# Chapter 3

# Hypotheses and testing of hypotheses

> "We and other animals notice what goes on around us. This helps us by suggesting what we might expect and even how to prevent it, and thus fosters survival. However, the expedient works only imperfectly. There are surprises, and they are unsettling. How can we tell when we are right? We are faced with the problem of error."
>
> W.V.O. Quine

## *3.1 Introduction*

What do we mean when we say that an argument or an activity is scientific? Can we find any general criterion which natural sciences, social sciences and the humanities alike meet, or ought to meet, in order to be called scientific? Some have said no, the natural sciences and the humanities are fundamentally different; these domains of discourse have completely different goals and methods and no common trait can be found. The German philosopher and historian Willhelm Dilthey (1833-1911) and his compatriot Max Weber (1864-1920), one of the founding fathers of sociology, both claimed that the goal of natural sciences is explanation, whereas the goal of the 'human sciences' (German 'Kulturwissenschaften') is understanding. In this context, 'understanding' means understanding people's thoughts, feelings and actions; to understand a causal mechanism in nature is something else entirely. Weber, Dilthey and many others claimed that understanding the inner life of other people requires empathy (German 'Einfühlung'), and therefore all sciences concerning human culture should use methods that take empathy into account. In this respect, they are fundamentally different from the natural sciences.

Some have criticized this view, arguing that understanding cannot give us reliable knowledge and that empathy is beside the point. For example, sociologists in the tradition stemming from Durkheim claim that there exist social facts that are just as accessible as facts of nature and which can and should be studied using quantitative methods.

Supposing that Dilthey and Weber are right, are there no common traits among all those activities we usually call scientific? Why, then, do we call them *scientific*, call their work *research* and hire people to do this special kind of work at *universities*? Some might say that it is just a matter of tradition that certain activities are performed at universities. Furthermore, they might argue, people in other fields can increase their own prestige by calling their activities science and research, basking in the reflected glory of the impressive accomplishments of modern natural science. There is some truth in both claims. However I do think that all sciences (in the broad sense of the word, including social and human sciences) have something in common.

But why bother about the concept of science? The answer is twofold. The first reason is that we need a general criterion for distinguishing between science, on the one hand, and pseudo-science, superstition, fraud, deception etc., on the other. By calling a statement scientific, we usually mean to say that it has good credibility and could be relied upon. Of course, the correct use of scientific methods is not in itself a guarantee for truth, but compared with other ways of acquiring knowledge, the scientific way is more reliable. Now, if each discipline had its own methods and rules, we would have no reason for dismissing e.g., astrology as a pseudo-science. The astrologer can claim that astrology has its own methods, and these are just as scientific as others. If there are no general criteria for what constitutes a science, each discipline can formulate its own criteria. Thus there is no reason why astrologers should not enjoy the same public support as other fields of inquiry. This argument has in fact been put forward by Paul Feyerabend. I think that Feyerabend's conclusion is unacceptable, and therefore a general criterion for science is needed.

Another reason for why we should seek a general criterion for science is that it can help us to understand one of the great changes in human history, namely, the scientific revolution. It is fair to say that the scientific revolution started a transformation of human life and society that has continued up until the present. If we can find a good characterisation of science, we are in a better position to understand scientific developments and thus to understand better a crucial component of long-term changes and developments in society.

### *3.2 The hypothetic-deductive method*

I believe that it is possible to find a general criterion for all sciences. There is something they have in common despite great differences; viz. that hypotheses are proposed and tested. In other words, my belief is that the hypothetic-deductive method is used in all sciences. Some

researchers in the humanities would immediately protest, saying that we in our discipline are not *testing hypotheses*, we are *interpreting* things such as texts and other artefacts, human actions, historical events etc. My answer is: of course, you interpret, but that is a kind of testing of hypotheses. I will return to this discussion later in this chapter, but first it is appropriate to be more specific about the concept of testing of hypotheses.

The best way of introducing the hypothetic-deductive method is to describe and analyse some examples from the history of science. I will give three: the rejection of spontaneous generation by Francesco Redi, an episode from Claude Bernard's seminal research in physiology, and Emile Durkheim's investigation of suicide rates in Switzerland.

*a. Refutation of spontaneous generation*

Spontaneous generation was the once common belief that worms, larvae and all kinds of animals found in rotting matter were generated without any progenitor, that new life emerged spontaneously in rotting matter. Given surface observations, the idea appears reasonable, not to say obvious. But Francesco Redi (1621-97), a physician in Florence, came up with the idea that larvae found on rotting meat came from flies and not from the rotting meat. In order to test this idea, he performed an experiment. He filled four jars with fish, veal and pieces of a snake, and covered the jars with waxed paper. Then he filled another four jars with fish, veal and pieces of snake, but left them open. After a few days, all the fish and meat was rotten, and in the open jars he found plenty of larvae. In the covered jars, however, he found no larvae. He drew the preliminary conclusion that larvae could not be generated out of rotting fish or meat.

Nonetheless, some doubts remained. Critics might argue that the spontaneous process required fresh air and, obviously, the air in the covered jar was not at all fresh. So he repeated the experiment, this time covering the jars with gauze; it would allow the air to circulate while at the same time prevent the flies from laying their eggs on the meat and fish. The result was the same, not one single larvae were found in the gauze-covered jars.

Redi published his findings 1668 and succeeded in convincing his readers that spontaneous generation was incorrect as a general mechanism. But it was not completely rejected. In the then newly invented microscopes (by Loewenhoek around 1650) people could see small animals, not visible with the naked eye, and these animals could not be shown to require ancestors. Thus belief in spontaneous generation survived with regard to microscopic organisms until Louis Pasteur's definitive refutation in the middle of the 19$^{th}$ century.

This little episode could be reconstructed as follows. Redi was out to refute the following hypothesis:

H: Larvae are generated spontaneously in rotting meat in the absence of any progenitor.

From this general hypothesis he drew the conclusion:

E: If meat or fish is put in a covered jar, larvae will be found in the jar when the meat is rotten.

The conclusion is called an empirical consequence (E) because it is directly empirically testable. And, as the test showed that E is false, the hypothesis must be false. The reasoning can be schematised in the following manner:

If H, then E
E false
———————
H false

The bar indicates the step from the premises to the conclusion. The form of this logically valid argument is well known and is called *Modus Tollens*.

But how should we bring in the further discussion about fresh air into this scheme? The argument about the fresh air would, if successful, show that the conclusion that H is false could be avoided. It is suitable here to introduce the concept of an *auxiliary assumption* (A):

A: Fresh air is required if the spontaneous generating process shall produce larvae in rotting meat or fish.

The more complete analysis of the argument is thus:

If H and A, then E
E false
———————
H or A false

It is obvious that one cannot logically infer that H is the false assumption if E is false; it could just as well be A. As there was no fresh air in the covered jars, we can explain the lack of larvae without concluding that H was false. Therefore, a second test, in which the auxiliary assumption is fulfilled, is needed in order to be able to conclude that H is false.

In sum, the first scheme displays the logical structure of the reasoning connected with the second experiment, whereas the second scheme displays the logic of the first experiment.

*b. Claude Bernard and the emergence of experimental physiology*

Claude Bernard (1813-78) worked in Paris and was the first to use animals in physiological experiments. He found that there is a connection between the quality of urine and what kind of food is eaten, as follows:

*T*. Carnivores have clear and acidic urine and herbivores have muddy and alkaline urine.

I label this result *T*, as an abbreviation of 'theory'. However, one day Bernard found that some rabbits that had recently been acquired from a breeder had clear and acid urine. As rabbits are herbivores, this finding contradicts *T*. What is the cause of this anomaly? Bernard came up with the idea that the rabbits might have been starving while they were being transported to the laboratory, and that this could explain the abnormal urine: the body uses its own flesh for nourishment when starved. This hypothesis can now easily be tested simply by giving the starving rabbits normal quantities of food and then observing the quality of their urine. The same test could also be used on horses, which were readily available in Paris 150 years ago. Another way of testing the hypothesis was to dissect starving animals and to look for signs of digestion of the muscles. The results were in all cases in accordance with the expectations. The structure in the reasoning can be displayed as follows:

H: Herbivores get acidic and clear urine when starving.
E1: Rabbits that have been starved will get muddy and alkaline urine when given food again.
E2: Horses get acidic and clear urine when starving.
E3: When dissection animals have been starved for some time, one can see signs of their digestion of the muscles.

If H, then E1, E2 and E3
E1, E2 and E3 are all true
— — — — — — — — — — —
H is confirmed

Notice that in this case the line dividing the premises from the conclusion is dotted, indicating that the conclusion is not a logically valid one. The hypothesis is not proved, or known to be true, but only strengthened: confirmation is not the same as valid proof. It is obvious that the general statement H might be false, although the particular empirical consequences are all true. In fact, no amount of experimentation will prove the correctness of the hypothesis,

because it is a general statement concerning an infinite number of cases. Bernard, however, was satisfied with these three experiments, and concluded that his general hypothesis is true.

When comparing this case with the former one, we discern an asymmetry: it appears that it is, at least sometimes, possible to prove that a hypothesis is false, whereas it is impossible to prove the truth of any general hypothesis. At most we can provide the hypothesis with supporting evidence. This has consequences for our understanding of science, namely, even if we have overwhelming evidence for a general hypothesis and all believe in its truth, we cannot really claim that it is *proved*. Therefore we cannot, as a matter of principle, dismiss the possibility that further research might show that our current belief in the hypothesis is wrong. We must always be prepared to accept that we have made a mistake and that our firmest beliefs are wrong. This position is called *fallibilism*.

### *c. Durkheim's analysis of suicide rates*

Emile Durkheim, one of the founding fathers of sociology, was convinced that there are sociological facts, and that these could be studied with exact quantitative methods. In his famous book, *De la Suicide* (1900), he suggested, among other things, that the incidence of suicide in a certain society has to do with the structure of the society in question, not with the particular individuals living in it. One hypothesis he tested was that the incidence of suicide is inversely proportional to the degree of interaction in the society. The concept *degree of interaction* is rather abstract and requires a standard of measurement. Durkheim thought that Catholicism and Protestantism differ as regards degree of integration. As is well known, the number of religious services, masses, processions, etc., is much higher in Catholicism than in Protestantism. As a result, people in a Catholic society come together and participate in many more communal activities that in an otherwise comparable Protestant society. The result ought to be, according to Durkheim, a lower incidence of suicide in the Catholic society. Was he right that this difference would result in different suicide rates? In order to test this hypothesis, he needed to compare a number of societies, which were reasonably similar in many respects other than religion. He found such comparable societies in Switzerland, where there are a number of cantons, many of which shared a similar structure, some of which were Protestant, some Catholic and some mixed. The figures he collected were as follows:

| Society | suicide incidence per 100 000 inhabitants |
|---|---|
| Catholic cantons | 86,7 |
| Mixed cantons | 212 |
| Protestant cantons | 326 |

One must concede that there is a very considerable difference between Catholic and Protestant cantons and that this difference supports his hypothesis. An analysis of his reasoning is as follows:

- H: The incidence of suicide in a society with a low degree of interaction is higher than in a society with a higher degree of integration.
- A: Catholic societies are more integrated that Protestant societies, given that they are comparable in other respects.
- E: Catholic societies have a lower incidence of suicide than Protestant societies.

As in the former case the logical structure is

If H and A, then E
E true
---
H has been confirmed

We can summarise the logic of hypothesis-testing as follows:

- Formulate a hypothesis.
- Deduce empirically testable consequences from the hypothesis and auxiliary assumptions.
- Determine the truth-value of the empirical consequences by comparing with experiments and observations.
- Depending on the truth of the empirical consequences, infer either that the hypothesis is false or that is has been confirmed.

There is one obvious complication (there are more), viz. that the auxiliary assumption might be false, which might result in a false empirical consequence. Therefore one cannot definitely say that the hypothesis is falsified if the empirical consequence is false. Further investigations and checking of auxiliary assumptions are usually required.

It is time to provide definitions for the central concepts:

- **Hypothesis** $=_{def.}$ A sentence about which i) the author is not sure whether or not it is true and ii) is a premise in the deduction of empirical consequences.
- **Auxiliary assumption** $=_{def.}$ A sentence about which i) the author takes its truth for granted and ii) is a premise in the deduction of empirical consequences.

- **Empirical consequence** $=_{def.}$ A sentence which i) follows from the hypothesis and auxiliary assumptions (if there be any) and ii) whose truth-value can determined by observation and experiment.

When analysing the logic of scientific reasoning, one should be careful to distinguish between *empirical consequence* and *observation sentence*. An empirical consequence is a sentence that follows from the hypothesis and auxiliary assumptions, whereas the observation sentence is just that, a report of what has been observed: if they contradict each other, the hypothesis or one of the auxiliary assumptions is false.

### *3.3. Interpretation as testing of hypotheses*

It is not controversial to claim that the testing of hypotheses is a common trait in the natural sciences. Furthermore, it is universally accepted that at least part of the social sciences, those using quantitative methods, also use the method of testing hypotheses. However, it is much more controversial to hold that even the humanities formulate and test hypotheses; that is however my view. I shall now, by way of an example, argue that the humanities can be said to test hypotheses. My example is taken from history: the disappearance of Raoul Wallenberg.

Raoul Wallenberg (RW for short) was a Swedish diplomat at the embassy in Budapest during the end of World War II. He worked hard to save Jews from the Holocaust by giving them Swedish interim passports and helping them to leave Hungary. When the Soviet Red Army came to Budapest in January 1945, he disappeared. A few weeks later, the Red Army in Budapest said that they had taken RW into custody. He has not been heard from since. The Swedish government asked the Soviet government about his whereabouts on several occasions. At first, the Russians denied having any knowledge of his fate, but, in 1956, the Soviet government admitted that he had been imprisoned and brought to Lubljanka, the headquarters for NKVD, Soviet security and intelligence, in 1945. The Soviet government claimed to have found a hand-written memo in the archives, which stated that a person named Walenberg had died there on the night of July 17, 1947. The Soviet government also explained why they earlier had denied all knowledge of RW; they blamed the former Security Minister, Abakumov, who had been found guilty of severe criminal activities (the persecution, torture and assassination of high-ranking Communists, among other things) and sentenced to death in 1954. Abakumov did not inform his colleagues in the government about RW, the Swedish government was told, and he removed all documents concerning the case, except for one that he could not find, the memo written by the physician on duty at the Ljublanka prison in July 1947. In this memo the physician writes that the prisoner Walenberg had died and asks

about instructions. An additional note says that order had been given about cremation of the body without autopsy. This is the main content of the memorandum delivered to the Swedish government in 1956.

The Swedish government accepted the Soviet account, and for a number of years the case seemed closed. However, around 1960, rumours about RW began to circulate; westerners who had been in Soviet prisons and released during the thaw under Khrushchev spread reports saying that RW still were alive in Soviet prisons, especially in the Wladimir prison. The case was reopened, and officials from the Swedish government interviewed several such former prisoners. Since then, the Swedish government takes the position that the fate of RW is unknown, and has made repeated inquiries to the Russian government concerning his case.

What is the truth? Two Swedish historians, Hans and Elsa Villius, have examined all of the relevant documents and testimonies, using standard critical methods when deciding the reliability of a historical source. They considered three possible hypotheses:

1. RW actually died in Ljublanka at July 17th, 1947, as was stated in the memo from the Soviet government from 1956.
2. The Russians have not deliberately lied to the Swedish government, but rather mistaken RW for another person; the person who died in July 1947 in Ljublanka was not RW, but someone with a nearly similar name (spelled Walenberg).
3. The Russians lied, and RW was alive after July 17, 1947.

Each of these hypotheses has consequences, which can be checked. For example, if 1 is true, the rumours about RW living in the fifties must be false. If so, those spreading these rumours must be lying. Another consequence is that there must be some reason why the Russians in the forties claimed that RW was not in Soviet Union, and that only one single document concerning his fate has been found.

If the second hypothesis is true, there must have been two different people with nearly identical names, and the Russian authorities have confused these two with each other.

If the third hypothesis is true, the Soviet government deliberately lied to the Swedish government, and must have had some reason for doing so.

Hans and Elsa Villius have investigated the plausibility of a number of consequences of these three hypotheses and published their results in *Fallet Raoul Wallenberg*, 1966. For example, if 1 is true, what is the source of the rumours about RW living after 1947? The Villius investigation showed that the rumours could be traced back to 4 people, among whom one claims to have met Wallenberg. This individual, however, has given vastly different versions of his meeting with Wallenberg when interviewed in a newspaper and when

interviewed by Swedish officials. His credibility is therefore not good, and there is reason to suspect that his information actually came from someone else, a person with the name Simon Goguberidze. This individual in turn claims that he had heard that Wallenberg was in the Wladimir prison in 1956. Interviews with the other three witnesses indicate that their information about Wallenberg presumably also comes from Goguberidze. In total, there is actually only one third-hand report of Wallenberg being alive in Wladimir prison in 1956. According to normal historical method, only first-hand reports can be used as reliable information. The conclusion must be that there is no positive evidence for claims that RW was alive in the fifties.

The second hypothesis, that the person who died in Ljublanka was another person with an almost similar name, is less plausible by the fact that double consonants often is pronounced as single ones in Russian language. This is actually evidence for the authenticity of the handwritten document.

The first hypothesis together with known facts about Abakumov and his actions, however, fits with all known facts, such us the disappearance of all documents except one about Wallenberg, the official denial of any knowledge about Wallenberg just after the war, etc.

Hans and Elsa Villius concluded that the only plausible hypothesis is the first one; RW really died 1947, and the rumours about him still living in a Soviet prison are false.

This is not the proper place to rehearse all the detailed arguments behind this conclusion. Let me just point out that all of these arguments are built upon the interpretation of texts, statements, interviews, actions, motives etc., of quite a number of actors. In short, this is a typical case of the interpretation of human actions, and fits well with the hypothetic-deductive model.

It may be worth noting that much later, in January 2001, a Swedish-Russian commission reported the results of their intensified investigations. Their conclusion was the same as that of the Villiuses, namely, that RW died the night of July 17, 1947. (However, the Swedish government officially still had some reservations; asked about that in a TV-program two other historians, independently of each other said that there were no historical reasons for doubting the conclusion; hence the Swedish government must have had political reasons not to fully endorse the conclusion).

New information is now available, e.g. that RW plausibly was poisoned in a top-secret operation. Understandably, Stalin and Abakumov did not want to have a Swedish diplomat in Soviet prisons, and he could not be released after having been imprisoned for two years, for

that would show that Soviet Union violated international diplomatic code. Thus it is understandable that Abakumov, and perhaps Stalin, ordered a covert assassination of Wallenberg without any trial and tried to remove all documents relating to him.

It is of course not completely certain that Wallenberg died in July 1947; the conclusion depends on a number of oral and written reports and interpretations of people's statements, motives and actions. However, there seems to be no further reasonable doubt, especially taking into account the report from the Swedish-Russian commission of 2001. The situation is quite similar to that in the natural sciences; a hypothesis can be tested and it can be confirmed, but it can in the strict sense never be conclusively proven.

### *3.4. Statistical testing of hypotheses*

To say that a hypothesis has been supported or that it is more probable than before testing is not very precise. We would very much like to be more exact, i.e., to provide some standard for the probability of a hypothesis. Could we use statistical methods for calculating the probability of a hypothesis after testing? The answer depends on the interpretation of probability.

The problem can be put in the following way: we want to know the conditional probability that a certain hypothesis H is true, given a certain outcome of an experiment. Assuming that we have only one alternative hypothesis, $H_0$, (called the null hypothesis, 'the pessimistic alternative') the conditional probability can be calculated using Bayes' theorem:

$$P(H/O) = \frac{P(H)P(O/H)}{P(H)P(O/H) + P(H_0)P(O/H_0)}$$

where P(H) is the initial probability for the hypothesis H, P(H/O) is the probability for H conditional upon the outcome O, P(O/H) is the probability for the outcome O conditional upon the hypothesis H, and $P(O/H_0)$ is the probability for the outcome O conditional on the alternative hypothesis $H_0$.

An example could be the testing of a certain drug as treatment for a particular illness. The hypothesis could be formulated as 'The drug has some effect on the illness,' and the null hypothesis as 'The drug has no effect on the illness'. (In real life the statistical methods are more sophisticated than that, but I simplify for pedagogical purposes.) To test the hypothesis, a random sample of people with the illness is selected. The sample is ramdomly divided into two groups, the test group and the control group. The test group is given the drug and the

control group is given a placebo. This method is called double blind testing, used in order to minimise the psychological effects of treatment. Now suppose there is a difference between the test group and the control group; what is the probability that this difference is caused by the drug, and not merely a chance event?

In order to calculate P(H/O), i.e., the conditional probability for the hypothesis H, given the outcome O, we need to know the probabilities P(H), P(O/H) and $P(O/H_0)$. The conditional probabilities are simply experiment results, but what about the initial probability P(H)? What is its value?

If we interpret the concept of probability as a measure of subjective certainty (or uncertainty), we can reason as follows. We have a hypothesis H that is either true or false. Initially, i.e., before testing, it could just as well be true as false and therefore its initial probability is 50%. By the same token, the initial probability for the alternative hypothesis is 50%. Then a test is performed and, using Bayes' theorem, new probabilities can be calculated. Hence the testing will increase or decrease the probability of the hypothesis as the case may be.

If, however, we do not interpret the concept of probability as a measure of the degree of subjective certainty, but rather as a measure of an objective property of the state of affairs (or event, or situation, or object), it is hard, not to say impossible, to give any reason why the initial probability should be 50%. The real but unknown probability could have any value.

Prima facie, it seems obvious that if probabilities are real properties and not just subjective states of minds, and if the real probability deviates substantially from 50%, the calculated conditional probability would be completely wrong, and we ought to notice this sooner or later. However, if a sequence of tests is performed and the initial probability in each is taken as the outcome of the former, (i.e., starting with a guess only in the first test) the sequence of probabilities will converge to a certain value, independently of the chosen initial probability. Adherents of this method (which is called Bayesianism) can thus defend the subjective approach by pointing out that, in the long run, the result is independent of the initial subjective probability.

This is, however, not the generally accepted view. Critics maintain that we are not justified in using any subjective assumptions in a truly scientific methodology and therefore we should not use any subjective measures. A more pragmatic argument is simply to say that we never have time nor money to perform a sufficiently long sequence of tests in order to see any convergence. Instead, it is argued, one should use the Neyman-Pearson method for making statistical inferences, which goes as follows. Suppose, as before, that we are interested

in the effect of a certain drug, and the two hypotheses are also the same: either the drug has some effect, which is the test hypothesis H, or it has no effect whatsoever, the null hypothesis $H_0$. Start by making a methodological decision, viz., to reject the null-hypothesis if the probability for the outcome conditional on the null-hypothesis ($P(O/H_0)$ is less than a certain predetermined value, the significance level. Usually one chooses 5%, 1% or 0.1 %. Suppose we have decided to reject the null hypothesis if $P(O/H_0)$ is less than 1%. Then we perform the significance test, that is, we calculate the probability $P(O/H_0)$ to see whether it is below or above the chosen significance level. Suppose the calculations confirm that the probability is less than 1%. We then follow our methodological decision, reject the null-hypothesis and accept the test-hypothesis.

It certainly appears rational to accept the test-hypothesis if the level of significance is 1%. However, it does *not* follow that the probability for the test hypothesis is higher than 99 %. Inspection of Bayes' theorem shows that such a conclusion can be drawn only if we know that P(O/H)=P(H/O). But that we usually do not know, and, consequently, no information about the probability for the hypothesis is available. There are certainly general qualitative arguments for accepting H if it passes a significance test, but the probability of H has not been measured.

Summarising, we can say that the Bayesian approach makes it possible to calculate the probability of a hypothesis, but this probability is not an objective property unless a sequence of repeated tests are performed. If, on the other hand, we use the Neyman-Pearson approach, we cannot calculate the probability of the hypothesis, but only make a practical decision. The vast majority of scientists use Neyman-Pearson statistics. To the epistemologist, the difference seems minimal.

### *3.5. Ad hoc-hypotheses*

As we saw in the spontaneous-generation example, one can save a hypothesis from refutation by inserting an auxiliary assumption at which to direct Modus Tollens. It is an obvious logical point that if several assumptions are needed in order to infer a consequence and the consequence is false, any of the assumptions could be false. Sometimes such auxiliary assumptions are obvious. A trivial example is the assumption that the measuring device is in order when testing a theory by measuring a certain quantity with this device. Sometimes, however, one can be quite suspicious about the motive for a certain auxiliary assumption, especially when it is formulated after the testing. If it is inserted merely for the purpose of

saving a hypothesis, it is called an *ad hoc hypothesi*s (lat. *ad hoc*= for this), and such hypotheses are forbidden within science. A little example might illustrate the situation.

The ancient world picture is, I think, well-known: the earth was believed to be at the centre of the universe, and the moon, the sun, the planets and the starry sky all rotate in a regular and perfectly circular motion around the earth. Using this picture, i.e., the ancient theory, one can explain a number of celestial phenomena, such as eclipses. According to this world-view, the moon and the sun move in concentric circles (being bound to different celestial spheres), and so it might happen that the moon comes exactly between the sun and the earth. However, there are different kinds of solar eclipses, depending upon which time of the year they occur. Sometimes an eclipse is total, the moon completely covering the surface of the sun, whereas sometimes the moon appears slightly smaller, thus not covering the sun entirely, but leaving a fine ring of the sun's surface visible around the moon. Thus, we observe either complete or ring-formed solar eclipses. This is not in accordance with the ancient theory, according to which both the moon and the sun move in circles around the common centre where the earth is situated. One might think it easy to accommodate this finding by suggesting that the sun moves in an elliptic trajectory, but that was not in accord with Greek thought, which held that bodies move in perfect trajectories, and that the circle is more perfect than the ellipse. Instead it was suggested that the moon sometimes constricts, viz. precisely in those cases when we observe a ring-formed eclipse. Moreover this constriction cannot be observed by means other than observing the eclipse.

From our present day vantage-point of astronomical knowledge, it is easy to laugh at the idea that the moon might shrink. But in discussions about methodology, we cannot use later and better theories as criteria for evaluating hypotheses, for we need criteria applicable at the moment. Popper defined an *ad hoc* hypothesis as an assumption that is not independently testable, i.e., not testable in circumstances other than those in which it was proposed in order do the explanatory job. This appears to catch exactly the decisive character of the hypothesis of the moon's constriction, since the constriction was said to occur *only* during a ring-formed eclipse and was not measurable by other means.

However, the criterion is not as clear-cut as one might wish. For today it is easy to measure the diameter of the moon (although it does require some equipment), and the moon-constriction hypothesis is now independently testable and readily refuted. It can thus be claimed that it was not an ad hoc hypothesis at all, but rather simply a falsifiable and indeed false hypothesis.

One is now tempted to modify the criterion by saying that a hypothesis is an ad hoc hypothesis if it is *in principle* impossible to test by independent means. But how do we decide that? The expression 'in principle' in this context cannot mean a purely logical qualification; rather, it means that, as a matter of physical principle it cannot be tested. But the physical principles are part of the problem, that is, the principles to which we adhere come and go with our theories.

A modern example might illustrate the point. The first experimental investigations of processes now called weak interactions apparently showed that the principle of conservation of energy was violated; the energy of the outgoing particles was lesser than that of the incoming particles. Pauli, however, suggested that energy conservation is a fundamental principle which should not be abandoned; hence the explanation for the loss of energy in the experiment must be due to some undetected new particle, later to be called neutrino. Calculations showed that if such a particle exists, it has no mass, no charge, and would therefore not interact with any at this time available instrument. It seemed reasonable to say that the neutrino was impossible to measure in principle and so Pauli's proposal could be accused of being *ad hoc*. However, further theoretical investigations of this type of interaction suggested ways of testing the existence of the new particle. 20 years later new experiments could be performed, and the existence of neutrinos was confirmed.

Should we now say that Pauli's idéa was not an ad hoc hypothesis after all, despite its initial appearance? It is hard to say. The difficulty resides in formulating a satisfactory definition of what constitutes ad hocness. In spite of this, scientists usually have a strong sense of what to count as ad hoc assumptions that should be rejected as illegitimate moves.

### *Exercises*

Here follows a couple of examples from the history of science. Analyse the episodes using the concepts of hypothesis, auxiliary assumption and empirical consequence.

1. For a long time, it was commonly held that the dinosaurs were cold-blooded. This assumption has been criticised, however, by Stephen Jay Gould, who claims that they must have been warm-blooded. His arguments are as follows:

a. The temperature of cold-blooded animals varies with the external temperature, and animals living in areas with great differences between winter and summer develop annual rings in the peripheral parts of their skeleton, just like trees. Warm-blooded animals do not

have such rings because they have a constant temperature. Dinosaurs from areas with great seasonal temperature differences do not have rings in their skeleton.

b. Large cold-blooded animals do not live near polar areas because they cannot be warmed sufficiently during the short winter days (summer days in the southern hemisphere), and they are too big to hibernate. However, there were large dinosaurs living far up north, who must have survived without any sunlight during extended winter periods and thus without any source of heat during the winter.

c. Modern reconstructions of the anatomy of dinosaurs show that many of the large dinosaurs are similar to walking mammals with regard to anatomy and the proportions of the extremities. And, as we all know, mammals are warm-blooded.

2. The Austrian doctor Ignaz Semmelweiss worked 1844-48 as obstetrician in Vienna. The hospital where he worked had two delivery units, of which both served approximately 3000 patients per year. It appeared that the mortality rates connected with puerperal fever differed vastly:

| year | Unit 1 | Unit dept 2 |
|---|---|---|
| 1844 | 8.2% | 2.3% |
| 1845 | 6.8% | 2,0% |
| 1846 | 11.4% | 2.7% |

A number of different explanations were proposed, such as bad diet, the wrong posture during childbirth, 'atmospheric disturbances', etc. All of these explanations were soon rejected. However, one day a colleague to Semmelweiss cut his finger during a post-mortem examination, developed a fever rather similar to childbirth fever, and died within a few days. This gave Semmelweiss the idea that childbirth fever had to do with the fact that medical students were trained in Unit 1, whereas midwives were trained in Unit 2. Since medical students came to Unit 1 after performing autopsies, Semmelweiss then surmised that a substance from the dead bodies was the cause of the fever. As a test, he ordered that all personnel should wash their hands in a solution containing calcium chloride after performing an autopsy. The result was as follows:

| | Unit 1 | Unit 2 |
|---|---|---|
| 1848 | 1,27% | 1, 37% |

(Despite this clear success, Semmelweiss failed to convince his colleagues of his findings, and was actually fired from his job at the hospital: he was accused of being uncooperative. It took more than twenty years before more hygienic routines were introduced. Semmelweiss'

idea was accepted only after it could be explained in the aftermath of Pasteur's observation in the microscope of the causal connection between bacteria and disease, a discovery that resulted in a complete transformation of pathology.)