Chapter 3
Hypotheses and Hypothesis Testing

We and other animals notice what goes on around us. This helps us by suggesting what we might expect and event 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 does it mean that an activity, or argument, is scientific? Is there any single criterion that the natural sciences, social sciences and humanities must fulfil in order to be called science? Many have answered 'no', on the basis that the natural sciences and the human sciences are fundamentally two different enterprises. One central argument for this view, offered by Dilthey[1] and Weber,[2] is that the natural sciences and the cultural sciences (which was their term for human and most social sciences) have different purposes: the purpose of the natural sciences is the explanation of natural phenomena, whereas the purpose of the cultural sciences is the understanding of human phenomena. Weber may be interpreted as holding that explanation is causal explanation and understanding is discerning the meaning of actions of individuals and collectives.

There are further differences between natural and cultural science, which we shall discuss in detail in Chap. 5; but despite the differences there are also similarities. In my opinion, there is a common denominator, a defining trait, of all that is properly called science, and it is our task here to identify it.

Is there a point to finding a good definition of science? Yes, and for at least two important reasons. Firstly, there is a need to distinguish between science and pseudo-science: the latter being activities that claim to be scientific but upon closer analysis prove to be based on wishful thinking, fraud, business interests or pure

[1] Willhelm Dilthey (1833–1911), German philosopher and historian who helped develop German idealism. He saw cultural phenomena as objectifications of people's mental lives.

[2] Max Weber (1864–1920), German sociologist, historian and philosopher. According to Weber, the purpose of sociology is to understand the meaning and/or values expressed in human action.

L.-G. Johansson, *Philosophy of Science for Scientists*, Springer Undergraduate Texts in Philosophy, DOI 10.1007/978-3-319-26551-3_3

superstition. In particular, it is essential for the medical field to be able to clearly distinguish between medical science and quackery. In public debate, success stories are often reported regarding things like homeopathy, herbal remedies, healing, etc. Even if one may sometimes believe in these success stories, the central question is whether that success was the result of the remedying method or some extraneous circumstances. Thus it can be of great value to be able to distinguish scientific methods from non-scientific, as only with scientific methods is there a reasonable chance that the conclusions one draws are true and can be generalized.

The other reason for discussing criteria for science is to make possible an understanding of the immense transformations that we call the scientific and industrial revolutions. These revolutions began a line of research that has since then accelerated to an amazing pace. If we can find a satisfying characterization of what science is, then we can better understand how the propagation of science works, as well as further our understanding of a central component of our own society's transformation.

3.2 Unity of Science?

What is the common characteristic of all the activities that we call science? The naive answer would be that science discovers new things and increases our knowledge. An immediate rebuttal to this characterization is the fact that, in the future, presumably some of our present scientific beliefs will sooner or later prove to be mistaken. We cannot know *now* what is amiss with our present beliefs (for then we would not believe they were knowledge); but it is quite reasonable to believe that at least some part of what we now believe will later turn out to be false. If we claim that the criterion for scientific research is that it increases our knowledge, we would be forced to say that some of what we now call research, though we do not know which part, is not actually research (since it does not guarantee knowledge). However, to condemn some part of scientific research after the fact is not particularly productive. What we need is a criterion for science and research that we can use without it being necessary to know exactly what is true or false. Therefore, the most common view among philosophers is that criteria for science should be formulated in terms of *methods* and not in terms of *results*. This means that a certain enterprise can count as science even though it later turns out that the theories this enterprise put forward are not quite correct. It also means that an enterprise that results in true propositions is not necessarily scientific; they could be arrived at merely by coincidence.

We come thus to the following central question: is there some common method used by all the sciences that can stand as a criterion, or even *the* criterion, for science? In response to this question, one can identify three possible answers.

Some rather cynical people would answer no: there is no such common property that distinguishes science from other activities. According to this cynical view,

what is called 'science' is arbitrary and the label is primarily used for enhancing the prestige of the enterprise so labelled.

According to another rather common view, inspired by Weber, scientific research is that which is typically practiced in universities, and there need be no significant similarity between all of these research enterprises beyond this. There is one crucial difference between the *natural* and *cultural* sciences, which is that the natural and cultural sciences use essentially different methods. The central activity of the cultural sciences is interpretation: interpretation of texts, events, actions, assertions, institutions, artefacts, etc. Such interpretation always contains a subjective element, as any interpretation must be based on the interpreter's prior understanding and cultural background. This in turn implies that interpretation is, in principle, not totally objective: i.e., not independent of individual perspective and culture. In contrast, in the natural sciences it is possible to be strictly objective in describing the phenomena one is studying in that two researchers can agree on an observation, irrespective of the cultural or theoretical background they might have.

Natural and cultural science are, according to this view, essentially different, and the fact that we call both types of enterprise 'science' does not imply that they share a method in common. On this view, those enterprises that do not fall into one of these groups, each of which is characterized by its methods, are deemed pseudo-sciences and hence are not sciences at all.

According to a third much disputed view, there is a least common denominator for *everything* that deserves the labels 'science' and/or 'research'. This is my view. The common denominator is the *hypothetical-deductive method*. Every enterprise that has the right to call itself science applies some variation of this method, though the details may differ from case to case. The hypothetical-deductive method, according to this view, should be used as the criterion for science; if an activity, taken in its entirety, does not abide by the requirements formulated in the hypothetical-deductive method, then it should be classified as a pseudo-science. I favour this third view.

Against those who claim that the cultural and natural sciences are fundamentally different, one may reply that the core activity of the major part of the cultural sciences–interpretation of texts, actions, and historical events–can be viewed as a sort of hypothesis testing where the interpretations play the role of hypotheses. These 'hypotheses' are tested against the types of evidence forthcoming in the texts, events, artefacts, etc. I shall explain this idea further in Sect. 3.5.

3.3 Hypothetical-Deductive Method

In order to more precisely describe what the hypothetical-deductive method is, I shall discuss a couple of examples from the history of science.

Example 1. The rejection of abiogenesis

For a long time people thought that worms, larva, etc. came into being spontaneously, rather than through the reproductive activity of parents. The idea was that these types of organisms were generated by decaying plant and animal matter. This idea is not entirely unreasonable given everyday observations of the natural world. Indeed, if one digs up a shovel's worth of topsoil, which is composed of plants in various stages of decomposition, one finds that it contains a large variety of small animals such as centipedes, spiders, larva, etc. (the existence of millions of organisms invisible to the human eye was, of course, not known in the mid 1600s or earlier). This view – theory is perhaps too pretentious a word – is called abiogenesis and was widely accepted up until the mid 1600s. However, a doctor from Florence named Francesco Redi (1621–1697) came to doubt abiogenesis. He tells (1668) how he

> . . ..began to believe that all worms found in meat were derived directly from the droppings of flies, and not from the putrefaction of the meat, and I was still more confirmed in this belief by having observed that, before the meat grew wormy, flies had hovered over it, of the same kind as those that later bred in it. Belief would be vain without the confirmation of experiment, hence in the middle of July I put a snake, some fish, some eels of the Arno, and a slice of milk-fed veal in four large wide-mouthed flasks; having well closed and sealed them, I then filled the same number of flasks in the same way, only leaving these open. It was not long before the meat and the fish, in these second vessels, became wormy and flies were seen entering and leaving at will; but in the closed flasks I did not see a worm, though many days had passed since the dead flesh had been put in them. . . ..
>
> To remove all doubt, as the trial had been made with closed vessels into which the air could not penetrate or circulate, I wished to attempt a new experiment by putting meat and fish in a large vase closed only with a fine Naples veil, that allowed the air to enter. For further protection against flies, I placed the vessel in a frame covered with the same net. I never saw any worms in the meat, though many were to be seen moving about on the net-covered frame.' (p. 33 ff.)

After these experiments, Redi (and eventually many others) became convinced that worms did not spontaneously arise (However, people continued to believe in abiogenesis of microscopic organisms until Pasteur successfully falsified this view).

The structure of Redi's argument is as follows: Redi begins by considering abiogenesis as a hypothesis (H), which he proposes to falsify with an experiment. We can formulate it as

(H) 'Worms arise spontaneously, in the absence of living animals, when meat and fish decompose.'

This is apparently one part of the abiogenesis view. From this hypothesis he draws the following conclusion:

(E) 'If I place meat and fish to rot in a covered jar, after a time I will be able to observe worms in the meat and fish.'

We call this sentence an Empirical Consequence (E) since, (1) it is a logical consequence of the hypothesis, and (2) its truth value can be decided by empirical experiments: i.e., we can prepare a situation such that we can show, through observation, whether the sentence is truth or false. Since the experiment shows

the E is false, we must infer that the hypothesis is also false; because E is a direct consequence of H. The reasoning can be formulated thus:

$$\frac{\begin{array}{l}\text{If H, then E}\\ \text{E false}\end{array}}{\text{H false}}$$

This is the age-old way of a displaying a logical argument; premises go above the line and the conclusion is placed below.

Once we have formulated an argument in this fashion, it is easy to see whether the argument displays *logically valid reasoning*: that regardless of what the letters H and E stand for, the conclusion follows from the premises. Put another way, an argument is valid only on the basis of its *logical form*, which in this case is called *modus tollens* (see Appendix).

Have we now definitively proven that H is false? No, we have not. That the form of an argument is valid means only that *if* we accept the premises, then we must accept the conclusion. Thus we can still doubt the conclusion of a valid argument by doubting the premises. In particular, if we read the latter part of Redi's text, we see that there is a possible objection to H. When Redi acknowledges that further experiments are required to check that excluding the air in his initial experiments did not unduly inhibit larval growth, he is indirectly saying that one can make the following objection to his initial experiment: it neglected a necessary condition for the production of larvae from rotting meat, namely the free circulation of air. This condition is not met in the first experiment. In this experiment, Redi tacitly assumed that air circulation was inessential for the production of larvae in rotten meat. Such a tacit assumption is commonly called an *Auxiliary assumption* (A). Thus it so happens that there is a hidden premise in the argument:

(A) The production of worms in rotten meat and fish does not depend on air circulation.

This auxiliary assumption is an essential step in the argument; and thus we modify our logical formulation to include it:

$$\frac{\begin{array}{l}\text{If H and A, then E}\\ \text{E false}\end{array}}{\text{Either H or A is false}}$$

Consequently, H can no longer be disproved by the falsity of E: it is possible that H is true and A is false. It was in order to exclude this possibility that Redi conducted the second experiment, in which the auxiliary assumption is not used.

This experiment is a beautiful example of an application of the hypothetico-deductive method. The term itself was not invented until modern times, but this way of thinking had already begun to be applied during the scientific revolution.

We have here seen that it is possible to falsify a hypothesis given that one has identified, and controlled, all auxiliary assumptions. But can one definitively confirm a hypothesis? This question leads directly to the next example.

Example 2. The link between diet and urine pH.

Claude Bernard (1813–1878)–a physiologist from Paris–conducted experimental studies of various physiological properties of mammals. He discovered that carnivores have sour and clear urine, while herbivores have alkaline and turbid urine. This apparent rule held without exception until he one day observed a deviation among some rabbits that had recently arrived from a breeder. These rabbits had sour and clear urine, even though rabbits are strictly herbivores and so should have alkaline and turbid urine by the apparent rule. Thus there must have been something special about these particular rabbits. This prompted Bernard to make the assumption that these rabbits had been starved during transport to his laboratory. Starvation causes the body to break down its own muscle tissue, which is the same as eating meat on the metabolic level and hence would account for the clear and sour urine.

In order to test his hypothesis, Bernard performed a number of experiments. The first was to feed the rabbits their normal diet and observe their urine. The result was that the rabbits' urine became alkaline and turbid again. Then he proceeded to observe the urine of starved horses and found that they also had clear and sour urine. Bernard later dissected the starved rabbits in order to see if they had indeed metabolized their own meat. After these various investigations, all of which confirmed his predictions, he was prepared to accept the hypothesis and claim that he had found the explanation to why the newly bred rabbits produced uncharacteristic urine.

Now let us analyse the structure of this argument. It is obvious that the hypothesis that Bernard decided to test is the following:

(H) Herbivores produce clear and sour urine when they fast.

This hypothesis was tested with respect to a general rule, or perhaps we should call it a theory; namely, that herbivores' urine is alkaline and turbid under normal circumstances, whereas carnivores' urine is clear and sour. This supporting theory is required in order to draw to any testable empirical implications from the hypothesis. As in the foregoing example, we can call this assumption, untested in this experiment, an auxiliary assumption:

(A) Herbivores' urine is normally alkaline and turbid, and carnivores' urine is normally clear and sour.

From these two premises, H and A, one can draw the following empirical consequences (and probably many others):

(E1) Rabbits with clear and sour urine produce alkaline and turbid urine when fed grass.
(E2) Horses who are starved produce clear and sour urine.

(E3) The inner organs of rabbits with clear and sour urine show signs that their bodies have metabolized their own muscle tissue.

All consequences could be observed to agree with his observations. Could we now say that Bernard has *proven* his hypothesis? The answer will depend upon what we mean by 'prove'. If by 'proof' we mean a definitive argument such that doubt of its premises and conclusions will never arise, then we cannot say that Bernard has proven his hypothesis. Another experiment with different herbivores could turn out differently and disprove the hypothesis. This is a general, purely logical, conclusion from the fact that the hypothesis–like every other scientific hypothesis–states something about all objects of a certain kind, while even the most thorough series of experiments can only cover a sample of that kind. In this case, Bernard's hypothesis is about all herbivores, though only a very few individuals of this kind were examined by him. Absolute certainty, even if a great number of experiments all agree with our hypothesis, cannot be had. What we eventually can say, at most, is that we have good reasons in believing that our hypothesis in correct. We can summarize the foregoing argument in the following way:

If H and A, then (E1, E2 and E3)
E1, E2 and E3 are all true

H is supported

I have drawn a dotted line (instead of a solid line as in the last case) under the premises to indicate that the inference is not logically valid. Thus, we cannot state that H is true; rather, only that H is supported to some degree: that it now has more credence than it initially had.

If we compare the two examples, we see an asymmetry; it is possible to absolutely falsify a hypothesis, but it is not possible to absolutely prove one. Indeed, we learned this earlier. There are many examples in science where a generally accepted theory had to be given up based on more careful experiments, even when the reasons for believing that theory for a long time were very good.

It is time to summarize what the hypothetical-deductive method amounts to:

- Put forth a hypothesis.
- Infer an empirically testable claim from the hypothesis and eventual auxiliary assumptions.
- Determine the veracity of the empirically testable claim via experiment and observation.
- Depending on whether the empirical implications are true or false, determine whether the hypothesis is supported or falsified.

There are complications with this model to which I shall soon return. Presently it is pertinent to present definitions of the key terms:

Theory$=_{\text{def.}}$ A set of statements whose relations are explicitly stated.

Hypothesis$=_{\text{def.}}$ A statement that (i) we are not entirely certain about, and (ii) which is used as a premise in inferring empirical consequences.

Empirical Consequence$=_{\text{def.}}$ A statement that (i) follows from the hypothesis and eventual auxiliary assumptions, and (ii) whose truth, under plausible circumstances, can be determined by observation.

Auxiliary assumption$=_{\text{def.}}$ A statement that (i) is necessary in order to infer an empirical implication from a hypothesis, and which (ii) is not tested in the given situation, but is rather assumed to be true.

Auxiliary assumptions can be of different kinds: results from other researchers, relevant everyday observations, assumptions about the working of measurement devices, or even whole theories. All can be called auxiliary assumptions, if they are used when inferring testable claims from a hypothesis. Sometimes the matter is so obvious that no one notices that there is an active hidden assumption until closer inspection of the argument. Thus, when analysing an argument it is sometimes useful to first identify the hypotheses and the empirical consequences. Once one puts these side by side, it is often the case that one notices a logical gap between the hypothesis and its supposed empirical consequences. One may then formulate the propositions needed to fill that gap, thereby making explicit the auxiliary assumptions on which the argument is built.

3.4 Hypothesis Testing in the Social Sciences

Many have claimed that the formulation, and subsequent testing, of hypotheses is unique to the natural sciences, and that the social and human sciences[3] use altogether different methods. Without going too deeply into this discussion, I would still like to give an example from the social sciences that fits quite well with the model for hypothesis testing previously described. It is contained in Emile Durkheim's book *Le Suicide (1897)*, where he discusses the social factors that govern incidence of suicide. One of his ideas was that suicide incidence is connected to the strength of social connections that exist between people. He called this strength of inter-personal connection within a society that society's *degree of integration* and he assumed that the higher the degree of integration, the lower the risk that one looses the desire to live and commits suicide. But how does one measure degree of integration in a society?

Durkheim thought that there was a difference in the degree of integration between catholic and protestant societies. Catholicism and Protestantism have

[3] The traditional English label for the humanistic disciplines is 'arts and humanities', but those who, like the present author, stress similarities among different sciences have began use the label 'human sciences' instead. In earlier times English used the term 'moral sciences' as contrast to 'natural science'. In German one usually distinguishes between 'Naturwissnschaften' and 'Kulturwissenschaften', i.e. using the word 'wissenschaft' as a general label for all systematic study at universities.

very different views as to the role of the church as a link between God and humans. There are many more church activities within Catholicism, and these have the effect of bringing people together, which culminates in a higher degree of integration than in Protestant communities (It should be kept in mind that it is the societal effects of Catholicism and Protestantism which are relevant, not any theological differences.). This assumption allows for the possibility of making comparisons. For example, in Switzerland many cantons are broadly similar in all respects except whether Catholicism or Protestantism predominates. If Durkheim's hypothesis is correct, then the suicide rate should be lower in the Catholic provinces than in the Protestant provinces. This was, in fact, shown to be the case via statistical data. We can reconstruct Durkheim's argument in the following way:

(H) The suicide rate is higher in societies with a low degree of integration than in societies with a high degree of integration.
(A) Catholic societies are more integrated than Protestant societies, all other circumstances being equal.
(E) Catholic societies have lower suicide rates than Protestant societies, all other circumstances being equal.

E is thus a prediction and we may compare with the actual records:
Observations:

Society type	Incidents of suicides (per million)
Catholic provinces in Switzerland	86.7
Mixed provinces	212
Protestant provinces in Switzerland	326

One has to admit that the empirical implications agree with these observations. Naturally, this does not prove that Durkheim's hypothesis is correct, since there could be many other factors involved in determining incidents of suicide. However, the observations strengthen his hypothesis.

3.5 Hypothesis Testing in History: The Wallenberg Affair

Testing of hypotheses is commonly viewed as completely distinct from methods in the human sciences, which all are different versions of interpretation of human artefacts. I will here exhibit a historical case, highly interesting in itself, which very clearly exhibit a structure that nicely fits into the schema of the hypothetico-deductive method.[4]

[4] I have not reconstructed the argument so as to exactly fit into the schema of HDM, it would make the section much less readable. But I hope the reader with only little effort is able to sort out the different points, in particular the auxiliary assumptions to the different hypotheses.

Raoul Wallenberg–a Swedish diplomat active in Budapest in 1944 whose job it was to help Hungarian Jews escape the Holocaust–was arrested by Russian troops when they arrived in Budapest. On January 16, 1945, it was reported that Wallenberg and his property had been taken into Russian 'protection'. On August 18, 1947, the Russians sent a note (after inquiries from the Swedish government) that stated that Wallenberg was not in the Soviet Union. Some years later, the Swedish government made another inquiry as to Wallenberg's whereabouts, since rumours that Wallenberg was alive and in the Soviet Union were circulating. These inquiries resulted in a memorandum that was sent to ambassador Sohlman on February 6, 1957, from Deputy Foreign Minister Gromyko:

> . . . In this matter, Soviet authorities have made the appropriate page-by-page review of archived records from assistant departments in certain prisons. As a result of this review, an archived document from healthcare services at Ljubljanka prison has been found, in which there is evidence that Raoul Wallenberg was admitted there. The document is in the form of a handwritten report – addressed to former Soviet Union Minister for State Security Abakumov and written by the chief of healthcare services at this prison, A.L. Smoltsov – containing the following information:
>
> 'I report that the prisoner of your acquaintance, Walenberg, died suddenly last night in his cell, apparently as a result of an induced myocardial infarction.
>
> In following your instructions to personally oversee Walenberg, I request instruction as to who will be responsible for performing an autopsy to determine the cause of death.
>
> 17.7.1947
>
> Chief of the prison's sanitation department
>
> Colonel of Medical Services – Smoltsov'
>
> On this report there is the following handwritten signature from Smoltov: 'Have personally informed the minister. Order had been given to cremate without autopsy. 17.7. Smoltsov.'
>
> There has been no more success in finding any other information in the form of documents or testimony, given the death of aforementioned A.L. Smoltsov on May 7, 1953.
>
> On the basis of what has been found, the conclusion drawn here is that Wallenberg died in July, 1947.
>
> Raoul Wallenberg was apparently arrested together with other prisoners in an area of Soviet war activity. At the same time, one can be certain that Wallenberg's later detention and the false information about him being sent to the Soviet Foreign Ministry for a number of years by former leaders of security agencies was the result of Abakumov's criminal activities. In response to his serious crimes, which aimed to cause all sorts of damage to the Soviet Union, Akabumov was, as you know, sentenced to death by the USSR Supreme Court.
>
> The Soviet government expresses its sincerest condolences in light of what has occurred and its deepest sympathies to the Swedish government and Raoul Wallenberg's relatives.

At the time of this letter, Gromyko's statement was generally accepted by the Swedish government and the Swedish people. However, over time this opinion changed and doubt of the truth of his statement gained traction. The reason for this was that reports began to come in from various freed Russian prisoners that Raoul Wallenberg was alive much later than the initial report indicated. Among these prisoners were four that had been stationed at Wladimir prison who stated that Wallenberg was there in the mid 1950s. These four prisoners included an Austrian whose name was not given, a Swiss named Emil Brugger, and two Germans named Horst Theodor Müller and Gustaf Rehekampff. The Austrian is the only person who

purported having personal contact with Wallenberg. Brugger said that he had been in the stockade with Wallenberg and the other two said that they heard from other prisoners that Wallenberg had been in Wladimir. In a letter sent to the Soviet government on July 17, 1959, the Swedish government wrote, 'Of course the Foreign Ministry must attach great importance to independent testimony of such a precise nature regarding Wallenberg's presence in certain prisons during certain years in the 1950s.'

What is the truth? Hans and Elsa Villius consider in their book *Fallet Raoul Wallenberg* ('The Case of Raul Wallenberg') three hypotheses:

H1. Smoltsov's report is correct and RW died July 17, 1947.
H2. Smoltsov's report is correct insofar as it describes death of a certain person, but it was not RW but another one, with an almost similar name.
H3. Smoltsov's report is a fabrication made by Soviet authorities.

The couple Villius conclude in their book that there is overwhelming evidence for H1, that Wallenberg actually died in Ljubljanka prison in 1947. They base this position on a critical analysis of the Russian letter, and transcripts of the hearings of the four persons who claimed to have certain information about Wallenberg at Wladimir prison, and a great deal of other information that I have omitted here for lack of space. Some of the arguments resulting from their analysis are the following:

- The only document regarding Wallenberg that was found in the Soviet archives was Smoltov's letter. According to Kosygin, who was prime minister of the Soviet Union at the beginning of the 1960s, there is no dossier on Wallenberg in the Soviet government's archives. This is explained by the fact that Abakumov tried to dispose of all traces of Wallenberg after his death. This fact is strengthened by numerous accounts of former fellow prisoners who were questioned about their knowledge of Wallenberg by high security officers, and were thereafter moved, isolated and forced to guarantee that they would never mention Wallenberg. Such evidence agrees with the claim that Smoltov's letter is the only document found, since it was never sent from the medical department of Ljublanka and Abakumov did not know that it existed. It is also in accordance with standard procedure that Smoltsov wrote down instructions given to him orally by his superiors.
- That Smoltsov received instructions from Abakumov to personally oversee Wallenberg agrees with Abakumov's later actions: as a diplomat, Wallenberg was a particularly sensitive case.
- It also agrees with Abakumov's order to cremate without autopsy.
- It also agrees with the fact that the Soviet Foreign Ministry, in 1947, claims that Wallenberg was not in the Soviet Union: Abakumov simply lied when the Soviet Foreign Ministry inquired.
- It also agrees with the fact that Wallenberg's dossier could not be found in the Russian government's archives. That it no longer existed in the beginning of the 1960s could only mean that a high-ranking official in the Russian government

disposed of it (and Abakumov was minister for state security during the years in question).

- The misspelling of Wallenberg's name is easy to understand, since in Russian there is a tendency to pronounce double consonants singly. The misspelling is actually an argument for the documents authenticity; for Smoltov it is natural to write 'Walenberg' when writing to someone from which he had just received verbal instructions. He wrote the letter as he had understood the name in a verbal context.
- The four people who came forth with information about Wallenberg's stay in Wladimir prison during the 1950s can be doubted. The anonymous man (his name is known to the Swedish Foreign Ministry, but has never been published) from Austria lacks credibility: among other things, he has given different versions of his connection with Wallenberg, he testified long after the time of the event in question and he was supposedly a cell spy according to another testimony.
- Brugger has reported having contact with Wallenberg in an article in Berner Tageblatt. However, this story differs substantially from that told in front of a Swedish representative less than 3 weeks after the Tageblatt article was published. The differences heavily diminishes his credibility.
- The two Germans relate their stories to the same source, a Georgian named Simon Goguberidze, who in turn said that he heard people say that Wallenberg was in Wladimir. There are circumstances making it probable that Goguberidze is the source of Brugger's claims as well, in which case Brugger's, Muller's and Rehkampffs' claims all originate from the same source. We thus essentially have one third-hand testimony, and nothing more.
- A series of prisoners from the Ljublanka and Lefortovo prisons in Moscow, who in various ways came to know of Wallenberg, were later moved to Wladimir. They were there isolated for a time, since they somehow knew of Wallenberg. Yet, none of these people, who were all questioned after being released, claimed that Wallenberg spent any time in Wladimir during the 1950s.

There is another series of arguments for the claim that Wallenberg died in 1947. However, in summary we can say that the hypotheses (i) that Wallenberg died in 1947 in Ljublanka, and (ii) that Abakumov tried to dispose of any trace of him, are greatly strengthened by these circumstances.

If we instead try the contrary hypothesis that Wallenberg did not die in 1947, we must assume that either Smoltsov's letter was a fabrication, or that this letter is about another person. If it was a fabrication, then we need the auxiliary assumption that the Russians have constructed the letter with the intent of deceiving the Swedish government. In such a case, is it really believable that KGB–the Soviet security agency–would have chosen to construct a single unofficial document, rather than producing nothing at all? If one had wanted to deceive the Swedes into believing that Wallenberg was dead, would it not have been more effective to fabricate an official autopsy report, written by some deceased medical officer, instead of a handwritten note relaying orders for cremation without autopsy? Is it

credible that KGB would have misspelled Wallenberg's name in such a fabrication? It seems the only reasonable answers to both questions is no.

If Smoltsov's note is about some other person, then there must have been someone besides Wallenberg, another sensitively handled prisoner, for whom Abakumov would require special attention, and with practically the same last name as Wallenberg. Is this auxiliary assumption really believable? No, it is quite improbable.

Elsa and Hans Villius thereby draw the conclusion that there is no reason to doubt the authenticity of the Smoltsov-document, and thus the hypothesis that Wallenberg died in Ljubljanka prison on the night of July 17, 1947, is strongly supported.

It can be added that the latest (last?) investigation of the case, carried out by a joint Swedish-Russian commission, came up with the same conclusion in January 2001 as had the Villius couple. The only significant difference between the two reports was some apparent evidence recorded in the commission's report that Wallenberg may have been poisoned and did not die a natural death.

3.6 Statistical Testing of Hypotheses

3.6.1 Bayesianism

To say that a hypothesis has been supported by an experiment is not very precise. It is, without doubt, desirable to have a measure for the credence a given hypothesis gains through experiment. Is it, then, possible to calculate the probability that a certain hypothesis is true, given a certain experimental result? This is a controversial question, involving the interpretation of *probability*.

The problem can formulated in the following way: we want to know the conditional probability (see Sect. 7.3) for a hypothesis being true, given a certain outcome of an experiment or measurement. Using the experiment's outcome, we can calculate this probability using Bayes' formula:

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

where P(H) is the initial probability for the hypothesis H, and P(H|O) is the probability for H given a statistical outcome O. Thus P(O|H) is the probability for the outcome O given that the hypothesis H is true, and $P(O|H_0)$ is the probability of an outcome O given that the null-hypothesis (the negation of the test-hypothesis) H_0 is true. I have here assumed that there are only two alternatives as regards hypotheses: the test-hypothesis and null-hypothesis. Investigations of the effectiveness of medicines against diseases provides good examples of this method. In such a case, we can formulate the hypothesis as

(H) The medicine has an effect on the progression of the disease.

The null-hypothesis is then,

(H_0) The medicine has no effect on the progression of the disease.

In order to calculate the probability of the hypothesis H, given an outcome O, we must determine the values of P(H), P(O|H) and P(O|H_0). The two conditional probabilities pose no fundamental problems, but how does one figure out the original probability P(H)?

If one interprets the concept of probability as a measure of the degree of subjective certainty–otherwise known as credence–one can imagine the following argument: We have a hypothesis H, which is true or false. If we do not have any specific information that argues for or against the hypothesis prior to our experiments, then we may assume that the initial probability that H is true is 50 %, and similarly that the probability of H being false (and thus the null-hypothesis being true) is also 50 %. We may assume that these probabilities represent our degree of belief in H and its contrary, and in the absence of any evidence one way or the other it is plausible that we have as much belief in one as in the other. Then we perform the experiments. With the help of Bayes' formula, we can now calculate the new probabilities; and thus generate a new probability distribution where the hypothesis is either strengthened or weakened due to its probability either going up or down, respectively. Furthermore, we now have a measure of the degree of strengthening/weakening of propositions.

Contrarily, if one does not interpret probability as a measure of the degree of subjective certainty but rather as an objective property, then it is difficult to motivate the claim that the original probabilities for a hypothesis being true or false is 50/50. The actual unknown probability could be anything, and Bayes formula is useless if the original probability is not known.

A fundamental question is whether it is scientifically justifiable to assume that probability is a subjective phenomenon. Proponents of this methodology often argue that no matter what the initial probabilities are, after a number of experiments, the probabilities of different individuals converge. This means that despite the subjective nature of initial probabilities, the long run results are still, arguably, scientifically objective.

However, this view is not universally accepted. The majority claims that we are not justified in any assumption regarding initial probabilities of a given hypothesis. What remains then is to restrict oneself to calculating the conditional probability of obtaining a certain statistical outcome, given the hypothesis.

3.6.2 Statistical Inference -Neyman-Pearson's Method

The more common procedure in statistical testing is to assume the null-hypothesis and calculate the conditional probabilities of the outcomes obtained. The following

(fictional) medical example may illustrate this approach. Suppose that a new medicine in the fight against AIDS has been created. Our hypothesis and null-hypothesis are the same as above:

(H) The medicine has an effect on the progression of the disease.
(H_0) The medicine has no effect on the progression of the disease.

(For the sake of simplicity, I will ignore the possibility of negative effects.) The null-hypothesis implies that the difference between the groups is exactly zero, which is not particularly reasonable. In a more realistic experiment, one would formulate the null-hypothesis in terms of the difference being less than some arbitrarily chosen value. However, I shall ignore this complication in this case.

In order to determine the possible effects of the treatment, nine patients were given the new medicine while nine other patients were used as a control group. The patients in both groups are chosen at random. After treatment the results were:

	Treated	Untreated
Alive	6	1
Dead	3	8
Sum	9	9

A statistical test shows that if the null-hypothesis were true, then the probability of the above outcome would be less than 1 %. Is this sufficient for abandoning the null-hypothesis, drawing the conclusion that the treatment is effective? Most would probably answer 'Yes', but where does one draw the line? How low a probability for the experimental outcome, conditional on the null-hypothesis, is required before one considers that null-hypothesis refuted?

The choice is determined by balancing two risks against each other: the risk of accepting a false test-hypothesis (by rejecting a true null-hypothesis) and the risk of rejecting a true test hypothesis (by accepting a false null-hypothesis). In practice, one often takes the threshold to be 5 %, 1 % or 0.1 %. If the probability of some experimental outcome, given the null-hypothesis, is at most 5 %, then one says that the result is one-star significant. If it is at most 1 %, then the result is called two-star significant; and similarly, 0.1 % is three-star significant. The methodological rule is that one should determine the level of significance before the experiment is done and then accept or reject the null-hypothesis as the case may be.

Plainly, this rule is not certain to give correct conclusions. The probability for the outcome, conditional on the null-hypothesis, can be higher than 5 % in spite of the null-hypothesis being false, and it is possible that the significance value is less than 5 % (or any other chosen limit) though the null-hypothesis is true. We talk about two types of errors:

1. We reject H_0 while in fact it is true: *error of the first kind*.
2. We accept H_0 while in fact it is false: *error of the second kind*.

Hence, making an error of the first kind means accepting a false test hypothesis, an error of the second kind means rejecting a true test hypothesis.

The conclusion of this discussion is that if we are not prepared to say that a hypothesis has some initial probability, then using statistical methods does not allow one to claim any more than in earlier examples: experimental results merely strengthen or weaken test-hypotheses in a rather imprecise and hand-waving fashion. However, if one allows that test-hypotheses do have known initial probabilities, then one can make precise the degree to which a test hypothesis is strengthened by some experimental result. For instance, in the immediately preceding example, if we, being subjectivists, assume that the probability distribution between hypothesis and null-hypothesis is 50/50 before the experiment, then we can calculate, using Bayes' formula, the hypothesis' probability given the above results to be 99 %. How sensitive is this probability for variations in assumptions about prior probabilities? Not much; if, for example, one is rather pessimistic and say that the prior probability for the hypothesis is only 10 %, its probability after the test is instead 87 %.

Comparing these two methods from a purely epistemological perspective, I see no substantial difference. Both contain an element of subjective decision. In both there is a risk of making a mistake, which as a matter of principle cannot be completely eliminated.

3.7 Unacceptable Auxiliary Assumptions: Ad Hoc-Hypotheses

As we saw in the Abiogenesis example, one can save a hypothesis from falsification if one can find an auxiliary assumption to which one can direct modus tollens. Recall that if two or more hypotheses are required in order to infer an empirical consequence, and if that empirical consequence is false, then one cannot logically determine which of the hypotheses are false. Thus, if one is loathed to reject a certain hypothesis, then it can be saved from falsification by adding new auxiliary assumptions. However, such an attitude is hardly conducive to good science; rather, one should be prepared to reject or revise even the most well established theories if there is enough evidence against them.

An auxiliary assumption that is proposed only to save another hypothesis from falsification is called an *ad hoc-hypothesis* (*ad hoc* = for this). It is obvious that one cannot accept ad hoc-hypotheses in scientific argumentation. Then how does one distinguish between acceptable auxiliary assumptions and unacceptable ad hoc-hypotheses? Karl Popper, among others, have argued that ad hoc-hypotheses differ from acceptable auxiliary assumptions in that it is possible to independently test the latter by conducting experiments with no connection to the original ones, but not possible to do the same for the former. A couple of examples from ancient astronomy should illustrate this idea.

Example 1 According to the ancient world-view, the Earth was the centre of the universe and all celestial bodies revolved around it in perfect circles at constant

speeds. In the heavens, all motion was perfect. Observation of the sun, moon and the stars seemed to agree with this theory.

However, the motion of planets, such as Mars, does not fit the theory. Viewed against the fixed stars Mars sometimes appears to move 'backwards'.

Every day Mars appears to move one revolution around the Earth, so that against the fixed stars it moves slowly from west to east. However, sometimes Mars 'backs up' in what is known as retrograde motion, as can be seen in Fig. 3.1. Does this mean that Mars sometimes slows down? No, such imperfect motion was not reasonable for the ancient astronomers.

Ptolemy (Greek astronomer, active in Alexandria circa 90–168 A.D.) construed an improved system that agreed with the observed retrograde motions. He claimed that each planet moved in a small circle, an epicycle, around a mathematical point located on the planet's main circular path around the Earth. He also placed the Earth a little bit away from the centre of the main circular path. When you combine these two circular motions–the epicycle and the main circular path–with suitably chosen radii, you get the apparent motion of the planet. Seen from Earth, Mars is in retrograde motion when it moves in a direction opposite to that of its main path. Since one has no restrictions on the choice of the epicycle, one can construct a system that agrees perfectly with the observed retrograde motion. However, given that there is no argument for why Mars should rotate around a mathematical point in space and since there is no independent way to test the hypothesis, the assumption of epicycles is an ad hoc-hypothesis (Fig. 3.2).

Another problem in ancient astronomy was to explain the existence of two kinds of eclipses, total and annular. It is noteworthy that the apparent sizes of the sun and the moon, as seen from Earth, are very nearly the same; hence when the Earth, Moon and Sun are all aligned, the moon is just the right apparent size to hide the Sun from the Earth in what we know as an eclipse. If, as the ancients believed, the Sun and Moon's orbits of the Earth were perfectly circular, with the earth at the centre of that circle, then every eclipse would be the same. Yet, all eclipses are not exactly similar; some eclipses are total–where the sun is completely obscured from the earth–and some are annular–where one sees a thin ring of the sun past the edge of the moon. We now know that this difference in eclipses is due to our Earth orbiting the sun in an ellipse with the sun at one of the ellipse's foci; however, this explanation was not available to the ancients as such elliptical motions were incompatible with the their divinely perfect heaven; according to ancient views, an ellipsis is a non-perfect, misshaped circle.

One hypothesis that was proposed at the time to explain this conflict between theory and observation was that the moon sometimes shrinks and expands! But the change in size is so small that the only way to notice it is to observe its effect during solar eclipses, which implies that there is no way of independently testing this assumption. Thus, this assumption provides another example of an ad hoc-hypothesis.

However, if we consider this matter from a present day technological point of view, the situation is quite different, since nowadays it is relatively easy to

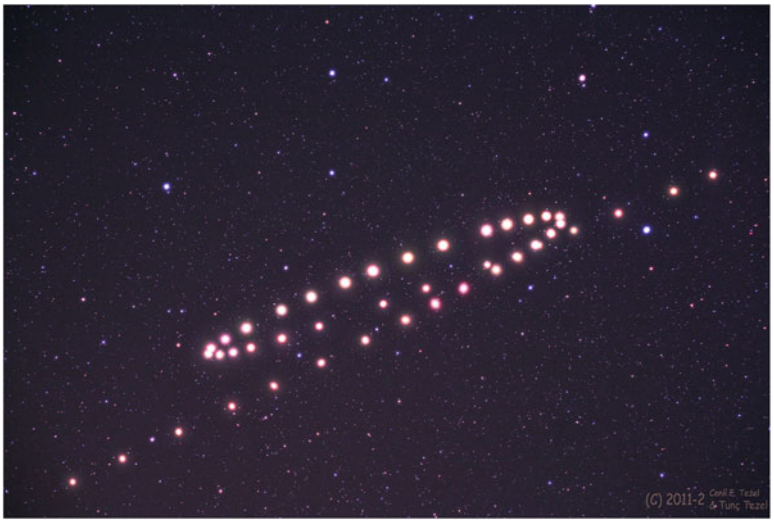

Fig. 3.1 *Retrograde motion*. *The motion of Mars* from late October 2011 (*top right*) through early July 2012 (*bottom left*) (Reproduced with kind permission of Tunc Tercel)

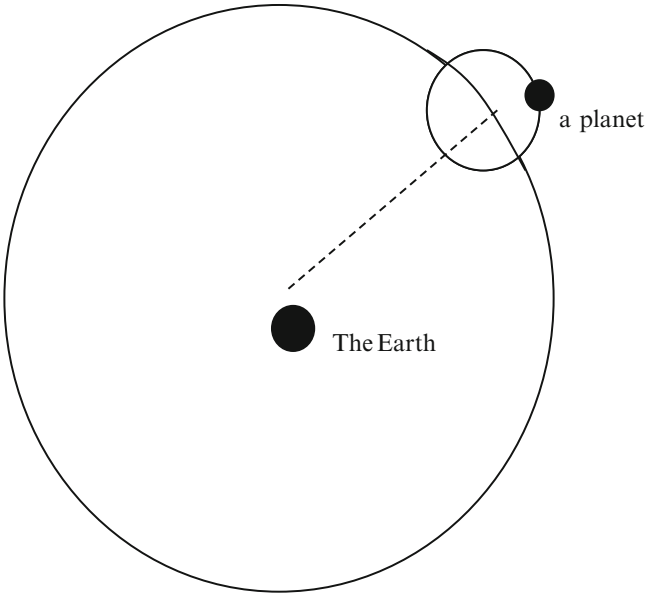

Fig. 3.2 The basic principle of Ptolemy's planetary system

accurately measure the diameter of the moon, for example using lasers. This shows that the question of whether the auxiliary assumption is testable or not is not so easy to determine as it might first appear. Criteria regarding independent testability do not seem to be absolute. Rather, they have to do with what is possible at a particular time. Even though all scientists have a strong intuition about what should count as an ad hoc-hypothesis, it is thus difficult to give a definite criterion.

3.8 Summary

The least common denominator for all sciences is that hypotheses are formulated and tested. This is meaningful only if one is prepared to change one mind after testing, to admit that even one favourite hypothesis was wrong. This state of mind is a crucial part of the scientific attitude.

Testing hypotheses against observations require auxiliary assumptions. They can also be tested, of course, although in a particular case of hypothesis testing they are taken for granted at the beginning.

The result of the test is either that the predictions and observation reports are compatible, or that they conflict. In the former case one may be justified to say that one's hypothesis is supported. In the latter case one must reconsider something; one must reject either the hypothesis, some auxiliary assumption, or the observation report. If one decides that the hypothesis is to be rejected, one has falsified it. It is thus clear that falsification of a hypothesis is no automatic inference from the test result; it is the result of a judgement, all aspects considered.

Exercises

Below are three short summaries of popular scientific articles. Give an analysis of the argumentation in each abstract using the concepts of hypothesis, auxiliary assumption and empirical consequence!

1. It has been generally assumed that the dinosaurs were cold-blooded animals. This assumption has been criticized by Stephen Jay Gould. He claims that they must have been warm-blooded because:

   (a) The temperature of cold-blooded animals' varies with outside temperature. Cold-blooded animals that live in areas with large changes in temperature between summer and winter get growth rings in the outer parts of their skeletons, similar to the growth rings of trees. Warm-blooded animals do not get such rings since they maintain, more or less, a constant body temperature. Dinosaurs that lived in areas of varying temperature did not have such growth rings.

   (b) Large cold-blooded animals do not live near the Polar Regions since they cannot get enough warmth during the short winter days and are too large to find secure shelter. Some large dinosaurs lived so far north that they must

have spent long periods of time without sunlight; and thus without an external heat source during the winter.

(c) Modern reconstructions of the anatomy of various dinosaurs show that many large dinosaurs are similar to present-day mammals (which are warm-blooded), both in regard to skeletal muscles and the proportions of various limbs.

2. The following text is an excerpt from the article 'The Confirmation of the Continental Drift' in *Scientific American*, April 1968. It discusses the possibility that Africa and South America were once joined, an idea that seems quite natural when looking at the contours of these two continents.

> Of particular interest for us was the sharp borderline between the 2000 million year-old geological area in Ghana, the Ivory Coast and further west, and the 600 million year-old area of Benin, Nigeria and further east. This borderline runs southwest towards the ocean near Accra, Ghana. If Brazil (and all of South America) had been part of Africa 500 million years ago, the borderline between these two areas would have run through South America near the city of Sao Luis on the north coast of Brazil. Therefore, out first task is dating the areas near Sao Luis.
>
> To our pleasure and surprise, the results fell into two groups: 2000 million years on the west side and 600 million years on the east side of the borderline where we had expected it to be. It appeared as if a 2000 million year-old piece of Africa had landed on the South American continent.

3. Over the years 1844–1848, Ignaz Semmelweis worked as an obstetrician in Vienna. The hospital where he worked had two maternity wards, each taking approximately 3000 admissions per year. It so happened that the proportion of deaths due to puerperal fever was markedly different between the two wards:

	Ward 1 (%)	Ward 2 (%)
1844	8.2	2.3
1845	6.6	2.0
1846	11.4	2.7

A great many explanations of this difference were proposed, such as incorrect birthing position, poor diet, 'atmospheric conditions' etc. All of these could be discarded quite reasonably. Then Semmelweis got the idea that some 'cadaveric material' from dead persons had been transported to Ward 1, which agreed with the fact that at Ward 1 medical students were instructed after having partaken in autopsies, while at Ward 2 midwives, who had no contact with cadavers, were instructed. As an experiment, Semmelweis had all who came to Ward 1 from an autopsy wash their hands in chlorinated lye. The results of this experiment on the number of deaths due to puerperal fever were the following:

	Ward 1	Ward 2
1848	1.27 %	1.37 %

(Unfortunately, Semmelweis' experiment was not accepted as a good argument for continuing to wash one's hands after autopsies. Semmelweis was dismissed and the use of disinfectants ceased. It took more than 20 years until the practice of disinfection became commonplace, thanks to Pasteur's discovery of bacteria.)

Further Reading

General

Hempel, C. (1966). *Philosophy of natural science*. Englewood Cliffs: Prentice Hall.

Hermeneutics as Hypothesis Testing

Chesterman, A. (2008). The status of interpretive hypotheses. In G. Hansen, A. Chesterman, & H. Gerzymisch-Arbogast (Eds.), *Efforts and models in interpreting and translation research: A tribute to Daniel Gile* (pp. 49–61). Philadelphia: John Benjamins.

Statistical Testing of Hypotheses: There Are Numerous Textbooks on Neyman-Person's Method, e.g.

Moore & McCabe. (2005, 2009, 2012, 2014). *Introduction to the practice of statistics*. New York: W.H. Freeman and Co, chapters 6 and 7.

Bayesian Inference, See e.g.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Burlington: Academic.