodological cheat, for it blinds us to possible counter-examples to conjectures.

As Bacon's story of the votive offerings of sailors illustrates (see Chapter I), no-risk data collecting plays an important role in superstitious reasoning.

* * * * *

The basic Popperian pattern of problem -- tentative solution -- testing is widespread. It is the method used by a child learning to walk, a drunk fumbling for the right key, and perhaps even by a rat learning a maze.

But in its details, the growth of science is quite different from learning in other domains. Scientific problems are in a sense "bigger". Scientists seek theories which cover a wide range of phenomena and provide deep explanations of them. And perhaps the most important difference between science and every day problem solving is the extensive use of stringent tests in science. If I solve the practical problem of crossing a stream by throwing a log over it, I do not test my solution stringently -- I don't drive a truck over it to see if it will break. If it will hold my weight (plus a little more), that's good enough.
But when a scientist proposes a solution to a theoretical problem (not an engineering problem), it is tested mercilessly. Parochial solutions to problems will not do in science. Here the aim is a theory which works perfectly everywhere! Stringent testing is the only method appropriate to that aim. Most conjectures do not survive stringent testing -- but if one does, we have at our disposal a precise, general theory which has stood up to the hardest hitting tests which scientists have thus far been able to dream up.

EXERCISES B:

1. Below is a conjecture and several proposals for testing it. Compare a with a', b with b', etc. Which test is more stringent, or are they equally stringent, or is it impossible to say? Defend your answer.

Conjecture: No one who takes Vitamin C regularly gets a cold.

(a) Study for one year 50 people in Jamaica who don't take Vitamin C. See whether any of them gets a cold.
(a') Study for two years 100 people in Jamaica who don't take Vitamin C. See whether any of them gets a cold.
(b) Interview 100 people in Minneapolis who have gone to the doctor for a cold. Ask whether they are regular Vitamin C takers.
(b') Interview 100 people in Minneapolis who have not had a cold all winter. Ask them whether they are regular Vitamin C takers.
(c) Interview 100 joggers who are Vitamin C nuts to see if they have had a cold this winter.
(c') Interview 100 patients in an old folks home who take Vitamin C regularly to see if they have had a cold this winter.

2. To test the theory "All metal rods expand at least 0.01% of their original length when heated 10°C" the following test was performed:

Ten samples of copper rods and ten samples of iron rods were measured with a ruler at room temperature, heated up 25°C, and then measured again. The % increase was then calculated.
(a) Give several different ways in which the test could be made more stringent and explain why each is more stringent.

(b) How could the falsifiable content of the hypothesis be increased?

3. In each of the following problems, assume the theory in question is testable (e.g., that we have an operational criterion for "losing one's temper"). Rank the proposed tests in order of stringency -- or if they are of roughly equal stringency, so indicate. Give a brief explanation of your answer.

(a) Theory: Jones' Irish grandmother never loses her temper.
   Test A: Tell her you hate the English.
   Test B: Tell her that scientists have shown that eating potatoes makes people dull-witted.

(b) Theory: Anyone can learn within two hours to vary their pulse rate using biofeedback techniques.
   Test A: Do the experiment with 200 IU students.
   Test B: Do the experiment with 200 Indian yogis.

(c) Theory: The IQ's of identical twins always fall within 10 points of each other.
   Test A: Test 20 pairs of twins selected randomly from the total population of twins.
   Test B: Test 20 pairs of twins selected from the population of twins orphaned at an early age and later adopted into different families.

(d) Theory: All farm kids are bare-footed.
   Test A: Pick a kid wearing shoes and ask him or her where he or she lives.
   Test B: Pick a city kid and see whether he or she is bare-footed.
   Test C: Pick some bare-footed person who lives on a farm and see if he or she is a kid or not.
   Test D: Pick a bare-footed kid and ask where he or she lives.
4. An epidemic of the screaming mimsies (call this condition SM) breaks out. Some scientists think that the SM condition is probably correlated with over-exposure to Archie Bunker (technically known as the AB factor). Others think the SM condition is connected with the consumption of a new snack food made largely from plastic, concentrated cholesterol, and mercury (with DDT added for flavor). The new product is called "Bummer Snacks" (or BS).

Preliminary studies are conducted in the town of Boobington. A sample consisting of the first 100 people with SM to enter the hospital are given depth interviews and it is found that 24% of them adore AB, 27% of them detest him, 39% couldn't care less, and the remainder think Archie Bunker plays for the Indiana Racers. All 100 SM patients had eaten large quantities of Bummer Snacks.

Newspaper headlines report: "Scientists prove anyone eating Bummer Snacks gets the screaming mimsies."

(a) Symbolize the generalization reported by the newspaper in a rough fashion.

(b) According to Popper's theory of methodology has that generalization been tested in a fairly stringent fashion?

(c) How might one give it a more stringent test?

5. Is there a demarcation between the sciences and what might be called the practical arts? (I am thinking of investigations into the way of cooking perfect soufflés, making clay pots which don't explode in the oven, welding strong, smooth joints, or judging when each step in the process of making champagne is completed.)

In answering this question first stress the similarities between the two domains. Then propose some plausible differences and subject each of them to criticism.

6. In the discussion of stringent tests, several dimensions of stringency were introduced -- extensiveness, variety, severity, etc.
(a) Which of these dimensions, if any, tend to be in conflict? For example, does increasing the extensiveness of a test work against making it varied? (Obviously not, but some of the other dimensions may conflict.)

(b) Which of these dimensions, if any, are (largely) redundant? For example, if we increase the extensiveness of a test sample do we automatically make it more varied?

e. What Does a Prediction Failure Refute?

As described so far, the logic of testing is simple and clear-cut:
(1) We derive a prediction from our conjecture which can be subjected to experimental check. (2) We do the experiment. (3) If the prediction is wrong, the theory is refuted. Period. Or so it would seem. In the typical scientific case, however, the situation is more complicated and the decision as to exactly which premise is to be given up is less straightforward.

Let me first illustrate the problem with a non-scientific example. Suppose you are bragging about your athletic dog: "I bet my dog would jump over that six-foot fence if there were something he liked on the other side." Your friend, tired of hearing about your Hairy Hoosier Hound, decides to call your bluff. She takes a piece of raw hide out of her brief case; HHHH sniffs at it; she then slowly tosses it over the fence. HHHH watches attentively and then goes off in the other direction to dig a hole. "Your conjecture is refuted," she says. "Your dog won't jump over that wall." "No," you say. "My conjecture may still be correct. In your test you made an additional assumption, namely, that HHHH likes to chew rawhide. As a matter of fact he's bored by it."

Your point about the logic of the testing situation is correct. We can analyze the situation as follows:

Conjecture: If there's something HHHH likes on the other side, he will jump over the fence.

Auxiliary Assumption: HHHH likes rawhide.
Experimental Prediction: (Therefore) If a piece of rawhide is thrown over the fence, HHH will jump.

Experimental Finding: A piece of rawhide was thrown over and HHH did not jump.

The experimental finding contradicts the experimental prediction. The experimental prediction was derived from two premises. Since the prediction is wrong, at least one of the premises must be wrong. You claim that the auxiliary assumption, not the conjecture, is false. Your friend obviously thinks otherwise.

In this case, it is probably easy to find out whether HHH likes rawhide chew bones or not. Does he whine piteously when one is placed just beyond his reach? Does he snarl if other dogs approach when he has a chew bone?

If HHH likes to chew rawhide, then your conjecture is indeed refuted (unless you can find some other auxiliary assumption to blame!). If HHH doesn't like rawhide, then your conjecture has not yet been properly tested and we can say nothing about its truth or falsity.

Let us now turn to a famous scientific example, the case of stellar parallax. After Copernicus put forward his theory that the earth revolved around the sun, astronomers noted that if his theory were true, one should be able to detect stellar parallax. If one is moving with respect to an object, then the direction in which the object appears changes. This phenomenon is known as parallax. As a race driver moves past the pit stop, at first it is ahead of him/her. Later it is behind. The angle $\alpha$ in the diagram below is called the angle of parallax.

![Diagram of car at time 1 and car at time 2 with pit stop and angle $\alpha$]
A similar diagram could be used to illustrate Copernicus' theory of the earth's annual movement with respect to a particular star.

But when 17th-century observers looked for stellar parallax, they couldn't detect any. Didn't this mean the theory was false? The supporters of Copernicus' theory decided to blame an auxiliary assumption instead. Their argument can be illustrated with the race-car analogy. Suppose the driver sights on a distant radio tower instead of on the pit stop. Now the angle of parallax becomes too small to be noticeable.

As the ratio of D to R increases, \( \alpha \) gets smaller. At very large values of D it will become too small to detect.

According to estimates of the distance between the earth and the stars available at the time, stellar parallax should have been observable. But the Copernicans argued that these estimates were wrong and claimed that the universe was about 1,000 times bigger than had previously been imagined. This bold move turned out to be correct, but 200 years passed before stellar parallax was detected experimentally.

The logic of the testing situation was as follows:

Copernican theory: The earth revolves around the sun, which is stationary relative to the stars.

Auxiliary hypothesis: The distance between the earth and the stars is about 20,000 earth radii.
Experimental Prediction: (Therefore) Stellar parallax should be easily observable with the apparatus available.

Experimental Finding: No stellar parallax is observable with the available apparatus.

Since the prediction failed, one of the premises had to be wrong. Copernicans blamed the auxiliary hypothesis; anti-Copernicans defended it and blamed the theory instead. With no good way at the time to test the auxiliary hypothesis, the status of the Copernican theory was left open.

The philosopher who first stressed that almost all tests involve a lot of auxiliary assumptions was Pierre Duhem, an early 20th-century philosopher, physicist, and historian of science. Hence, we will call the following the Duhemian problem:

When an experimental prediction turns out to be false, should the scientist blame the theory under test or the auxiliary assumptions (or both)?

There is no simple solution to the Duhemian problem, but a few guidelines can be laid down.

First of all, one should not use the Duhemian problem as a general excuse for one's pet theory. It is not good methodology to say, "My theory's prediction failed? Well, not to worry. I probably made a false auxiliary assumption somewhere along the line." If one wants to keep the theory despite the prediction failure, one must point to a specific auxiliary assumption and then design tests of that auxiliary assumption. If the auxiliary assumption passes the tests, then we should conclude that our theory and not the auxiliary was false.

Sometimes, however, it is not possible or practical to test auxiliary hypotheses. (We saw an example of this in the Copernican case.) In such instances, we can draw no firm conclusions about the original test situation. If a theory in conjunction with a variety of auxiliary assumptions makes a lot of false experimental predictions, though, we tend to decide that the theory is false, even though we can't conclusively test each auxiliary.
In the history of science, it is fairly rare to find a case where a theory is refuted by a single, decisive experiment. More often theories come to be rejected through a variety of prediction failures. Theories are rarely struck down by a blow from one type of crucial experiment, no matter how many times that experiment is repeated. Rather they are eroded away by an accumulation of anomalous results.

The Duhemian problem situation can be analyzed as follows:

The theory under test (T) when conjoined with one or more auxiliary hypotheses (A) makes a prediction (p).

Experiments show that p is not the case. By modus tollens we know that either T or A (or both) must be false, but logic doesn't tell us which.

\[(T \& A) \rightarrow p\]

\[^\neg p\]

(Therefore) \[^\neg T, \text{ or } ^\neg A, \text{ or } ^\neg T \& ^\neg A\]

Note that in the pure Duhemian problem situation there is no controversy about the experimental result, \[^\neg p\]. Furthermore, all parties agree that T & A imply p. The disagreement arises about whether to revise A or to revise T.

Of course, there are also cases in which people cannot agree on experimental results or on what exactly the implications of the theory are. These latter disagreements can usually be settled either through further experimentation or by means of logical analysis. The Duhemian problem is often more recalcitrant.

**EXERCISES C:**

1. The conjecture under test is Galileo's law of falling bodies: All freely falling bodies move towards the center of the earth with a constant acceleration. Suppose the following highly confirmed, reproducible experimental result is published: Steel ballbearings dropped off the top of Ballantine Hall always swerve towards the south as they fall.
Does this finding refute Galileo's law? Are there auxiliary hypotheses which could be blamed instead? (What are T, A, p and \( \neg p \) in this example?) Can the auxiliary hypotheses be tested?

2. You are trouble-shooting a malfunctioning stereo unit and record the following observations:

(i) It doesn't work when hooked up in the normal way using the turntable as the signal source.
(ii) It does work with the tapedeck as the signal source.
(iii) It does work with the FM tuner as the signal source.

(a) In which component would you most likely conclude the problem lies?
(b) Must you conclude this? Are any unchecked auxiliary assumptions involved in your reasoning?

f. Eliminating Alternative Explanations--Controlled Tests

Even if we are lucky enough to find a hypothesis which gives correct predictions, our process of testing and criticizing it is not complete. We are still faced with the problem of searching for and eliminating alternative explanations of the same phenomena.

Here are some rather simple (and in some cases silly) examples of what can be a complex problem:

(i) Your dizzy roommate says, "Wow, have I got an electrifying personality. Everytime I even look at the light bulb in the fridge, it lights up."

(ii) Your paranoid neighbor says, "You know why I don't like scientists? Because they keep all those barometers in their lab. Everytime their barometers go down we have a storm."
(iv) Your fraternity brother has a get-rich-quick scheme: "Let's market beer as a contraceptive. Look it—all the Zeta boys drink beer and not a single one has gotten pregnant."

The above examples are alike in two respects: First of all, in each instance the underlined generalization is true. (It's literally true that no fraternity boy who drinks beer gets pregnant.) Secondly, each generalization is seriously misleading. (Drinking beer has nothing to do with it!)

Most of the errors in the above examples involve mistakes about causality—it is not a low altitude, per se, which brings on malaria; it is the mosquitos who live there. The barometer's behavior is caused by the same atmospheric conditions which cause the storm. Etc.

Rather than trying to analyze the concept of causality, which is a knotty philosophical problem, let us simply say that in each of the above four examples there exist alternative explanations for the same phenomena, and furthermore, it is possible to design tests which will support the alternative and undermine the silly interpretations given above.

For example, you could show that the fridge lights up whenever the door opens, regardless of whether your dynamic roommate is around. Etc.

One good procedure for eliminating alternative explanations and showing that our hypothesis has isolated a crucial factor is the method of controlled tests. Suppose we wonder if factor F (e.g., the roommate) is in any way responsible for phenomenon P (the light going out).

What we could do is set up two situations: In the so-called experimental situation we would have F (the roommate) present. In the so-called control situation, we would make everything the same (i.e., the position of the fridge, the way the door is opened, etc.) except that the F factor would be absent. We can then readily see whether P (the lighting of the bulb) depends on F (the roommate's presence).
Controlled tests are also the appropriate sort to use when a conjecture is of the form, "The presence of Q causes an increase/decrease in the frequency/intensity of R." For example, to test the conjecture, "Regular doses of vitamin C decrease the frequency of colds," we would look at two groups: an experimental group whose members take vitamin C and a control group which is similar in every relevant respect except that its members do not take vitamin C.

Notice that the respects in which the members of the control group are required to match those of the experimental group depend on our background knowledge. Obviously the groups will differ in some ways -- even "identical" twins are not really identical! But which differences might be relevant? For example, should we try to match the groups with respect to eye color? Probably not, although it has been shown that people with blue eyes have more trouble wearing contact lenses so it is conceivable that they might have different susceptibilities to colds.

One thing we definitely should control for is the phenomenon known as the placebo effect. Because the incidence of colds may depend in part on psychological factors, it is important that both groups have similar beliefs about how likely they are to get a cold. For this reason every time the experimental group is given a vitamin C tablet, the control group should be given an innocuous pill (known as a placebo) which they believe is vitamin C.

It is often also necessary to use an experimental design known as a double-blind. In this case neither the subjects nor the experimenters know which group is getting the placebo until the experiment is over. (If the experimenters know but the subjects do not, we will speak of a single-blind experiment.) This device insures that the expectations of the experimenters do not affect the results. For example, we do not wish the experimenter to say "Feeling good today?" to the vitamin C group and "Got a little sniffle, huh?" to the control group.
EXERCISES D:

1. Suppose you are testing the efficacy of acupuncture and want to rule out the possibility that it works through psychological suggestion. How could you set up a single-blind test? (What could be used as a placebo?)

2. Consider the conjecture: "Talking to plants causes them to grow faster." Design an appropriate controlled experiment to test this conjecture.

3. List a variety of practical difficulties in testing the following conjecture: "Increased violence in TV shows and movies is one cause of the increasing frequency of violent crimes."

4. As I have described the single-blind test above, the control group is told a lie, namely that they are getting vitamin C. Could one re-design the experiment so that no deception is required?

5. Design controlled tests for the following two conjectures:

   (a) Beating a drum during an eclipse causes the sun to reappear.
   (b) Penicillin prevents syphilis patients from developing paresis (general paralysis and brain damage).

6. Suppose that before conducting the controlled tests in (5) above, you are rather inclined to believe that the conjectures are in fact true, but of course you don't really know because they haven't been tested. What moral problem would you face as an experimenter? How could it be resolved?

7. Suppose you are on an auto camping trip with a Martian who is quite sophisticated in scientific method but knows very little science, especially science which deals primarily with terrestrial phenomena. You are not always able to keep the ice chest filled and occasionally things spoil. After several such experiences the Martian remarks, "Whenever the bacon turns green, the cottage cheese turns green." How would you convince the visitor from Mars that the bacon's color change did not cause the cottage
8. Your dog barks at people walking down the street. Whenever your dog barks the neighbor's dog starts barking. The neighbor complains, "I don't like my dog to bark, and your dog is making her do it. Stop provoking my dog or I will sue." What evidence would you need to prepare your court case?

g. When Has A Theory Been Tested Enough?

We have discussed what happens when our theory's prediction is refuted --either we revise it or adjust an auxiliary hypothesis. What happens if we are fortunate and our theory passes every experimental test with flying colors? Can we then declare it proven true and move on?

The history of science suggests that we should never feel completely certain about any scientific generalization, no matter how frequently or stringently it has been tested. Newton's theory of classical mechanics had perhaps the best track record ever; yet it was superceded by Einstein's relativistic mechanics. Here are a few other examples of well-established claims which eventually had to be corrected or rejected:

(i) Matter cannot be created or destroyed. (Not true in nuclear fission or fusion processes.)

(ii) The sun rises once every twenty-four hours. (Not true at the North Pole.)

(iii) All molecules of water are made of the same stuff. (Not true for heavy water, deuterium oxide.)

(iv) The major difference between homo sapiens and the lower animals is that man can use language. (Not true for chimpanzees which can use sign language.)

(v) Living matter can only come from living matter; it cannot be formed from inanimate substances. (Not true--amino acids can be synthesized from ammonia, methane, hydrogen, etc.)

So the history of science warns us that any scientific claim is fallible. Logic and philosophy of science can help us understand why this is so. Here are some of the reasons:
(i) Generalizations cover a potential infinity of cases. But we can only check on a finite number of predictions. We can never be sure that the next case won't violate the rule (e.g., a black swan may turn up in Australia).

(ii) Scientific theories make infinitely precise claims. But we can only make measurements of finite accuracy. (For example, Newton's law of gravitation says the force of gravity varies inversely with the square of the distance, i.e., the exponent is $r^{-2.00000...}$ but our measurements cannot discriminate between $r^{-2}$ and $r^{-2.0000000001}$.)

(iii) Many of our scientific laws only hold under idealized conditions—to give two very simple examples, the law of the lever assumes no friction at the fulcrum, and the law of the pendulum assumes there is no air resistance. Of course we can try to minimize such interferences when we conduct tests, e.g. by resting our lever on a point or setting up the pendulum in a vacuum, but our experiments never achieve the perfect conditions which are assumed in our ideal laws.

(iv) There may be alternative theories which we have not even dreamt of yet which account for all of the data we have in hand.

For all these reasons, theories are underdetermined by our observational results and can never be proved through any amount of observation and experiments. There are no rules for deciding when to accept a theory (for the time being) and move on to new problems, but what we can do is to answer each of the above sources of fallibility as best we can.

(i) By testing in widely scattered domains, we guard ourselves against parochialism, e.g., the black swans in Australia.

(ii) By making our tests as precise and ideal as possible, we can approach the infinite precision and perfection of our theories.

(iii) And the best way to rule out alternative explanations is to deliberately try to imagine radically different ways of explaining our results. If we can devise a new alternative, we can then set up a crucial experiment between the two competing accounts.

If we have tested our theories stringently and eliminated all the alternative hypotheses we can think of, then we have good reasons to accept our conjecture—for the time being until new data or alternatives come along.