Experiments, models and methodology

Part 1 & Part 2

By Till Grüne-Yanoff
Version: 2019-09-06

Contents

Part 1 - Methodology

1.1 Introduction – Goals of science

Welcome to Theory and Methodology of Science with Applications. This text will discuss methods: methods for producing evidence, methods for representing phenomena, methods for making inferences and for testing theory. However, the purpose is not to discuss the technical detail of these methods. Rather, the focus is on **methodology**, the systematic assessment and justification of method choice. Scientists typically must choose between alternative possible methods when doing their work. What the relevant alternatives are depend partly on what goal a scientist wants to achieve with their work. **Typical goals in science** include prediction, explanation and design.

Methodology: The systematic assessment and justification of method choice

Typical goals of science: Prediction explanation and design

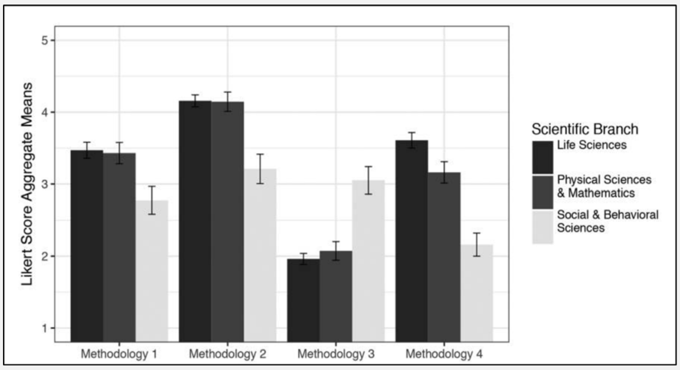
However, you cannot expect that specifying your goal immediately determines what method to choose. Instead, it will require considering the reasons why one method might serve one's goal better than another, which, in turn, might require specifying one's goal more precisely, or learning more about the methods and the context in which the methods are supposed to be applied. A couple of examples might illustrate this point. For example, your goal might be to compare how much wear different road covers can sustain under real-world traffic conditions. Which method for producing evidence should you choose? One alternative is to design a laboratory experiment. This gives you a lot of control over background variables. But the kind of use you can simulate in the lab might be quite different from the real-world traffic conditions.

Alternatively, you could run a field experiment, constructing real stretches of road with the different road covers you want to test, and let public traffic do its damage. Here, the test conditions are realistic, but you might not be able to control all background factors to your satisfaction. These different types of experiments and their respective advantages will be discussed in much more detail later on, so do not worry if this sounds strange to you right now. The case just illustrates how a scientist must choose between different methods – in this case, different types of experiments – and that this choice requires considering the advantages and disadvantages of each of these methods for one's goal.

Take another example. Your goal might be to model a benzene molecule in order to simulate a chemical synthesis with it. Should you do this by constructing a structural formula, or by building a quantum model? The former is simpler, and this makes it easier to analyze. But the latter is more accurate. Again, understanding how important these respective advantages are will depend on how well one understands one's goal. Or take this example. You want to test the claim that street level concentration of exhaust fumes causes respiratory diseases. Which

Practice Snapshot: Multiple Conventions

A recent study asked 346 scientists from different disciplines the following four questions about their method choices. Their answers show clearly that there are different conventions of what methods to use in different scientific disciplines.



Robinson, B., Gonnerman, C., & O'Rourke, M. (2019). Experimental Philosophy of Science and Philosophical Differences across the Sciences. *Philosophy of Science*, *86*(3), 551-576.

method for hypothesis testing should you choose? Should you use a statistical test, or not? And if yes, should you perform a significance test, or employ a Bayesian approach?

These are typical method choices that scientists face. How do they decide between these alternatives? Is there a way to determine what method is rational to choose? There are three common answers to this question. The conventional answer suggests choosing the methods your teachers and your peers are using. So according to the **conventional view**, within a discipline, or maybe just within a research group, the question of method choice does not arise. Everybody just follows the established convention, just like everybody in Sweden drives their car on the right side of the road. An example of such a convention, even today, is perhaps that most scientists use statistical tests for all their hypothesis evaluations and would never even think about writing a paper without such a statistical analysis.

Conventional view of methodology: Choices between different methods are justified by what is seen as "common practice" in the field.

Methodology 1: "Scientific research (applied or basic) must be hypothesis driven."

Methodology 2: "In my disciplinary research, I employ primarily quantitative methods."

Methodology 3: "In my disciplinary research, I employ primarily qualitative methods."

Methodology 4: "In my disciplinary research, I employ primarily experimental methods."

The problem with the conventional view is that it renders scientists less flexible in finding alternative viewpoints that could support criticism of their own discipline or group. If the use of a method contributes to an error in the research result, then it is much harder for a group that uses this method by convention to discover this mistake than for a group that does not follow a conventional method. An additional problem is that the conventional answer makes interdisciplinary work really hard. If two disciplines have different conventions about what methods to use, then bringing these disciplines together is like letting people drive on the right and on the left on the

Practice Snapshot: Funding for Denial

To get a feel how much money industry is paying for the services of sowing doubt, read this article:

Sharon Begley "Global Warming Deniers: A Well-Funded Machine". *Newsweek* August 3rd, 2007. Link

same street – that inevitably leads to mayhem, as long as people have no other reason for their choice than "but we've always done it this way!". To be able to collaborate with people from different disciplines, who follow different conventions, one must go beyond conventional methodology and instead make transparent the underlying justifications for choosing one's methods – which are understandable to people from different disciplines, and perhaps may even convince them.

The best-results view of methodology: Choices between different methods should be determined based on what produces desired results.

An alternative answer to the question of how one should choose methods is to say one should choose the method that gives the **best results**. This this answer is on the right track, but it suffers from excessive vagueness. The idea here is that one chooses that method that serves some purpose best. But this purpose is not sufficiently specified. In particular, science often involves long-term plans, so that the final outcome from choosing one method instead of another are not known or uncertain. Take, for example, the Large Hadron Collider (LHC) near Geneva. Its construction so far has cost more than 7,5 billion euros – and the next-generation collider, now in planning stage, will cost even more (see box). What material benefits the research done here will eventually have, however, is entirely unclear – CERN's director Bordy is very explicit about the deep uncertainty about the potential benefits. Thus, choosing this method of investigating subatomic particles might not be justifiable by its material outcomes.

Furthermore, who judges the best in "best results"? Some scientists have been making very good business in creating doubt about scientific results in the service of lobbyists and industry. Their choice of methods does not stand up to scientific scrutiny. But their choice presumably yielded results they considered best for themselves. Hence, we need to be more precise with regards to what we mean by justifying one's method choice.

The epistemic tool view of methodology: Choices between different methods are determined on what one has the most reason to believe will satisfy ones epistemic goals.

Instead, I will argue for another justification of method choice, which I call the **epistemic tool account**. According to this account, one chooses that method one has most reasons to believe is the best tool for one's goals. Scientists might pursue different types of goals, for example prediction, explanation, or design. For a given method choice problem, once these goals are sufficiently specified – e.g. what is to be explained or predicted, what rate of error is permissible in the prediction, etc. – one can ask which conditions one must satisfy in order to reach these goals. For example, to explain a phenomenon Y requires identifying the cause X that produced Y. To understand this, we need to analyze what the concept "explanation" means (I will do this later in this course, so do not worry if you disagree

Practice Snapshot: Costs of the LHC

"Cern's director for accelerators and technology, Dr Frédérick Bordry, said that he did not think that £20bn was expensive for a cutting-edge project, the cost of which would be spread among several international partners over 20 years.

He added that spending on Cern had led to many technological benefits, such as the World Wide Web and the real benefits were yet to be realised.

"When I am asked about the benefits of the Higgs Boson, I say 'bosonics'. And when they ask me what is bosonics, I say 'I don't know'. But if you imagine the discovery of the electron by JJ Thomson in 1897, he didn't know what electronics was. But you can't imagine a world now without electronics."

Cern plans even larger hadron collider for physics search, BBC News, January 15th, 2019. Link.

or don't understand yet why explanation requires identifying causes). With this analysis, we can then ask what advantages (and disadvantages) the available tools have for identifying causes in the given situation: for instance the tool of experimental methods (if experimenting is possible) or perhaps specific statistical analyses of observational data? Giving the reasons to believes that a certain method M is the most advantageous for one's goals **justifies** choosing M.

Justification: Reasons for believing a certain **proposition** to be true

Instrumental reasoning: Providing justification to fulfil a clearly stated goal.

Proposition: The information expressed in a statement or a claim. "Snow is white" and "Snö är vitt" contains the same proposition, but are two different statements.

The epistemic tool account is a version of **instrumental reasoning**: it justifies choosing a method M by identifying the reasons we have to believe that M is the best tool is for a given goal in a given situation. Instrumental reasoning is widespread in everyday thinking: when choosing between colleges programs, or between burger and broccoli for dinner, people reason instrumentally, considering what goals these programs or diets serve (or do not serve). That this reasoning is so widely and successfully used is an advantage for the epistemic tool account – it indicates that this form of justification is tried and widely accepted. Note however that the outcome-oriented approach also is instrumental. The difference is that the outcome-oriented approach remained unclear on what goals it supports, while the epistemic tool account makes scientific goals specific, and it focusses on the reasons one has for choosing a method, *before* implementing it.

Additionally, in practice, there are often other considerations that are relevant for such an argument, that are not about fulfilling scientific goals. For instance, say you want to explain a phenomena Y and you know that an explanation requires to accurately identify the cause, X, that produced Y. You furthermore know that experimental methods allow for you to determine the cause of Y. If there are no other relevant considerations, then you have a good overall reason to choose the experiment as your method. However, imagine that the phenomena Y which we want to explain causes certain human patients to die. While we of course remain interested in explaining Y, there are now *legal* and *ethical* restrictions on what experiments one can perform that might cause human deaths. In this case,

Further reading

You find a detailed discussion of the epistemic tool account – where methods are described as heuristics, and methodology as guidelines for the appropriate use of such heuristics – in:

Hey, S. P. (2014). Heuristics and meta-heuristics in scientific judgement. *The British Journal for the Philosophy of Science*, *67*(2), 471-495. Link.

some reasons are favoring one method, some another. We now need to weigh these reasons against each other – for example, by writing a pros and cons list for each of the alternative methods in order to justify choice between them.

To conclude, this course focuses on methodology, the systematic assessment and justification of method choice. Methodology ask questions like: what relevant methods do we have available to reach a scientific goal? What reasons speak for or against choosing any of these alternatives? How should one weigh these reasons to form an overall decision? Methodology must be distinguished from describing methods on the one side and doing general philosophy of science on the other. *Descriptions of methods* concern the design and implementation of particular research strategies, and often focus on technical skills. Examples include how to program simulations, how to set up measurement instruments, how to calculate variables, or how to implement experimental design. It does not include method comparison, choice, or justification. On the other side, *general philosophy of science* asks many questions that are not directly related to method choice and its justifications either. These include questions about the nature of scientific knowledge and scientific theories, or deal with various forms of skepticism. While undoubtedly fascinating, this text only addresses such questions to the extent that they are relevant for methodological issues.

Reflecting on what you just read

- Consider a recent scientific article you have read. What was the authors' scientific goal(s)? What alternative methods do *you* think the authors had a choice between? Did the authors describe such a choice in their article?
- Describe a typical methodological convention in your field. Check with students from other disciplines whether they have the same convention or not!
- Scientists are often skeptical about philosophers' views on how scientist should do their work – arguing that scientists should know better than philosophers (who after all typically do not work in science). That seems to be a reasonable criticism, at least against those philosophers who propose fundamental and unchangeable methodological principles. Do you think that the epistemic tool account is affected by this criticism, too?

1.2 Scientific Knowledge

The main focus of this text is methodology, and rational method choice depends on the goals that scientists want to reach with these methods. Crucially, methodology, as it is presented here, does not consists of a set of *first principles* how science should be done – rather, it indicates which method is rational to choose, depending on the researchers' goals and the conditions under which they operate. The most important goals in science are prediction, explanation, and design. Being able to **predict** an event X means to know that X will occur at some time, t, in the future. Being able to **explain** why X happened rather than something else means to know what causes produced X. Finally, being able to **design** an artifact X means to know that X satisfies certain desirable functions.

Prediction: Knowing that an event will occur at a future time.

Explanation: Knowing what causes produced an event.

Design: Being able to design an artefact that satisfies certain functions.

These goals all share a common necessary ingredient: scientific knowledge. One must *know* to predict, explain, or design. And this knowledge is acquired by applying the best scientific methods. But what is knowledge? Philosophers have offered many accounts, but the answer the ancient Greek philosopher Plato gave is still considered one of the most useful. Plato argued that **knowledge** is true, justified belief. He answered the question "What does it mean that a person knows something?". The answer consists in giving necessary conditions for a person, let us call her K, to know something (that the moon is spherical, for instance), which we can call P. Then Plato's definition states that K knows P if and only if: (1) that K beliefs P to be true; (2), that K is justified in believing P to be true; and (3), that P in fact is true. These are called necessary conditions, because each of them must be satisfied for K to know P.

The classical definition of knowledge: True, justified belief.

However, even together they are not sufficient. There are cases that satisfy all these conditions, and yet most people would intuitively say that K does not, in fact, know P – search "Gettier Problem" if you are interested in such cases.

Plato's answer in the first place is just a definition, a statement what it means to know. One shouldn't just accept definitions from anybody, not even from an authority like Plato. Rather, one should ask what the arguments are for accepting such a definition. Sometimes the argument for a definition of a term is that almost everybody thinks that that term has this meaning. While such **lexical** definitions are useful to understand what others are saying, philosophers are rarely content with them. For example, even if everybody thought that the definition of a whale includes categorizing it as fish, philosophers – or scientists, for that matter – would not accept such a definition. Instead, they would argue that the definition of "whale" should not include whales being categorized as fish, even if everybody else thinks so. Such a **stipulative** definitions rest on reasons and arguments, not on common usage. Plato's definition is clearly stipulative. It is not because of this common usage that knowing requires satisfying these conditions, but because Plato and many philosophers since thought that there are good reasons for these conditions.

Lexical definition: A definition that intends to **c**apture common usage of a term.

Stipulative definition: A definition made for a specific purpose, in a specific context.

When investigating these conditions in more detail, however, it is clear that there are good reasons to adopt them as a definition of knowledge, and in particular for *scientific* knowledge. The first condition requires that for K to know P, she must **believe** it. Belief is the state of mind in which a person thinks something to be the case. Thus, this condition ties knowledge to individual state of mind. Consequently, P cannot be known without someone believing that P is true. This seems to be right – knowledge is something a person has, not something out there in the world. Scientists, when investigating the world, do not discover knowledge. Rather, they collect information, identify evidence, and craft arguments that produces knowledge by convincing themselves and others to believe that certain claims are true. Through the belief condition, it also becomes clear how knowledge affects people's behavior. People hold beliefs about what

Belief: A state of mind where a person considers a proposition to be true.

consequences their different choice options have. And based on these beliefs–combined with values, desires, and goals – they decide how to act. Knowledge has this guiding influence on actions because it is constituted by belief. But beliefs come in degrees: sometimes our belief that a certain claim is true is very strong. We mark this by saying something like, “I fully believe that”. In other cases, we hold less certain beliefs, which we mark this by saying something like: “I tend to believe this”, or even, “this is hard to believe”. It seems that many of our beliefs are less than certain. There often is the possibility, and sometimes the very real probability, that they are wrong. In that case, we believe a claim up to a certain degree. The question is, however, whether such uncertain beliefs are compatible with the idea of scientific knowledge. To put it differently, does science require certainty?

The first thing to note is that Plato's definition is compatible with **uncertainty**: it only says that a person has to have a belief, not that it has to be certain. If knowledge is belief and beliefs come in degrees, then knowledge comes in degrees too. But is this how it should be? The answer is yes. Scientists often make claims that they are less than certain about–whether it concerns direct observations, measurement instruments, results of calculations, or claims about theory. These claims, if justified and true, nevertheless constitute knowledge. If you think that this sounds strange, be clear that *certainty of belief* and *truth* are two completely different things. One can have certain beliefs about a false claim. Most pre-modern people, for example, presumably believed with near certainty that the sun revolves around the Earth. And one can have uncertain beliefs about true claims. For example, if a fair coin is to be flipped 100 times, you might believe that it will fall heads 50 times. And you might later find out that indeed this is true: having tossed it 100 times, you observed it fall heads 50 times. But before you checked, you would not likely believe this with certainty. That is, you have an uncertain but true belief. In fact, acknowledging uncertainty in one's knowledge is very important in science. It is possible that even our best theories are wrong. But that does not mean what various self-declared anti-scientists want it to mean when they go on about evolution or climate change just being a theory. We have good reasons to be very confident in these theories without denying that there is a (remote) possibility that they might be wrong. We can acknowledge that we are less than certain in our knowledge without this denying its truth.

Uncertainty: A belief is uncertain if it is not absolutely justified.

Moving on, the second condition is that a belief must be **justified** to constitute knowledge. A justification is evidence or some other reason that can be put forward in defense of believing a claim to be true. Only beliefs properly justified might constitute knowledge. Many beliefs that people hold lack a justification. A lottery player presumably believes that the numbers she records on the slip will win, but she has no justification for this belief. Thus, even if the number of wins – thus her belief was true – she did not know that this number would win. Other cases are more complex. If my upstairs neighbor waters her plants and splashes my window, then if I believed that it rained because of the drops on my

Justification: Reasons for believing a proposition to be true.

window, my belief is not justified and does not constitute knowledge - even if it actually rained! The observation of *those* drops (coming from my neighbor's watering can) does not constitute evidence for the claim that it rained. Thus my belief is true (as it actually rained), but the kind of evidence that I point to in defense of that claim is not an actually justification. So my belief that it rained is true but unjustified belief, and hence not knowledge.

Furthermore, not all relevant reasons for beliefs are justifications for scientific knowledge. If one strongly wishes something to be true, the happiness one feels if one believes that it is true might be a strong reason why someone believes it. But such wishful thinking does not constitute a justification of scientific knowledge, nor do religious convictions. One can, of course, have these beliefs. But they do not constitute scientific knowledge.

A belief is justified if evidence or some other reason can be put forward in its defense. But do we actually have to have this evidence or other reason "in hand", so to say? Most Swedes, for example, know that Gustav Vasa fled on ski through Dalarna from men loyal to king Christian. Yet what kind of justification can they show for this belief? Of course, no living person today has witnessed this event, so they instead must rely on documents or reports about it. But most Swedes learn about this event in elementary school – and while they remember the story, they most likely don't remember where they learned it from. So they cannot point to the actual reason that made them form this belief – and that might seem to imply that their belief is not justified. The position that I defend here, called "weak internalism of justification", contradicts this impression. Someone's belief is justified only if one can become aware by reflection of what the belief's justifiers are. This reflection might involve considerable time and effort (for example, researching what history textbook you had in fourth grade, etc.); the point is that one is justified even if one does not have evidence or reasons "in hand" – it is enough that one could get hold of them, if one invested sufficient effort. Thus, when challenged on a certain claim, one must in principle be able to produce the justifiers for them. It will not do to insist that these justifiers are to be found somewhere, even though one cannot, even with one's best efforts, locate them anymore.

Practice Snapshot: From uncertainty to falsity?

Here's an example of an argumentative strategy that implicitly assumes that knowledge must be certain:

"How did life originate? Evolutionist Professor Paul Davies admitted, 'Nobody knows how a mixture of lifeless chemicals spontaneously organized themselves into the first living cell.'[1] Andrew Knoll, professor of biology, Harvard, said, 'we don't really know how life originated on this planet'.[2] A minimal cell needs several hundred proteins. Even if every atom in the universe were an experiment with all the correct amino acids present for every possible molecular vibration in the supposed evolutionary age of the universe, not even one average-sized functional protein would form. So how did life with hundreds of proteins originate just by chemistry without intelligent design?"

Source: https://creation.com/15-questions-for-evolutionists

Scientists are expected to publicly defend their scientific beliefs, and they are expected to make public their justifications for these beliefs. Numerous scientific practices and institutions are designed towards this public justification goal, be it public thesis defenses, seminar and conference presentations, the peer review journal system, the tenure track system, insistence on data transparency, or attempts to reproduce experimental results. All these designs share the basic idea that although justification is crucial for science, few claims can be justified conclusively so that no doubt remains. Instead, science encourages us to continue scrutiny of justifications and allows for any reasonable challenge to be heard and responded to. It is worth reminding ourselves that such a system did not always exist. Just 400 years ago, Giordano Bruno was burned on the stake for not bowing to the authority of the church in astronomical questions. Galileo only avoided that same fate because he did bow. It is only rather recently that the exchange of criticism and the respect for reasons and arguments has replaced such authority-driven systems, and it is still not universal.

Further reading

If you are interested in reading more on why the heliocentric model devised by Copernicus was no more justified by 16th and 17th century data than Ptolemy's system, read:

Gingerich, O. (2011). Galileo, the Impact of the Telescope, and the Birth of Modern Astronomy1. *Proceedings of the American Philosophical Society*, *155*(2), 134-144.

The third and final condition of the classical definition of knowledge is that knowledge is **true** justified belief. How to characterize truth is one of the big philosophical problems. Let us instead use a very simple conception: true statements describe (part of) how things really are. As argued before, belief and truth are two different things. Even if I believe with certainty that P is true, P might still not be true. Perhaps more interesting, *even* if I am justified in believing that P is true, P might still not be true. This is most obvious when looking at historical examples. Astronomers before the invention of the telescope believed that the earth was in the center of the universe and that all heavenly bodies moved around it. Moreover, they were justified in this belief. The observations they could make with their technology were not sufficiently precise to reveal planetary movements that would have contradicted that belief. Based on the available data at the time, the geocentric model was justified. Pre-telescope astronomers thus held false justified beliefs.

Truth: A proposition is true if it describes how things really are.

This example shows the importance of adding truth as a third condition for knowledge. We do not want to say that those astronomers *knew* the positions of the earth in the solar system, but we do not want to deny that their beliefs at the time were justified. The problem is that today, we might not be in a much better position than those pre-telescope astronomers. Some of our scientific beliefs might be justified and nevertheless false. We therefore might not have scientific knowledge. But how do we find out? We cannot access truth directly. All we can do is to justify our beliefs and hope for the best.

One possible solution is to start from the view that we might at least have access to the truth of some of our beliefs, and then argue from their truth to the truth of others. One such approach that many scientists agree on is **empiricism**, along with the assumption that we can tell whether some **observational reports** are either true or false. Such reports are more about us, the observers, than about the world. And because most humans share the same observation abilities, they might be able to determine the truth even of other people's observational reports. But this leaves the possibility of metaphysical skepticism. Even if we can tell the truth of observational reports, the skeptic might say, how do we ever learn about the world itself? Science, after all, talks about trees and cars and electrons, and forces and the regularities between them – not about observations of cars or cars, or electrons, etc. How do we ever determine the truth of statements about these things from the purported truth of observations about these things?

Empiricism: Knowledge arises from evidence gathered via sense experience.

Observational report: A statement about sense experience.

So, on a rather fundamental level, our discussion of knowledge has led us to the question what science and scientific knowledge is really about. Philosophers have different opinions on this question. **Instrumentalists** argue that creating a theory is just to rearrange and order observational reports. The theories might be deemed useful, but they are never strictly speaking true or false. **Realists** in contrast argue that theories about trees, cars, but in particular about not directly observable objects like electrons or forces, are either true or false. This realist position is appealing to many, including many scientists. This so-called *scientific realism debate* is an important part of philosophy of science, but it has few if any methodological consequences – it therefore is enough to barely sketch it here, and move on!

Instrumentalism: A scientific theory is not strictly speaking true or false, and the entities it proposes are conceptual tools, rather than something that exist.

Realism: A scientific theory is true or false, and the entities it proposes either exist or do not exist.

Part 2 - Scientific inference

2.1 Scientific inference rules

I have given you examples of methods and method choice. These examples were largely related to specific observations, such as observations in an experiment or when manipulating a model. But science is not generally interested in the particular: scientists want to generalize. They want to say something beyond the particulars that they observe and predict new phenomenon or explain phenomena that they did not directly investigate. In order to do so, they need to make **inferences**: starting from something they know and going to something else that they do not know yet. This is what we call an inference – an act or a process of reaching a conclusion from some known facts or evidence. Think of this as a simple relation between what we call a **premise** and a **conclusion**.

Inference: An act or a process of reaching a conclusion from a set of **premises**, which can express, for instance, known facts or evidence.

Premise: a statement in an argument that justifies a conclusion.

Conclusion: A statement that follows logically from premises.

There are many different kinds of inferences, and there are even more specific inferences or inference rules. When we have a lot of different inference rules that we could use, we have to make a choice, and we come back to the general theme of method choice. We have a menu of methods to choose from, and we have to justify which of these methods we are choosing. So let us go and have a look and see some of the examples of inferences. A very common one is a **generalization**. In a generalization, infers from a sample to a general claim, for example: “since these one

Generalization: Inductive inference from a sample to a general conclusion.

hundred and fifty blackbirds have been observed to be nest-making, we infer that *all* blackbirds are nest making". Another type is this one: "since the last five rockets reached a top speed of 12 km/s, the next one will also reach a top speed of 12 km/ s". We are projecting what we have observed onto future cases that we consider to be of the same type. This is another form of inference: **projection**. With this kind of inference, we are still basing it on a *finite* number of observations, but contrary to the first case, we are not drawing a general conclusion from it (to a potentially infinite number of cases) but only a conclusion about another small finite number of objects.

Projection: Inductive inference from past samples to future samples.

Further reading

For further discussion of different types of inductive inferences, of the need to justify the choice of inference rules, and it dependence on facts, read:

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*(4), 647-670.

Now, these two types are sort of somewhat similar. They all start out with observing a finite number of objects and then draw conclusions about objects of this type above and beyond the original observed ones. These are called **inductive inferences**. They are commonly contrasted with a different type of inferences, where you start with a general assumption and infer something particular: called **deductive inferences**. Such inferences are common in mathematics, but also in physics or economics.

Inductive inference: In an inductive inference, the premises support the conclusion but does not guarantee its truth.

Deductive inference: In a valid deductive inference, true premises necessitate the truth of the conclusion.

A common characteristic of this other category of inferences is that they often involve a **conditional claim**. This is not as complicated as it might sound, if you are saying to a friend: "If John bought ice cream, James bought strawberries" you are technically making a conditional claim. If it is true that John bought ice cream, you can infer that James bought strawberries (but God knows what happened if John did not buy ice cream!). More formally, a conditional claim is saying that *if* a number of assumptions are true, *then* the conclusion is true. If one then finds (or assumes) that the assumptions are true, one can infer the conclusion is true. That is a type of deductive inference called **modus ponens**. It is a standard type of inference in propositional logic. Here is another type of deductive inference, a type which will be very important for later, when we later on discuss *falsification*. We say, well, we start out with this conditional: "If the hypothesis is true, then certain consequences must be true". And then we observe that the consequence is false. And thus, we infer that the hypothesis must be false. The conditional claim could be "If it rains on this Saturday, then the music festival will be canceled". If this claim is true, and it is also the case that the music festival was not canceled, you can draw the conclusion (i.e. infer) that it did not rain on that Saturday. That is what we call a **modus tollens**.

Conditional claim: A claim involving the logical operator "if", for instance of the form "If A then B".

Modus ponens: A deductive inference of the form: *If A then B*, *A*, therefore *B*.

Modus tollens: A deductive inference of the form: *If A then B*, *not B*, therefore *not A*.

I have here given you a number of examples of types of inference rules. The first two types of inferences are inductive. They are inductive in the sense that they *amplify knowledge*. And here, **amplifications** – strengthening, boosting, or other similar terms – mean that we extend the knowledge that we have in the premise to something beyond itself. We are observing a certain number of individuals (or events, or phenomena etc.), and then we are making claims about new individuals that were not

Amplicative: Inferences that go beyond what is stated in the premises – in particular, inductive inferences are amplicative.

actually in the original premises. Because we are amplifying in this way, we are also running the risk of making an error, of making the wrong kind of conclusion. It cannot be excluded in using any of the inferences that we might be wrong: even if it is correct that all observed blackbirds are making nests, the conclusion that *all* blackbirds are nest-making might still be incorrect. The methodological question here is thus: which inductive inference rules are justified in this specific situation – and why? When are we correct in projecting from our sample to the population? This is what a lot of the discussion regarding what good inference rules are, versus what bad inference rules are, is about.

In contrast to that, deductive inferences are not amplifying. Instead, they are **explicative**, or they *explicate knowledge*. The idea is that all the information is already contained the premise. When you are making a deductive inference, you are just unpacking the information that is already contained in the premise. Think of the modus ponens example where James bought strawberries if John bought ice cream, and John did buy ice cream. Some of you might have said "well, this is very trivial". The conclusion can practically be found in the premises. This trivialness, in a way, is the strength, as we will see soon. And in any way, it is certainly the character of deductive inference. Namely, it only unpacks what is already at least implicit in the premises. And the good thing about this is that if you are explicating the premise, and you are doing that with a rule that is said to be *valid*, then you cannot go wrong. Deductive inferences are said to be **truth-preserving**: they ensure that the conclusion is always true if the premises are true. And that distinguishes them from inductive inferences, which are **fallible**: the conclusion might be false even if the premises are true.

Explicative: Inferences that do not go beyond what is stated (implicitly) in the premises – in particular, deductive inferences are explicative.

Truth preservation: The conclusion must be true if the premises are true, see **deductive inference**.

Fallible: The conclusion can be false even if premises are true.

2.2 Hume's problem and justifying inference rules

In the last section I presented two *types* of inference rules and categorized them as either inductive or deductive. For each type, there are many *particular* inference rules. Take, for example, generalization which is an inductive inference rule. One particular instance of a generalizing inference rule is rule A: "Whenever you have observed at least 150 objects of kind X to have property R, then conclude that *all* objects of that kind have property R". For example, you might ask me, why do you believe that all blackbirds are nest-making? And I might answer, I observed 150 blackbirds and, using inference rule A, I am justified in believing that all blackbirds are nest-making. This is probably not a very good rule – it is unclear why one should sample 150 objects, what the role of kind X is in this rule, and why one can draw a universal conclusion.

Although inference rules that scientists actually use look a bit more complicated, they have the same structure. Rule A might be very simplistic, but scientists do use something akin to rule B: "Let Q be evidence against hypothesis H. Whenever the probability of observing Q, given that H is true – this is the so-called p-value – is smaller than the significance level of 0.05, then reject H". Don't mind the details of this rule at this point – I will discuss this later in the statistics section.

However, there are there are many other particular inference rules. In fact, there is a continuum of such rules. An alternative to our simple rule A is to insist on having made at least 200 observations (or 250, or 300 etcetera), or an alternative that this holds for properties of kind Y. An alternative to rule B is to use significance levels of 0.01 or 0.1 or, in fact, any other number between 0 and 1. Now recall what these rules are used for. They are used to justify inferences. They are used for the purposes of justifying *why* they accept or reject certain hypotheses. But B is just *one* rule on a continuum of rules. So why would scientists use B instead of some of the infinitely many alternatives? In other words, how is the choice of an inference rule justified?

Further reading

You can read Carroll's allegorical dialogue discussing problems for deduction here:

Carrol, L (1895). *What the Tortoise Said to Achilles*. Mind. Link.

Particular inference rules thus justify conclusions *only if they themselves are justified*. Be careful to distinguish these two issues – first, that inference rules justify conclusions; and second, that the choice of a particular inference rule itself must be justified. I will now focus on the justification of inference rules, in particular inductive inference rules. The justification of deductive inference rules is also sometimes discussed, notably by Charles Dodgson (who, besides being an eminent scholar, also wrote under the pen name Lewis Carroll), but I will instead focus on the skeptical argument against induction. The most famous version of this argument was developed by the Scottish philosopher David Hume. Hume published his so-called problem of induction in 1748, but it remains as fresh and unresolved as it was then. You find the structure of his argument in the box below, but I will discuss it in detail here.

Hume's argument has five steps: (1) It starts by assuming that there are only two kinds of inferences, inductive and deductive. He uses these categories in pretty much the same way as I presented them earlier, so there is nothing surprising here. In his second step (2), Hume assumes that to justify any inductive inference rule **I**, such a rule must itself be inferred from some premise. Again, this seems pretty plausible, and it is often claimed that all justificatory arguments take on the form of inference. Hume then assumes (3) that such a justification of **I** cannot proceed deductively. Making a deductive argument would require a premise like "If an inductive inference rule has worked in the past, it will work in the future". But this premise is quite clearly not true, and Hume argues that there is no other deductive type of inference rule that would work. Or, as Hume himself puts it, just because an inference rule has yielded true conclusions in the past does not necessarily imply that it will yield true conclusions in the future. Again, this seems pretty plausible as many of us have the experience of an inference rule that fails even though it performed well in the past.

1. Every inference is either an induction or a deduction
2. To justify an inductive inference rule **I**, this rule itself has to be inferred from some premises
3. **I** cannot be inferred deductively, because there are no *necessary* connection between past and future inferences
4. Thus, **I** must be inferred *inductively*
5. When inferring **I** inductively, we must appeal to another (inductive) inference rule **J** to justify this induction. But that raises the issue of how to justify **J**, which would require appealing to another inference rule **K**, ... [*infinite regress*]

Consequently, no inductive inference rule can be justified

From assumptions one and three, Hume then concludes (4) that the justification of **I** must proceed inductively. This is just a modus ponens conclusion, so no doubt here, but it is also pretty plausible as an intermediate result. Our only chance to argue that an inductive inference rule is a good one is by showing that it has worked well in the past *sufficiently often* so that we inductively infer that we can be quite confident, though of course never sure, that it will work well in the future. But now comes the decisive blow (5). If the justification of **I** must proceed inductively, then there must be another rule **J** according to which this inductive inference is performed. And then that rule **J** also requires a justification for which we could need another inference rule **K**, and so on, leaving us with a never-ending quest for more justification. Philosophers call such a runaway chain an **infinite regress**. Ending in such an infinite regress demolishes any hope of obtaining a justification for what one wanted to justify. Hence, Hume concludes that no inductive inference rule can be justified. This is a very powerful argument which seems to demolish our hope to justify inductive inferences. But science use inductive inferences all the time – does not this then mean that scientists engage in practices that are not justified?

Infinite regress: A never ending chain of propositions being justified by other propositions which in turn are justified by other propositions and so on.

Earlier, I argued that basing one's claims on reasons is one of the hallmarks of science. In the light of Hume's argument, does this characteristic then disappear? Is science irrational because it lacks justifying reasons? Let consider this claim about irrationality and science in more detail. Scientists make inductive inferences. But by Hume's argument, any attempt to justify inductive inferences ends in an infinite regress. Thus, scientists employ unjustified methods and science is therefore irrational – quite an unsavory conclusion indeed. If we do not want to accept this, then we need to dispute either of the two premises. One possibility is to deny that scientists need to employ inductive inferences altogether. That is exactly what the 20th century philosopher Karl Popper argued. I will come back to that later.

Another possibility is to deny the *impact* of Hume's argument, rather than the argument itself. Remember, Hume's argument as such has not been disproven. The infinite regress is here to stay. However, perhaps there is an interpretation of justification which is not completely destroyed by such an infinite regress. So, what is the nature of justification? What is that which offers justification? Similar to the construction of buildings, justification, as envisioned by Hume and others, seeks an **ultimate foundation**, a set of basic claims, onto which all the other claims can be built on, or inferred from. Hume's argument denies an ultimate foundation for inductive inferences: at any level, we can always dig deeper into the ground, searching for a firmer foundation. But we never hit bedrock – that is the infinite regress. If there is no proper foundation, then one can neither erect a building, nor can one ground justification for any claim.

Foundationalism: Propositions are justified by being inferred from foundational premises which do not need additional justification, for instance necessarily true premises.

But perhaps we need no ultimate foundation to justify induction? That way, we could sidestep Hume's argument, without needing to deny its validity. But this would require a different idea of what justification is – one that is not affected by Hume's argument. Luckily, in philosophy, the foundationalist view is not the only account of the nature of justification.

An alternative to foundationalism is so-called **coherentism**. Its basic idea is that the justification for a claim increases the better it fits into a coherent system with other beliefs one holds, or claims one accepts. To illustrate this position, the Austrian philosopher Otto Neurath used the metaphor of the ship at sea. Scientists are like sailors who, out on the open sea, must reconstruct their ship to keep it afloat but who are never able to return to dry dock to lay a new foundation. Coherentism might not suffer as much from the impact of Hume's argument as foundationalism since inductive-inference rules might not function as foundations of inductive practices. Rather, they might serve as abstract descriptions of inductive practices, and they might serve as tools that can connect inductive practices with each other and with other, more-abstract, arguments and intuitions. Then both the inductive practices and their corresponding rules are justified if they cohere with each other, i.e. if they form a coherent system. This holds for claims held by one scientist holds, as well as for claims made by different scientists.

Coherentism: Propositions are justified by being compatible with a coherent set of propositions, where each proposition in the set is compatible with every other proposition in the set.

We thus have a way of dealing with Hume's problem, but only by substantially changing our view of what justification is. If we accept this new view, then formulating inductive-inference rules, as, for example, statistics does, is not a deeper or more fundamental endeavor than conducting everyday science. Rather, it is a strategy for increasing coherence between practices and rules and thereby contributing to the justification of both.

2.3 The Hypothetico-Deductive Method

We have discussed different types of inference rules, particularly deduction and induction, and we have dealt with Hume's problem of induction. One conclusion from the discussion so far is that if we want to counter Hume's argument, we better pay close attention to the actual practices of induction that scientists use. We cannot say that the actual practices are only justified if they cohere with the foundational principles. Rather, as I have argued, from the coherentist perspective it is the inferential practices themselves that we should be looking at, and consider which of them are justified in which situation. Therefore, I will discuss one such practice in particular, namely the Hypothetico-Deductive method. The Hypothetico-Deductive (HD) method is widespread family of inductive inferences in the sciences. It features both deductive and inductive steps – but because the inductive steps introduce fallibility and thus prevent truth-preservation, it must be, according to the above distinction, counted as a type of *inductive* inference rule. Here are the 5 steps of the HD method:

1. Formulate a hypothesis H
2. Deduce observable consequences $\{C_i\}$ from H.
3. Test whether $\{C_i\}$ is true or not.
4. If $\{C_i\}$ is false, infer that H is false.
5. If $\{C_i\}$ is true, increase confidence in H

To summarize, the HD methods starts (1) by proposing a novel hypothesis H that might be true. From this hypotheses, (2) one or more observable consequences $\{C_i\}$ are deduced. Through some empirical practice (for example, an experiment), some of these consequences are (3) tested: it is checked whether these consequences of the hypothesis are true or not. The result of this will either increase or decrease one's confidence in the hypothesis. If a consequence of H turns out to be false, (4) one reduces one's confidence in H, or even rejects H altogether. This latter, stronger, inference is called **falsification**. If (some of) the consequences turn out to be true, (5) one increases one's confidence in H, called **confirmation**.

Falsification: Rejecting a hypothesis as a result of an empirical test.

Confirmation: Increasing the confidence in a hypothesis as a result of an empirical test.

But let look at step 1 in more detail. What is a **hypothesis**? The Hypothetico-deductive method starts with the formulation of a hypothesis. Does that mean that we can formulate anything? No, it does not. There are requirements for what a hypothesis is, and certainly for what makes a good hypothesis. The first thing to note is that a hypothesis must be either true or false. That excludes a lot of statements that you can make, or questions that you can raise. In particular, it *excludes* research questions. There is a difference between asking: "what are the consequences of experiment X" and proposing: "the consequence of experiment X will be Y". A question cannot be either true or false. Questions do not have a truth value. A proposition, a statement, however, does have a truth value. Only those can be proposed as hypotheses.

Hypothesis: A proposition that can be true or false but is not necessarily true or false, and that preferably either has some generality or is about something not directly observable.

The second requirement is that a hypothesis should not be a **tautology**. Tautologies are claims that are *necessarily* true or necessarily false. For example, in English, the term 'bachelor' is often defined as an unmarried man. Therefore, a claim that bachelors are unmarried is necessarily true, i.e. there is no evidence that can render it false (or increase our confidence in it, because what it is to be a bachelor is to be an unmarried man by definition! Inductive inference do not affect our confidence in tautologies – they are either true by definition or inconsistent (like the claim "the sum of angles in a square is 180°"). Therefore hypotheses should not be tautological.

Tautology: A proposition which is necessarily true or false.

A third requirement is that a good hypothesis should either have some amount of generality and/or be about something that is not directly observable. Many hypotheses are not about some particular thing, but about things of that type generally speaking. If a hypothesis nevertheless is about something particular, that particular something should be something not directly observable (I will discuss the notion of direct observability in the next section). Consider the hypothesis H: "Object X smells distinctly of rotten eggs". This is a hypothesis about a particular – object X. That is already quite uncommon in science – hypotheses are typically formulated about types, not particulars. For example, "Hydrogen Sulfide smells distinctly of rotten eggs" says that *every* object that is Hydrogen Sulfide has this odor. But H only focusses on a particular object X. Furthermore, H has really only one empirical consequence – itself. It reports a directly observable property – furthermore one that can arguably only be tested through direct sense experience, namely by smelling it. To allow a hypothesis like H into the HD method makes the whole exercise trivial. Of course, hypotheses about particulars that are not directly observable are perfectly ok – as for example the claim "This object has an electrical resistance of X Ω". But hypotheses should consist of claims about directly observable particulars.

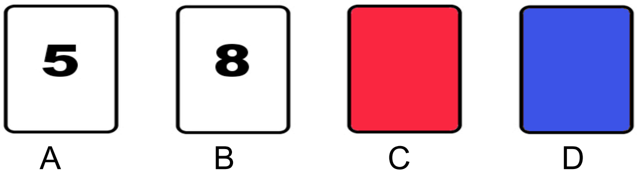
Back to the Hypothetico-Deductive method, step 2. Once we have formulated a good hypothesis, we set out to deduce some observable consequence from the hypothesis that we can test. The consequence should be about something **directly observable**, or it should be about something observable with the help of some instrument, some aid. Alternatively, the observation might involve some **operationalization**. In any case, we need to have deduce a clear, empirically observable consequence that we can test. It would be pointless to derive consequences that cannot be observed, and therefore, cannot be tested, since we test hypotheses by means of testing observable consequences in the Hypothetico-Deductive method.

Direct observation: Observations of objects and properties that are accessible through the use of human senses.

Operationalization: A way to measure something which cannot be directly observed or that cannot be observed directly with sufficient precision, by connecting this feature with something causally connected to something that can be observed directly.

The deduction of the observable consequence must, of course, be valid, and furthermore the consequence must be *relevant* to the hypothesis and the research question. Formulating a general criterion of relevance for the *confirmation* of a hypothesis is difficult, although most people find it to be reasonably easy in practice when looking at particular hypotheses. Sometimes one can derive consequences from a hypothesis, but observing them is insufficient for increasing confidence in the particular hypothesis. These are consequences that we would expect to turn out true regardless of the truth or falsity of the hypothesis at hand. For example, from the hypothesis "Mammary cancer in mice is caused by X" one can correctly deduce the observable consequence "Mice have mammary glands". This is a valid observable consequence derived from the hypothesis. But observing such mammary glands in mice does of course not confirm the hypothesis about cancer, which was what we was interested in, and is thus not relevant.

The above relevance problem does not arise for falsification, as the falsity of *any* valid consequence from a hypothesis by Modus Tollens implies the falsity of the hypothesis. If mice do not have mammary glands, they cannot have mammary cancer. But there is a related problem, namely knowing what the relevant consequences are from a hypothesis are. To **falsify** a hypothesis one must check *all* relevant hypothesis. Consider the following scenario. Four cards are presented to you like this:



Now consider this hypothesis: "If a card shows an even number on one face, then its opposite face is red." What consequences do you need to observe, i.e. which cards do you need to turn, in order to check whether this hypothesis is false or not? Try to give an answer before you continue!

The hypothesis is formulated as a conditional statement; it says, *if* the number is even, *then* it will be red. In order to show it to be false, we therefore have to figure out when the statement is false. The only circumstance under which the statement is false is when the number is even but the color is not red. Therefore, you need to turn over those cards that show an even number and those cards that are not red. If a card is red, then whatever number is on the other side is fine. But if a card is not red, then we need to make sure that it does not have an even number, because if it did, that would show that the hypothesis is false.

The card example shows that it can often be tricky to figure out what the relevant observable consequences are, that you need to deduce from your hypothesis in order to be able to obtain test result data that allows us to say something about our hypothesis.

Once we have performed our deduction, derived our observable consequence, we proceed to testing it. To test it we make a direct observation, a direct aided observation, or an indirect observation to determine whether the consequence is true or not. This will be discussed more in the lectures on measurements and on experiments, when we look into scientific empirical practices.

A test of the observable consequence can, ideally speaking, yield two outcomes. Either the observable consequence is shown to be true (i.e. the observation that we had derived, or expected, did occur), or it is shown to be false (i.e. the observation that we had derived, or expected, did not occur). In the first case, we increase our confidence in the hypothesis (I will discuss confirmation in section 2.5) In the second case, it is usually argued that we can infer the falsity of the hypothesis. This is called falsification, and I will discuss this in the next section.

Remember to distinguish **falsification** in the context of hypothesis testing, where it means to prove a hypothesis false, from **falsification** in the context of ethics, where it means the unethical behaviour of changing data to suit your opinion or target.

2.4 Falsification

Many authors claim that there is a fundamental difference between confirmation and falsification in the HD method. In the final step of the hypothetico-deductive method, falsification of the hypothesis is inferred from the falsity of even a single consequence. In cases of confirmation, in contrast, we do not strictly infer the truth of a hypothesis even in the face of a very large set of true consequences. Rather, we only infer that one might be more confident in the hypothesis if suitably many consequences of this hypothesis are found to be true. This is often called the **asymmetry** between falsification and confirmation. It is widely noted amongst practicing scientists.

Asymmetry between falsification and confirmation: No amount of confirming observations can deductively confirm the hypothesis, but one falsifying observation can deductively falsify the hypothesis.

Albert Einstein succinctly described the asymmetry as follows: “no amount of experimentation can ever prove me right. A single experiment can prove me wrong”. One must read Einstein carefully here. By proof, he meant valid deductive inference. Thus, he did not deny the possibility of confirmation. But in the case of confirmation, we make an inductive inference, not a deductive one. Falsification, in contrast, is an example of a valid deductive inference.

This asymmetry between confirmation and falsification also forms the basis of Karl Popper's **falsificationism**. Popper's method closely follows the structure of the hypothetico-deductive method. According to Popper, scientists (should) conjecture **falsifiable** hypotheses. They (should) seek to falsify these hypotheses by checking whether any of their implications conflict with empirical data. They reject any hypothesis thus falsified.

Falsificationism: The view that science should proceed only through valid falsification, and never use confirmation.

Falsifiable: A hypothesis is falsifiable if it is possible to show that it is false, even if it has not yet been shown to be false.

Note the distinction between a hypothesis being falsifiable and a hypothesis being falsified. Falsifiability is a property of a hypothesis, or more broadly, of a theory. Falsifiability in this sense, means that the hypothesis or theory has implications whose truth or falsity can be determined by means of observations. Some theories have no relevant empirical implications whatsoever. Popper claimed that theories in, for example, astrology, 20th-century Marxism, and Freudian psychoanalysis, were not falsifiable. He argued that the falsifiability of a theory **demarcates** it as scientific. A theory that is not falsifiable is not scientific according to Popper. Amongst hypotheses that are falsifiable, some might be more falsifiable than others because they have more empirically observable implications. Popper thought that a more falsifiable hypotheses is better than a less falsifiable one.

Demarcation of science: Distinguishing science from non-science by providing criteria for counting something as science.

Falsification, in contrast, is the event of observing that an implication of a hypothesis is not true, from which it is inferred that the corresponding hypothesis is not true. Falsifiable hypotheses need not be falsified, but hypotheses can only be falsified if they are falsifiable. Importantly, Popper restricted his testing method to the falsification of hypotheses. In his hypothesis testing method, confirmation plays no role at all. He was very explicit in his belief that one should never infer the truth of a hypothesis from observing its implications, not even increase one's confidence in the truth of a hypothesis. All that falsificationism allows, is to record that hypotheses, and by extension theories consisting of such hypotheses, have *so far not been falsified*. This is really quite a radical departure from scientific practice. Most scientists talk and act in ways that indicates that they have differencing degrees of confidence in various claims. Popper suggests that this is all bad method. But why does he insist on such an austere program?

Popper's reason for this strict aversion against hypothesis confirmation was the hope of avoiding Hume's problem of induction. Recall that Hume's argument seriously damaged any foundationalist justification of inductive inferences. Popper aimed to circumvent this challenge by proposing that science could proceed *without induction* altogether, and indeed, a scientist who strictly follows Popper's falsificationist program can do so. In the HD method, testable consequences C of a hypothesis H are *deductively* inferred modus ponens: If H, then C; H; hence C. The falsity of the hypothesis is inferred from the falsity of the implication in accordance with the deductive inference rule modus tollens – if not C, then not H. By arguing that science does and should proceed by deductive inferences *only*, Popper concluded that the lacking justification for inductive inferences is not so damaging for science after all.

However, one might worry that Popper's cure for Hume's problem have worse consequences than the disease it is intended for. This is because Poppers falsificationist program forces science into a very austere program that scientists do not actually follow in practice, and, which they arguably should not satisfy. Two arguments in particular have been presented against Popper's version of falsificationism.

The first worry is that Popper's deductive program does not allow scientists to separate between non-falsified hypotheses based on our level of confidence in them. At any given time, there are many non-falsified hypotheses. Some of these will contradict each other. Some will be connected to several observations made by scientists. Some will have been used in many explanations. According to Popper, scientists must nevertheless suspend their judgement about such hypotheses being more likely to be true than other, competing hypotheses: all that they can say is that some hypotheses are false, nothing more. That however, does not accord with how scientists act in practice. On the contrary, scientists not seldom make precisely such judgments, and it is not obvious that they shouldn't. Popper himself has proposed a measure of **corroboration** which distinguishes among non-falsified hypotheses according to the number of unsuccessful falsification attempts that they have been exposed to. I find this to be an interesting proposal, but it is unclear how this can be part of a purely deductive program. So, as a purely deductive program, Popper's falsificationist account is too restrictive to fit scientific practices.

Corroboration: A hypothesis is corroborated if it has withstood multiple falsification attempts.

A second worry concerns the applicability of modus tollens for the purposes of rejecting a hypothesis. In scientific practice, hypotheses rarely have immediately observable consequences. Instead, they often require the correct application of measurement instruments or experimental designs. Take, for example, a simple hypothesis such as: this liquid contains three chemical substances. What are the observable consequences of such a hypothesis? Perhaps one might use distillation, observing that three fractions condensate at different temperatures. Or one might use column chromatography, observing that compounds move through the column at different rates, allowing them to be separated into three fractions. In order to deduce such observable consequences from the hypothesis, we need to make additional assumptions about the techniques involved. Let's assume that the hypothesis is true. Even so, one can expect to observe three fractions condensating only if the distillation operates properly. Similarly, one can expect the compounds to move through the medium at different speeds only if the column is properly prepared. The need for auxiliary hypotheses such as these, changes the HD picture considerably. In the second step of the method, the deduction of observable consequences is only possible from the main hypothesis in conjunction with various **auxiliary hypotheses**. The falsity of an observable consequence therefore, via modus tollens, only implies the falsity of that conjunction. And a conjunction is false if any of the conjuncts are false. For example, if we do not observe three fractions condensating at different temperatures, all we can conclude is that the conjunction of the main hypothesis *and* the auxiliary hypotheses are false. The **conjunction** is false either if the main hypothesis is false or if the distillation apparatus does not work properly. Which of the conjuncts is false cannot be concluded only on the basis of the observations we make. More generally, we can say that we never test a single hypothesis alone, but only in conjunction with various auxiliary hypotheses. This is known as the **Duhem-Quine** thesis.

The Duhem-Quine thesis is named after the French physicist Pierre Duhem and the American philosopher Willard Van Orman Quine. The thesis considers not only assumptions about measurement instruments, but also, for example, basic assumptions such as that there exists a world external to the human mind, or that our sense organs do not systematically mislead us. We might be quite confident in some of these assumptions, but without them, we could not derive any observable implications from hypotheses. Occasionally, even such basic assumptions are shown to be false – as, for example, visual illusion tests teach us. The Duhem-Quine thesis implies that in order to falsify a hypothesis, we need to be *confident* that it is that hypothesis, and not its associated auxiliary hypotheses, that are responsible for an observation showing that the observable consequence is false. This leads us back to making judgments about confidence, however. Deductive inferences will clearly not deliver on these judgments. Consequently, the asymmetry between falsification and confirmation begins to break down. This, I think, is an important argument against Popper's suggestion for circumventing Hume's problem.

Falsification, although an important type of inference in hypothesis testing, faces the same problem of making sense of judgments of confidence and confirmation as inductive practices do. The previous arguments do not reject falsificationism altogether however. They only

Auxiliary hypothesis: A hypothesis used to test another hypothesis, but which one does not intend to test, for instance background assumptions necessary to infer the empirical conclusion.

Conjunction: Two propositions joined by the logical operator AND. The conjunction is true if and only if both propositions are true

Duhem-Quine thesis: No hypothesis can be tested without the use of auxiliary hypotheses.

point out its limitations. Falsification might still be a viable inference method among others that scientists can use.

If scientists use falsification as an inference method, they must be particularly aware of the pitfalls of **ad hoc** modification. Consider the historical example of phlogiston theory. This theory, proposed in the 17th century, postulated that a fire-like element called phlogiston is contained within combustible bodies and released during combustion. One observable consequence of this, would be that metals will lose weight if they burned, and thus release phlogiston. In conflict with this, it was observed that some metals gained mass when they burned, seemingly falsifying the phlogiston theory. But defenders of the theory replied that phlogiston might have negative weight, thus saving the theory from falsification. This reply from the phlogiston defenders is an example of ad hoc modification, as it does not serve any other purpose but to prevent the falsification of the theory. Of course, not all modifications are ad hoc. Many modifications of theories that were previously falsified often generated important new scientific insights. The criterion that Popper himself proposed is that a modification of a theory is ad hoc *if it reduces that theory's falsifiability*. It is those kinds of modification that must be avoided.

Ad hoc: The modification of a claim is *ad hoc* if (i) the claim has previously been falsified, (ii) the modification saves the claim from this falsification and (iii) it makes the claim less falsifiable – i.e. it does not allow deriving any new testable consequences.

Thus, to conclude, although Popper was right that, in the absence of ad hoc modifications, falsification is an important inference type in science, he overstated the importance of falsification in at least two ways. First, as seen from the Duhem-Quine thesis, scientists cannot avoid making confidence judgments about hypotheses so the strict asymmetry claim between falsification and confirmation is not correct. Secondly, despite its appeal in avoiding Hume's problem, falsification seen as a purely deductive program, is too austere a framework to fit the needs of scientific inferences made in practice. Instead, we need to explicitly address inductive inferences that aim to confirm hypotheses. This I do in the next section.

2.5 Confirmation

From the facts that H implies C and C is true, nothing can be deduced about the truth of H. This is not a deductively valid form of reasoning. It does not correspond to neither modus ponens nor modus tollens. Rather, any inference allowing us to say something about H, must *amplify* the information contained in the premises. Any inductive inference rule is fallible. You are likely to have made inductions that failed, be it when predicting the weather, when self-diagnosing an illness, or investing in the stock market. The interesting thing, though, is that different inductive inferences seem to have different degrees of fallibility. Some make us quite confident that the hypothesis is true, others less so.

Imagine, for example, you have a job testing products for a large sports retailer. You are currently checking a large batch of baseballs, and your hypothesis half have a cork center and the other half have a rubber core.

Before you start looking at the batch, you hear from one of your assistants that he saw workers at one of the factories sewing baseballs both with cork and rubber centers, but couldn't make out the proportion of each. If you have no other information about this issue, a plausible guess is that the batch contains 50% cork and 50% rubber cores. But then you launch a large-scale test: you select 5,000 of these balls at random, open them, observe their core, and find half of them to be made of cork and half made of rubber. Now you still believe in the 50/50 hypothesis, but with more confidence. You still can't be certain, of course: it isn't impossible that the balls in the batch that you didn't check have a different proportion of cork and rubber. But your degree of belief in the truth of the hypothesis, your *confidence*, should have gone up. Confirmation thus comes in degrees. This degree depends on several factors, for instance on the kind and quality of the evidence, as well as on the inference rule used.

Degrees of confirmation might be characterized *qualitatively*. The most recent assessment report of the International Panel on Climate Change (IPCC), for example, offers five qualifiers for expressing confidence in hypotheses ranging from "very low" to "very high". Alternatively, people have attempted to *quantify* confidence. The most prominent route here is to use probabilities to say that an observation O confirms hypothesis H. This means that the probability of that the hypothesis is true, given the observation, is higher than the probability that the hypothesis is false given the observation. For example, the probability that it has rained, given our observation that the street is wet, is higher than the probability that it has not rained, given our observation that the street is wet. This is, however, beset with a number of problems. First, not everybody agrees that it makes sense to assign probabilities to hypotheses because they interpret the concept of probability differently. One prominent position, **frequentism**, defines probabilities as the frequencies of repeatable observable events. Because hypotheses are not events, neither observable nor repeatable, frequentists argue that one cannot meaningfully assign probabilities to them. They are the wrong kind of thing to be said to have probabilities, the argument goes. Another problem is that probabilities are already used to express a property different from confidence. I might say, for example, that the probability of drawing a rubber core ball from the above batch is one half. You might then ask how confident I am in *that* claim, "The probability of drawing a rubber core ball is one half", and that is a separate question from the first one. Furthermore, in the above example, the probability of drawing a rubber core ball stayed the same, but the degree of confidence that this was the correct probability changed. Of course, one might answer the second question with a probability also, and have two distinct types of probability. But then one must keep these two numbers separate, which people often fail to do.

Frequentism: Probabilities are frequencies of repeatable observable events.

So specifying what we mean by degrees of confidence isn't trivial. But even if we agreed on a framework, it is still unclear what justifies confirming inductive inferences. To understand this more clearly, let's go back to the HD method. There, observing C to be true increases our degree of confidence that H is true. But why? What makes it so that observing C increases our belief in the truth of H?

The first answer is because H is *compatible* with C. When checking whether C was true or not, we could have made observations that contradicted H, thus leading to it is rejection. But we did not. Therefore, we are now more confident in H. This compatibility argument offers only very weak justification of confirmation. First, one can deduce indefinitely many implications from a hypothesis, even though the truth of many of them seems highly irrelevant. For example, one can always deduce **tautologies** from anything: from the hypothesis “it has rained today” we can infer that “it has rained or it has not rained today” – and tautologies are always true. But surely, deducing a tautology from any H does not lead to an increase in our confidence in H. For another example, recall that from the hypothesis “Mammary cancer in mice is caused by X” one can deduce the observable consequence “Mice have mammary glands”. But observing such mammary glands in mice does not confirm the hypothesis about cancer.

Tautology: A proposition or an inference which is necessarily true.

An additional concern with the compatibility argument is that very many hypotheses are compatible with any given observation. This is called the problem of **under-determination**. Take, for example, the problem of curve-fitting. In curve-fitting, we connect a finite number of observation in state space through a continuous line that is supposed to represent not only those observational points, but also the infinitely many points in between. We can think of the curve as a hypothesis that inductively generalizes beyond the observations. Now, does the fit of the curve to the observational points confirm this hypothesis? Well, maybe, but we need to admit that there might be another curve, distinct from the first, that is also compatible with the observations. And so are indefinitely many other ones. Each of them are substantially different hypotheses compatible with the observations but making different claims about the world otherwise.

Under-determination: An inference is underdetermined if there are multiple conclusion that would be equally supported by the premises.

A possible solution to this problem is to select the *simplest* hypothesis that is compatible with the observations. For the curve fitting case, this might work, as one might be able to rank these different curve hypotheses according to their degree of polynomial. But in many other under-determination cases simplicity ranking is not available. Consider, for example, the hypothesis that I have stomach cancer. Theory indicates that one relevant consequence of stomach cancer is heartburn. So, when I experience heartburn, should I substantially increase my confidence that I have stomach cancer? The answer is no. Very many hypotheses are compatible with heartburn. Thus, heartburn alone is not a strong sign of stomach cancer. A proposed solution for this problem is to introduce some criterion of *relevance* for observable consequences. Only the truth of relevant implications in the stomach cancer case, such as various effects of the cancer, would confirm the hypothesis. Specifying such relevant conditions is tricky, however, and requires a lot of background knowledge.

These examples shows that the compatibility of hypothesis and observation is not a sufficient basis for confirmation. We need something stronger. One proposal is that C confirms H because C would have been *very unlikely if H had been false*. That is, the possibility to observe C would have been remote if H were false and some other hypothesis true. This dependence of C on H however does not rule out that C is compatible with other hypotheses. Rather, it acts as an additional requirement on whether a consequence can be confirming evidence of a hypothesis: only if there is such a relationship between C and H, does the observation of C confirm H.

Severe test: A hypothesis test is a severe test if the possibility to observe a consequence would be low if the hypothesis was true.

The philosopher Deborah Mayo calls hypothesis tests that satisfies this condition **severe tests**. Severe testing solves our heartburn problem. Although stomach cancer is compatible with heartburn, the chance of suffering from heartburn if one has stomach cancer is not higher than the chance of suffering from heartburn if one has reflux disease or some other gastrointestinal disorder. Testing the cancer hypothesis by observing heartburn is thus not a severe test and observing heartburn does not substantially confirm the cancer hypothesis. Severe testing thus considerably improves the quality of inductive inferences.

Severe testing, however, requires a lot of information, in particular about alternative hypotheses and the probability of making the observation that one has made if those alternatives where true or false, respectively. Furthermore, this proposal does not solve all problems associated with confirmation either. What, for example, if one first made an observation and then constructed one's hypothesis in such a way that it would imply this observation? Such a hypothesis might meet the severe test condition, but would, arguably, not confirm the hypothesis. Consider curve-fitting again. It seems illegitimate to construct the curve in such a way that it fits all the available information, then claim the curve as one's hypothesis and then use *the same data points* as confirming evidence for this hypothesis. Instead, only observations that have not already been built into the construction of the curve should be used in confirmation.

Finally, consider this case where severity seems insufficient to justify strong confirmation. The police used new breathalyzers, which falsely display drunkenness in 5% of the cases in which the driver is sober. However, the breathalyzers never fail to detect a truly drunk person. 0.1% of the population is driving drunk. Suppose a police officer stops a driver at random and forces the driver to take a breathalyzer test. The test indicates that the driver is drunk. How *confident* should the officer be that the driver is *actually* drunk? Try finding the solution yourself before continuing!

Further reading

You can read more about Mayo's account of statistical inferences in her 2018 book:

Mayo, Deborah G. (2018). Statistical inference as severe testing: how to get beyond the statistics wars. Cambridge, United Kingdom: Cambridge University Press

The correct answer is the probability that the stopped driver is actually drunk is about 2%. This might seem surprising. After all, the breathalyzers which falsely display drunkenness in only 5% of the cases in which the driver is sober – that is, C would have been pretty unlikely if H had been false, and the breathalyzer still counts as a fairly severe test.

Yet, because the base rate of drunken driving is so low in the population, the relatively small error of a wrong indication becomes very influential. For example, consider a sample of 1,000 drivers. According to the information given above, only 0.1% - one out of these 1,000 - is driving drunk. And the breathalyzer correctly identifies him or her. But the breathalyzer also falsely identifies 50 out of these 1,000 as drunk although they are sober. Altogether, the test identifies 51 people as drunk, of which only one is actually drunk. One out of 51 equals roughly 2%. Thus, the probability of someone identified by a breathalyzer to be actually drunk is about 2%.

This is a case that meets the condition of a severe test, but it does *not* confirm the hypothesis *strongly*. It does confirm it a little but - just not as much as you probably thought. Mistaken judgments of confirmation based on these kinds of scenarios are called **base-rate fallacies**. Because the initial confidence in the hypothesis was so low, even such a severe test could not increase the confidence very much. In order to deal with such cases, we need a Bayesian approach to hypothesis testing. I will discuss such approaches later in the statistics section.

Base-rate fallacy: The initial confidence in a hypothesis not taken into account when performing a statistical hypothesis test.

To conclude, confirmation of a hypothesis by an observation comes in degrees. To simply note that the hypothesis is compatible with the observation is not enough however. Rather, we need also to ensure that the hypothesis made the observation very likely, and that the observation would not be very likely if the hypothesis had been false. Such a severe test condition improves the quality of inductive inferences substantially, but even it faces a number of problems.