ÜLO MAIVÄLI

METHODOLOGY

Introduction

Scientific method: the short version

Science has a single goal: to produce useful knowledge. "Useful" does not necessarily mean "monetized", as there is more to human experience than business. It means any kind of knowledge that is interesting for any reason. Scientific product is used in technology, in medicine, in business, in mass murder, in politics, in music, and, frankly, in all human affairs. Without science the culture, and even thinking, would be different.

Scientists create and test theories on data. This means that science is a belief system, where existing beliefs are continually updated on new evidence. So science is evidence-based, but its more than that. It's like a spider's web that is continuously re-spun in response to a changing environment. Every piece of new data causes ripples across this world wide web of beliefs, and thus the web is never static, never finished. The main goal of methodology, the science of method, is to ensure optimal updating of beliefs.

Rationality: the short version

We live in an uncertain world. While the world itself may or may not behave with true randomness, our attempts to collect data on it are hampered by measuring noise, random variation, and bias. Thus any attempt on making sense of the world must come to grips with these sources of uncertainty.

What makes a rational entity? Perhaps it is its goal of making the most of limited, noisy and biased information to formulate degrees of belief; and as a step further, to convert these beliefs into optimal decisions. We can make sense of the concept of rational beliefs by thinking of a spectrum of uncertainties, beginning with radical uncertainty, where we have no relevant information to limit the possibilities and on the other end of the continuum, completely understood uncertainty exemplified by games of chance. If we think of uncertainty in terms of the risk of something happening, then we immediately get

that the radical uncertainty side of the continuum is deeply unpleasant, because the risks it describes are essentially unknowable – they can be very large, very small, or anything in between. In contrast, the games of chance are much more satisfying, as they are perfectly understood in the sense that we can from simple models calculate the exact long-run relative frequencies of any conceivable outcome (and these frequencies can be very close to truth in a real casino).

Somewhere inside of this continuum lies every scientific question that we might want to answer on empirical data. What scientists often do is to try to identify interesting questions that are as far away as possible from the radical uncertainty side of things, and then to apply simple stochastic models that are close to those employed for the games of chance, and hope that the modeling is somehow relevant in the real world. Every statistical model that uses binomial distribution to predict binary outcomes (dead/alive, etc.) is an example of this approach to rationality.

Example 1: Covid-19 in the spring of 2020. The first epidemiological models of spread of COVID-19 in Europe (and in Estonia) used imperfect and scanty data and assumed that the Chinese containment measures would be replicated in Europe; and the models themselves were created on flu epidemics. These models failed on two levels: (i) their predictions were wrong, and (ii) their uncertainty intervals were too narrow, disregarding actual uncertainty. Most models do not "know" anything about the inherent uncertainty present in their own structures – they have no self-awareness.[1] Thus models inappropriate for the radical uncertainty side of the spectrum were (mis)used on a problem that was rather close to radical uncertainty.

[1] the models had no notion that covid-19 is not flu, because the modelers did not know how to tell this to their models.

Example 2: Covid-19 in the spring of 2021. With a years worth of good European data and appropriately tweaked model structures, everything works much better. Now the problem is quite far from radical uncertainty and modeling has become very useful indeed.

What are the assumptions behind rational thought?
Very general assumptions:

1) The constancy of Nature. The mechanisms of Nature do not change over time, neither does nature try to deceive us.

2) The consistency of explanation: any true description of Nature is self-consistent, meaning that a claim cannot be true and false at the same time, and that two mutually inconsistent claims cannot be both true (but they can be both false).

3) The consistency of inference: we have identified the full set of mutually exclusive and exhaustive states of the world (or hypotheses

describing these states of the world, a.k.a. sample space, a.k.a. hypothesis space). The rules of inference guarantee that on correct application the conclusions will always be the same (regardless on which particular way they were reached).

4) Methodological reductionism: the great diversity of observed phenomena is generated by a relatively small number of underlying causal mechanisms. That is, we can get meaningful insight into nature by studying these elementary causal mechanisms.

These assumptions can never be proven to be satisfied.
General modeling assumptions:

1) Stationarity: the data generating process that is being modeled does not change during data collection.

2) Exchangeability - the order on which the data was collected has no influence on the results.

3) The theories and models that we use come close enough to reality to matter.

This is how it works

Lets say that we want to find out, whether Madame Marie has extrasensory perception (ESP). We present her with playing cards and let her guess, what is on each card. The data is in the form of a list of TRUE-FALSE and so on. Now we compare our data with simulated datasets that assume no ESP and see that our data are significantly different from the simulated data. Lets say that only 1 in 10 000 simulated datsets gives a fraction of TRUE guesses that is equal or greater than Madame Marie's ($p = 10^{-4}$). We can interpret this result as showing that the data are pretty inconsistent with the hypothesis that Marie picks cards completely randomly. But can we interpret this as weighty evidence for ESP? And, if so, does this mean that we should now put a lot of confidence on this hypothesis? The answer is "it depends" on both counts. It depends on:

(1) Our prior beliefs on whether Marie is clairvoyant. If we think this very unlikely (our prior probability is a lot less than 10^{-4}), then the evidence obtained is not enough to change our minds. To arrive at a prior probability, we should ask both about how likely is ESP as a general physical/biological mechanism (is it excluded from our current scientific worldview the way as a *perpetum mobile* is?), and then, whether we know something about the individual that would make her more or less likely than average to have such abilities. If we decide to treat ESP analogously with a perpetum

mobile, then we need to put a probability on how likely it is that our current scientific wordlview is completely wrong and, specifically, wrong in a way that would allow for the existence of ESP. This is obviously complicated, and there is no guarantee that two reasonable people would arrive at the same prior probabilities. This does not mean that we could ignore our prior beliefs, because we find it hard to precisely quantify them. It does mean that we introduce a new source of uncertainty, which means that in practical analysis we might want to try out a range of priors to see how robust our inference is to the choice of prior.[2]

[2] For any uncertainty in our data, models and priors, we must strive to carry it forward all the way into our conclusions, rather than to pretend that it does not exist.

(2) the inference from data also depends on how we choose to divide the *hypothesis space* into testable hypotheses. The hypothesis space (also known as the *sample space*) is defined as the set of all mutually exclusive and exhaustive hypotheses, exactly one of which must be true. So if we define our hypothesis space as consisting of 2 hypotheses, on of which (H1) is that "Marie has ESP", then the other one (H2) must be that "Marie has no ESP". If we start by giving H1 the prior probability of 1/100 000 then that means that the H2 must have a prior probability of 1 - 1/100 000. The probabilities always sum to unity accross the hypothesis space. This means that if after taking account of the data the P(H1) is determined as 1/10, or 0.1, then the P(H2) must be 9/10the or 0.9. But what happens when we decide to slice the cake differenty, into 3 hypotheses: H1 (ESP), H2 (no ESP) and H3 (bias), the last one encompassing all mechanisms that would shift the results away from completely random card picking. Some of these will be conscious (fraud) and some unconcious (some cards might have distinctive marks on their backs, or the experimenter might give involuntary signals to Marie), so the H3 is itself a hodgepodge of very different hypotheses.[3] The the priors would change: P(H1) = 1/100 000, P(H3) = 1/10, say, and P(H2) = 1 - 0.1 - 1/100 000. Now it becoes crucial how well data corresponds to not only H2 (very poorly), but also to H3 (quite well). And is data is equally compatible with H1 and H3, and the prior probability of H3, P(H3) » P(H1), then the same data turn out quite useless in promoting H1 for us. And again, the posterior probabilities will sum to unity accross the 3 hypotheses.

[3] Into how many slices to divide the hypothesis space depends on both the scientific importance of each hypothesis and on how well you can determine the probability of data under that hypothesis. So its a compromise between your scientific needs and technical ability.

For the philosophically minded, re-slicing the hypothesis space would be akin to a Kuhnian paradigm shift where it becomes possible, and indeed necessary, to re-evaluate what the data says.

(3) From the previous point we can surmise that the inference also depends on how good was Marie in guessing cards? Was she only a little bit better than chance, or maybe near-perfect? If the data

are too good to be true, then they are less likely to support the hypothesis of ESP over that of bias. So not only the incomatibility of data with the null hypothesis of no effect, but also with other relevant hypotheses counts, and there sure can be too good data that reduces the weight of evidence of data in relation to our favorite hypthesis.

The previous 3 points are meant to be logical necessities, which have the following consequences: (1) Evidence is always relative - it is a quantitative measure of how much more data supports one hypothesis over another hypothesis.[4] There is no sense in collecting data if you only have a single defined hypothesis to check it against (then you end up with a p value, which is by itself inferentially useless). (2) the weight of evidence depends on data and data contributes to evidence alone, but to form rational beliefs on the truth of hypotheses you also need prior beliefs, which are independent of data in hand. (3) How you divide the hypotheses space makes a big difference, and if a relevant hypothesis is excluded from consideration, that could easily invalidate your inferences. Sadly, in science we can never be sure that we have even an inkling of what the single true hypothesis looks like. The logic of rational thought is not a truth machine – it only shows how to optimally combine information that you have. Thus, given poor-quality input in terms of data, models and/or hypotheses, it can in no way guarantee a march towards truth for science.

[4] Technically, evidence is the likelihood ratio $LR = \frac{P(data|H1)}{P(data|H2)}$, which can take values from 0 to infinity. LR=1 means that data is equally likely if H1 is true and if H2 is true and therefore has no evidentiary weight on H1 vis-a-vis H2. LR<1 means that evidence supports H2 over H1 and LR>1 means that evidence supports H1 over H2. Of course, it is always possible that there exists H3, which is much more supported by data than H1 and H2.

Decision theroy: the short version

Narrow rationality can be separated into two interlocking parts:

1. epistemic rationality maps our beliefs to reality.
2. instrumental rationality uses these beliefs to chart an optimal course of action. Technically speaking, it maximizes expected utility.

Maximizing expected utility entails: (1) mapping of all possible actions that are under considerations; (2) evaluation of the consequences of each possible action in each possible state of the world.[5]

Obviously, some people are more rational than others. Psychologists believe that a rationally thinking person is characterized by a specific type of thinking: he collects information before making up his mind, he seeks various points of view, he thinks extensively about a problem, he calibrates the degree of strength of his opinion to the degree of evidence available, he thinks about future consequences

[5] To calculate expected utility we take the product of the utility of each outcome with the probability of that outcome occurring, and then sum those products.

before taking action, he explicitly weighs pluses and minuses of situations before making a decision, and he seeks nuance and avoids absolutism.[6]

[6] see Stanovich, K. E., West, R. F., and Toplak, M. E. (2016). The Rationality Quotient: Toward a Test of Rational Thinking (pp. 1–479). The MIT Press.

Causality: the short version

Causal hypothesis means that we believe that some specified cause produces or helps to produce some effects. Causal effects may be produced by the cause either always (colliding billiard balls), or with some probability/frequency (smoking causing cancer). Also, causes can be direct or indirect and a given effect may have several causes which must either operate all together or in some combination to produce this effect. So it is often not a trivial task to squeeze the world into a causal mold.

Causal thinking seems to be how our brains impose order onto universe, and as such it is important to us. At the very least, it is the major tool of scientists and conspiracy theorists. However, the concept of causality is essentially metaphysical: we cannot be sure whether or not it corresponds to anything real in nature (if you are in doubt about this, ask a quantum physicist). Nevertheless, by mutual consent, we do have an indirect measure of causation. This is the counterfactual principle according to which removing a cause should also remove (or at least modify) the effect(s). "Counterfactual" means that a cause cannot be present and absent at the same time, unless we live in a multiverse. Thus, to test a causal hypothesis, we must employ a model wherein we can generate *in silico* both states of the world, and by comparing the predicted value of the "effect" at different values of the "cause" we can modify our beliefs on the hypothesis.

Example 1: By doing a randomized placebo-controlled drug study and employing a simple model *outcome = intercept + slope x drug=1/0 + error model* we can compare the modeled outcome for each group and if the drug causally influences the outcome, these model predictions will be substantially different. Question: why are we randomizing subjects into the two study groups?

Example 2: Was the extreme storm that we experienced caused by global warming? Now we build a really complicated model and run it many times under a global warming scenario and in an alternative form that lacks global warming. If the global warming scenario produces storms like ours with twice the frequency of the sans warming scenario, then we can conclude that the particular storm was caused by global warming with the probability of 2/3.

Modeling: the short version

Methodology does not deal directly with genesis of theories, which is art more than it is science. This means that a methodologist does not try to build a full-fledged robot scientist. Rather he tries to build a simple robot that thinks simple thoughts very quickly, all this to help the flesh-and-bone researcher with the easy part while leaving the hard part – formulating novel theories – to the genius of the scientist. Science is hard, and cannot be automated. Another way of saying this is that the goal of methodology is to ensure the rationality of scientific argument. Rationality in this narrow, normative sense prescribes an optimal way of working with information. Anything that deviates from the prescribed optimality, is thus by definition "irrational", which does not mean that it is "bad", or even "worse than rational under the circumstances". We must accept that scientific thinking in the real world is by necessity a mixture of the rational and the irrational. Especially we must note that rationality in this narrow normative sense is blind to ethics.

Behind every model is our understanding of the **data generating mechanism**, which is the collection of causal elementary mechanisms that generated the data that goes into model. It comprises the laws of nature, the experimental design (controls and such), the actual data collection (whether it is collected in a single centre or not, and so on), and every other conceivable mechanism that can affect how the data come out. The major assumption of modeling is that our system is **stationary**, that is the data generating mechanism does not change during the collection of data. This does not mean that the data itself cannot systematically change over time, or between centres, only that the underlying mechanisms do not change. For example, if we want to model the prize of gas, the model loses stationarity as soon as a new mechanism that affects this prize emerges (electric cars, bioethanol). That would require building of new models that take this new mechanism into account.

Our models should strive to reflect some aspect(s) of the data generating mechanism, and by building several of them and comparing the results, there is a chance that we will learn something useful about the whole complex mechanism, and by implication the world it represents.

Causal inference and Experiment

Ancient theories of causality

Aristotle (384-322) offers 4 types of answers to the question "why?"

1) the **material cause** – this what things are made of.

2) the **formal cause** may be expressed in a definition. It is a form and a pattern of a thing. the proportion of the length of strings is the reason for different notes played on a lyre being an octave away.

3) the **efficient cause** – this the orgin of a change or state of rest. A person reaching a decision, a father who begets a child, a sculptor carving a statue, a doctor curing a patient.

4) the **final cause** is the end goal why something is done (doctor works to make the patient healthy, a lung works to oxygenate the body, etc.). For living things the formal and final causes coincide (the realization of a mature organism is the goal towards which the an organism strives).

Aritotles causes and effects are usually not single events, like causation is commonly seen in modern philosophy, but more complex entities, like human beings or proportions or the skill of the sculptor. Effects can also be states, actions, substances, events.

Stoics saw causation differently.[7] They disregard material, formal and final causes, but espouse several different forms of efficient causes. They also offer a **universal law of causation**: *A* causes that *B* is *F*, where *A* and *B* are bodies, and *F* is an abstract non-bodily entity, a *lekton*. For example, A - scalpel, B- flesh, F - predicate "being cut". *A* is the active element, while *B* is the passive element (also called *matter*).

[7] Stoicism is a school of Hellenistic philosophy founded by Zeno of Citium in Athens in the early 3rd century BC. It lasted for 500 years.

- Joint causes (*sunaita*) – two oxen in fromt of a cart, if neither alone can move it
- Auxiliary causes (*sumerga*) – if either ox can move the cart alone.

Causes form a network, not always a chain.

- sustaining causes (*aitiai synektikai*) hold things together in time. Your body is held togehter by an active fluid, *pneuma* (breath), which causes the cohesion of the universe. You are kept alive by your soul, which is your sustaining cause. As soon as your sustaining cause ceases, you will die and start to decompose.

- other antecedent (*prokatrktikai*) causes retain their effects after ceasing to operate. Like a house that remains after the builders leave. An antecedent cause leaves something material to the matter that is itself a sustaining cause – this makes it possible for its effects to linger. When a patient catches a cold, the coldness of air is the antecedent cause and the subsequent fever is the sustaining cause of symptoms.

The stoics accepted a strong causal determinism: (i) everything is caused by something, and (ii) every effect is fully determined by its cause(s). So, nothing happens without a cause and every cause is a necessitating cause (the effect cannot not happen, when there is cause). This unified series of causes and effects is the doctrine of fate. This, of course, brings on the question about free will. On way to get to free will is to posit with Epicurus (341-270) that the world is made of tiny atoms, which in addition to causal interactions have some amount of random unpredictable swerves. Or, alternatively, maybe the mind is somehow has freedom of tought, although it is influenced by external causes (Chrysippus; b. 279 B.C.), or maybe there actually are no external antecedent causes for our will (Carneades; 214-129).

A modern framework for causal analysis, the counterfactual model or potential outcomes model was proposed by Donald Rubin ca. 1980. According to this each individual has a potential outcome under each treatment state, even though each individual can be observed in only one treatment state at any point in time. Because it is impossible to observe both outcomes at the same time for any individual, causal effects cannot be observed at the individual level. Thus, we need to estimate average causal effects for groups of individuals defined by specific characteristics. To do so effectively, the process by which individuals of different types are exposed to the cause of interest must be modeled. If the modelling assumptions are defendable, and a suitable method for constructing an average contrast from the data is chosen, then an average difference can be given a causal interpretation. Experiment is an investigation where the system under study is under the control of the investigator. The individuals or material investigated, the nature of the treatments or

manipulations under study and the measurement procedures are all selected by the investigator, whereas in an observational study some of these features, and in particular the allocation of individuals to treatment groups, are outside the investigator's control. This brings us to the **fundamental problem of causation**: as comparison with control is neccessary, how can we both do and not do the treatment? We could do treatment first and then the control (or vice versa). But this assumes temporal stability, causal transience (system response is not influenced by its history), and unit homogeneity. Or then, there is always the statistical solution, where we assume that there are many different units of observation (a population), where units are not identical, but as a whole still represent the population. We also assume that the variable that we manipulate (the cause) is independent of the variables that we measure – only then can we measure experimental values in some units and control values in others to calculate the average causal effect.

This is a lot of assumptions. Remeber that: experimental results do not speak for themselves. All experimental systems are simplifications of reality: they involve physical models and often statistical models as well. All experimental results are relative to control. This means that experimental results must be put into wider context and they need interpretation.

The randomization to groups to balance the groups means that increasing the sample size reduces the probability of it not being representative of the population. The goal is that any confounding factor besides the experimental manipulation, which could influence the result of your experiment, would be present in equal "concentration" in the study group and control group. If this is the case, their influences on experimental results will cancel out. For this to work, we need (1) not too many confounding factors to partition and (2) not too small a sample. If the sample size is comparable with the number of confounding factors, some imbalance will likely be present and randomization cannot remove bias. Sadly, usually we have no idea about how many nuisance variables there might be in our experiment. As a rule of thumb, for RCTs with a sample size of at least 200, randomization should be "an insurance in the long run against substantial accidental bias between treatment groups". The ´long run´ means that even with large samples randomization will not guarantee elimination of bias from a particular sample, it will only make it more likely.[8]

Biological experiments are usually not randomized and the independence of repeat experiments is achieved by varying as many components of the experimental system (buffer batches and enzyme preps) as possible and by trying to keep the compositions and treat-

[8] In most biochemical experimentation it is unclear what "random sample" means because the underlying population is imaginary.

ments of the experimental and control groups as similar as possible. This leads naturally to paired designs where experimental and control groups are treated together as much as possible, and in parallel. The idea behind a paired design is that substantial fraction of random variation resides within the pairs, and we can take this into account in our analysis model structures.

So how do we know that by treatment we treated the variable that we wanted treated, rather than something else? We simply need a simple enough experimental system, which means that we cannot study the world directly – need a simplified system that we can consistently disrupt with treatment. As every simplification constitutes an extra hypothesis, this makes it correspondingly harder to generalize back to the world.

An experimental system consists of the biological object of measurement, reagents, machines, the experimenters, and the act of measurement. Experimental system has long-run quality properties: it has sensitivity and specificity. It has a measure of complexity and a measure of relevance to the real world. There is trade-off, by which a simple system is easy to interpret, but at the expense of external validity (genralizability). For instance, simplified *in vitro* systems presuppose that we understand *in vivo* conditions, less there be serious problems with external validity. And the in vivo sytems, which tend to have better external validity, are often harder to interpret (theres a bigger chance of confounding).

Early mordern experiment

Experiment has meant different things to different people. Sometimes it is used to test a theory, sometimes it is used to help formulate theories, and sometimes it is done just to get new experiences, or just for fun. It is also done to figure out why something does not work (testing new enzymes to figure out, why my PCR reaction doesnt work), or how to improve some process (component titration to improve some reaction). There is not a single reason for experimentation.

For different takes on experiments compare two great chemists, Davy and Leibig:

> The foundations of chemical philosophy, are observation, experiment, and analogy. By observation, facts are distinctly and minutely impressed on the mind. By analogy, similar facts are connected. By experiment, new facts are discovered; and, in the progression of knowledge, observation, guided by analogy, leads to experiment, and analogy confirmed by experiment, becomes scientific truth. To give an instance. - Whoever will consider with attention the slender green vegetable filaments (Conferva rivularis) which in the summer exist in almost all streams, lakes, or pools, under the different circumstances of shade

and sunshine, will discover globules of air upon the filaments that are shaded. He will find that the effect is owing to the presence of light. This is an observation; but it gives no information respecting the nature of the air. Let a wine glass filled with water be inverted over the Conferva, the air will collect in the upper part of the glass, and when the glass is filled with air, it may be closed by the hand, placed in its usual position, and an inflamed taper introduced into it; the taper will burn with more brilliancy than in the atmosphere. This is an experiment. If the phenomena are reasoned upon, and the question is put, whether all vegetables ofthis kind, in fresh or in salt water, do not produce such air under like circumstances, the enquirer is guided by analogy: and when this is determined to be the case by new trials, a general scientific truth is established - That all Confervae in the sunshine produce a species of air that supports flame in a superior degree; which has been shown to be the case by various minute investigations. Humprey Davy (1778-1823)

and

In all investigations Bacon attaches a great deal of value to experiments. But he understands their meaning not at all. He thinks they are a sort of mechanism which once put in motion will bring about a result of their own. But in science all investigation is deductive or a priori. Experiment is only an aid to thought, like a calculation: the thought must always and necessarily precede it if it is to have any meaning. An empirical mode of research, in the usual sense of the term, does not exist. An experiment not preceded by theory, i.e. by an idea, bears the same relation to scientific research as a child's rattle does to music. Justus Liebig (1803-73)

Or the philosopher Karl Popper (1934)

The theoretician puts certain definite questions to the experimenter, and the latter by his experiments tries to elicit a decisive answer to these questions, and to no others. All other questions he tries hard to exclude. . . . It is a mistake to suppose that the experimenter [. . . aims] 'to lighten the task of the theoretician', or . . . to furnish the theoretician with a basis for inductive generalizations. On the contrary the theoretician must long before have done his work, or at least the most important part of his work: he must have formulated his questions as sharply as possible. Thus it is he who shows the experimenter the way. But even the experimenter is not in the main engaged in making exact observations; his work is largely of a theoretical kind. Theory dominates the experimental work from its initial planning up to the finishing touches in the laboratory

Or Joan Fisher Box (the daughter of R.A. Fisher)

The whole art and practice of scientific experimentation is comprised in the skillful interrogation of Nature. Observation has provided the scientist with a picture of Nature in some aspect, which has all the imperfections of a voluntary statement. He wishes to check his interpretation of this statement by asking specific questions aimed at

> establishing causal relationships. His questions, in the form of experimental operations, are necessarily particular, and he must rely on the consistency of Nature in making general deductions from her response in a particular instance or in predicting the outcome to be anticipated from similar operations on other occasions. His aim is to draw valid conclusions of determinate precision and generality from the evidence he elicits. Far from behaving consistently, however, Nature appears vacillating, coy, and ambiguous in her answers. She responds to the form of the question as it is set out in the field and not necessarily to the question in the experimenter's mind; she does not interpret for him; she gives no gratuitous information; and she is a stickler for accuracy. In consequence, the experimenter who wants to compare two manurial treatments wastes his labor if, dividing his field into two equal parts, he dresses each half with one of his manures, grows a crop, and compares the yields from the two halves. The form of his question was: what is the difference between the yield of plot A under the first treatment and that of plot B under the second? He has not asked whether plot A would yield the same as plot B under uniform treatment, and he cannot distinguish plot effects from treatment effects, for Nature has recorded, as requested, not only the contribution of the manurial differences to the plot yields but also the contributions of differences in soil fertility, texture, drainage, aspect, microflora, and innumerable other variables.

Fisher's take on causality and experiment seems to be in direct succession with that of Francis Bacon, whose Novum Organum (1620) argued rather masculinely for "taking nature by the forelock" and that real learning occurred not when nature was "free and at large," but when nature was "under constraint and vexed; that is to say, when by art and the hand of man she is forced out of her natural state, and squeezed and moulded". For Bacon experiment was a "lawful marriage between the empirical and rational faculty".

Or Robert Hooke (Memorandum on the Royal Society, 1663)

> The business of the Royal Society is to improve the knowledge of naturall things, and all useful Arts, Manufactures, Mechanick practices, Engynes and Inventions by Experiments (not meddling with Divinity, Metaphysics, Moralls, Politicks, Grammar, Rhetorick, or Logick). . . . to attempt the recovering of such allowable arts and inventions as are lost. . . . to examine all systems, theories, principles, hypotheses, elements, histories, and experiments of things naturall, mathematicall, and mechanicall, invented, recorded, or practised, by any considerable author ancient or modern. In order to the compiling of a complete system of solid philosophy for explicating all phenomena produced by nature or art, and recording a rationall account of the causes of things. In the mean time this Society will not own any hypothesis, system, or doctrine of the principles of naturall philosophy, . . . nor the explication of any phenomena . . . not explicable by heat, cold, weight, figure, and the like, as effects produced thereby; nor dogmatically define nor fix axioms of scientificall things, but will question and canvass

> all opinions adopting nor adhering to none, till by mature debate and clear arguments, chiefly such as are deduced from legitimate experiments, the truth of such experiments be demonstrated invincibly.

There are two types of experiments:

1. In an uncontrolled experiment we create some artificial conditions, which we would normally be unable to observe, and then see, what happens. Robert Boyle's *New Experiments Physico-Mechanical, Touching the Spring of the Air and Its Effects* (1660) described such experiments in artificial vaccuum, which were done to understand the nature of air, its effects on animals and respiration, and the relationship between the pressure ("spring of the air") and volume of gas in a closed system. Essentially, such an experiment creates new experiences, and its detractors not unreasonably asked, why bother studying these new things that are generated, at great cost, by poorly understood unnatural means, when the world is filled with wholly natural phenomena, waiting for explanation.

Figure 1: Joseph Wright's An Experiment on a Bird in the Air Pump (1768) housed at the National Gallery, London.

For example, when people but a bird into a vaccuum pump, it died; and adding a burning candle quickens the process. But building a really big vaccuum pump and putting a child inside did not result in any harm to the child. So, without controls, what can we conclude from these experiments?

2. In a controlled experiment we compare (at least) two sets of conditions.

In 1644, Evangelista Torricelli performed the famous experiment of "argento vivo" (mercury). Several glass tubes of different diameters, sealed at one end, were filled with mercury. After placing a finger over the opening, they were upturned into a mercury basin. The level of mercury in the tubes initially fell, then stopping at a height of 'one braccio, a quarter and a finger more', leaving an empty space at the top. Its emptiness was confirmed by pouring water into the basin. When the bottom of the tube was lifted to the level of the water, the mercury drained out, while the water 'rushed in with horrible force' until the tube was full. Torricelli proclaimed that the space created by the descent of the mercury in the tube was vaccuum, and that the air exerted on the mercury in the basin held up the column of mercury. He declared that his experiment proved two fundamental concepts: that nature did not abhor the void, and that the air had weight.

Torricelli further documented that the height of the mercury in a barometer changed slightly each day and concluded that this was due to the changing pressure in the atmosphere. He wrote: "We live submerged at the bottom of an ocean of elementary air, which is known by incontestable experiments to have weight".

By 1646, Blaise Pascal had learned of Torricelli's experiment. A confirmations of the influence of atmospheric pressure was given by the so-called "void within a void" experiment, successively performed in 1648 in several variants by Pascal, Robert Boyle, and others. If Torricelli's experiment was performed inside a vacuum pump, the mercury would not remain in the vertical tube, but would descend completely into the basin. They also observed that when, in a control experiment, air was allowed back in, the mercury rose back up the tube.

To decisevely check Torricelli's conclusions, in 1648 Pascal had a barometer carried up the mountain of Puy de Dome, where his brother-in-law made several measurements at different altitudes (Pascal himself was too sickly to clibm any mountains). Importantly, control measurements were done at appointed hours at the foot of the mountain by a local clergiman. Indeed, as the elevation increased the level of the mercury column gradually fell below that at normal ground level, while the control measurements were very stable, despite rapidly changing weather.

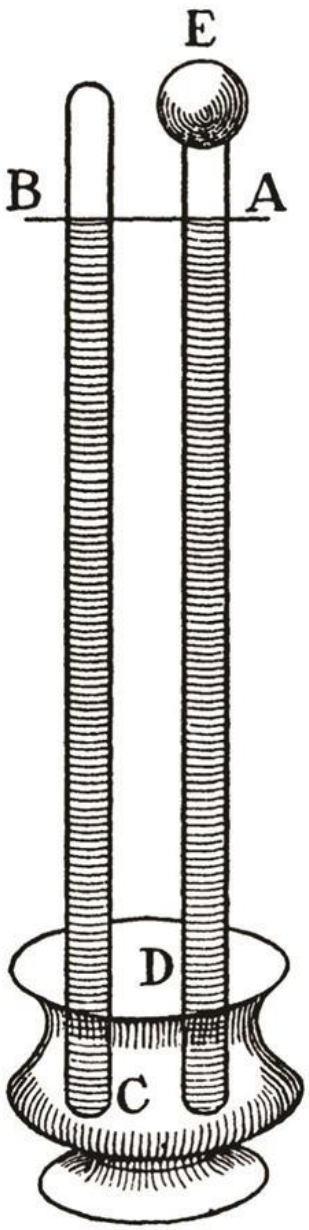


Figure 2: Torricelli's barometer.

Modern experiment

Modern clinical trials got their first theoretical exposition in 1921 when Austin Bradford Hill proposed that otherwise identical patients put into two groups, which would then be treated differently. If outcomes differed, then, he argued, we would know that the dif-

ference would be caused by the difference in treatment. As identical people are hard to find, the additional ingredient was to randomly assign subjects to the groups, hoping that their natural differences would balance out on average. The first such trials were conducted only after World War II (on streptomycin???) and the movement for evidence based medicine started in earnest only in 1990-ies, when the efficiency of drug discovery was already falling (the heyday of drug discovery was from 1950-ies to 1970-ies when most drugs now in oral use were discovered).

Modern experiments in biomedicine are nearly always of the controlled kind (although uncontrolled clinical studies do still happen, and should generally be ignored). The important parts of an experiment is that it should be 1) controlled, 2) randomized, 3) blinded.

All true experiments have this in common:

- They try to answer a causal question
- They contain an experimental treatment that modifies a cause, while not directly affecting the effect (or any other building block of the system)
- They give results that are relative to control.
- They should be randomized and/or matched, so that the treatment is the only causal factor that varies between the experimental and control arms.

A good experiment requires both appropriate design and data analysis. Specifically, the structure of data analysis models depends on the experimental design, but also on the known pre-treatment variables, on the quality of the actual experiment & measuring apparatus (including study dropouts and missing data), on blocks and clusters in the data, and on the size of the dataset.

Every experiment entails a clearly defined **experimental treatment**, which is done to the **experimental system**. The experimental system consists of the experimental protocol plus materials that go into the experiment, plus the measuring apparatus and the people who do the actual experiment. The reason behind this is to see, whether the treatment changes the behaviour of the system. The application of treatment is random, so we can believe that in a large enough sample all other properties of the system that might affect the outcome of the experiment are equally proportioned between the treatment and control groups. From the comparison of the treatment arm of the experiment and the non-treatment (control) arm, an **experimental effect** is derived as the mean value of some repeated measurements in experimental arm minus the same in the control

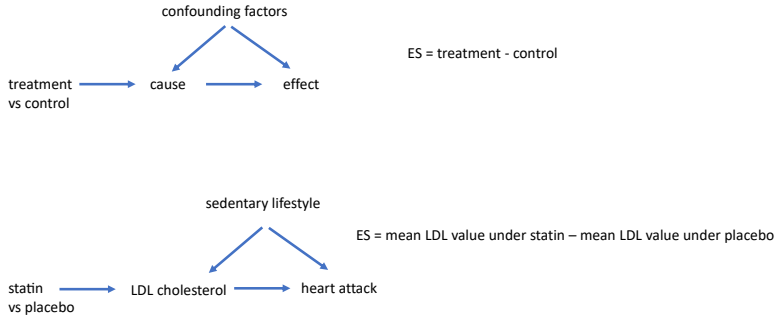


Figure 3: A general scheme of a controlled experiment and an example, where statin is the treatment that lowers LDL cholesterol, which is a causal agent of heart attacs. Sedentiary lifestyle also causes heart attacks both directly and indirectly, through increasing LDL cholesterol accumulation. This makes it hard to disentangle the causal effects of LDL and lifestyle, unless we do a randomized experiment, where people with different lifestyles have an equal chance of being given statin or placebo.

arm. Usually by "treatment" we try to very specifically disrupt a specific component of the system, and nothing else, and if we are successful, then by observing a strong experimental effect, we can conclude that whatever it was that

Useful concepts for understanding controlled experiments

- Experimental system - everything that is needed to produce the data that you got. The experimenters job is to infer from data and her knowldge about the experimental system, the data generating mechanism. For this, she uses statistical modeling. Experimental system includes study subjects (cell lines, mice, patients), chemicals, measuring apparatus, and experimenters. [9]

- Experimental treatment - some action that disrupts the experimental system according to a scientific causal hypothesis. By disrupting/modifying a hypothetical cause we aim to change an hypothetical effect/outcome. Experimental treatment should not affect the outcome directly, or by any other way than through the specified cause. Also, there should be no hidden versions of treatments. That is, for treatment effects to be analyzable, all units must receive the same well-defined treatment(s). Also, there should be no interference, of the type of treatment of unit A affecting the treatment of unit B.[10] For a study that is likely to have interference, one option is to assign the treatment at the level of a group beyond which interference is not likely, and then randomize at this level. Thus the subjects who are likely to transfer knowledge or behaviors among themselves, all get the same treatment. Then one should conservatively interpret results only at the group level.

[9] Sometimes the term is used instead to denote a commonly used study organism, for which exist worked-out experimental systems in the previous sense. Examples include *E. coli* (genetical and biochemical methods), Mouse (genetical methods), and rat (physiological methods), *X. laevis* (cell biology methods), and dalmatian dog as a narcolepsia model.

[10] Testing the effect of a fertilizer by randomly assigning adjacent plots to treatment or control is a classic example of interference. Vaccines that reduce the probability of a disease within a community could easily lead to violation. Note that an epidemiologist might well hope for interference to increase the impact of an intervention.

- experimental subject/object of measurement is whatever is measured in the experiment. Often we have different kinds of information on each subject (measured variables and non-measured variables, like sex and age). There are also repated measurement designs, where a single sibject is measured several times over the course of study, and designs where both experimental and control treatments are applied in some sequence to the same subject.
- primary outcomes - pre-defined measurements, whose results will be discussed as the main result of the study
- secondary outcomes - results of *ad hoc* subgroup analyses, or of *ad hoc* measuring protocols, which will be discussed as needing independent study to verify. Secondary analyses are often done when the results of primary analyses fail to excite. They often suffer from multiple testing problem and tend to result is false alarms.
- Control treatment - a mock treatment that provides the baseline, against which the experimental effect is estimated.
- Treatment arm/sample - a group of subjects or measurment objects, to whom experimental treatment is applied.
- Control arm/sample - a group of subjects or measurment objects, to whom control treatment is applied.
- Experimental effect size ES = mean treatment arm value - mean control arm value.
- Positive control - a sensitivity control experiment of the experimental system, which shows experimental system generating data in the absense of experimental treatment.[11]
- Negative control - a specificity control experiment of the experimental system, which shows experimental system, which lacks a necessary component, not generating data in the absence of experimental treatment.[12] Positive and negative controls are often used for lab experiments, where managing the integrity of the experimental system can be a challenge, independent of aplication of a specific experimental/control treatment.
- randomization - choosing treatment and control arm subjects using a random number generator. Full randomization means that it is impossible to predict from any attribute of the subject, into which arm the subject will fall. It does not mean that actual samples will contain an equal mix of subjects by such attributes.

[11] A PCR reaction with a standard template/primers combo

[12] A PCR reaction lacking template. Question: Why is omitting template a better idea than omitting the enzyme?

- Pre-treatment variables are the variables that are known before the act of randomization. Randomization will make them independent of the experimental outcome variable. They include both variables that do not change value over time and variables such as participant age, which change over time but cannot be affected by a treatment. One can check whether the randomization worked by comparing means or distributions of observed pre-treatment variables across randomized treatment groups.

- blocking - to increase the accuracy of ES estimation, one can divide the sample into blocks that contain subsamples of reduced variablity. There are equal number of blocks between the study arms, but not necessary equal number of subjects within paired blocks. The paired blocks are compared between the study arms. A typical block consists of a sex, age group, or comorbidity. Blocking variables are added to regression models as regular fixed effects.

- matching is an extreme case of blocking, where each block contains 1+1 subjects (one in each study arm). The subjects can be matched for several characteristics (age, comorbidity, sex, etc.). The most effective paired or blocked designs are when groupings arise naturally in the data, such as twins. It is good to match on a wide variety of characteristics using a dimension-reduction strategy such as Mahalanobis or propensity score matching.

- clustering - While in blocking we try to reduce variability within each cluster, in clustering, we try to group data such that they have the same variability. If a study is made in several centres, then each one might have its own bias and variablity and should be analysed together in a multilevel model. Incorporating cluster to the model reduces bias, but increases the uncertainty around our point estimates. Adding natural clusters (school, study centre, experimenter, chemical batch) is considered the honest thing to do in that the estimates from such models better reflect true uncertainty around the estimates.[13]

- blinding is a bias reduction measure that blinds the assignment of subjects into study arms for (i) the people who collect data and/or (ii) data analysts. Another form of blinding is to keep the group assignments in the open, but to code the data, so that its analysis is done using pre-specified protocols and without intermediate access to results, which are decoded at the very end of the analysis.

- technical replicate - experiment is replicated at technical level, often by running the same samples through measurement several

[13] Blocking and pairing are effective to the extent that units within the same block or pair are similar to each other in their pre-treatment characteristics, so that splitting each block or pair ensures some level of pre-treatment balance between treatment and control groups. In contrast, clustered designs are typically less efficient than complete randomization. NB! Blocking, pairing and clustering are welcome additions to randomized experiments in the analysis stage where they help to improve the overall accuracy of the results. In general, adding pre-treatment variables to regression models often helps to reduce uncertainty around the estimates in randomized experiments.The exceptions are when a pre-treatment variable is causally irrelevant, or when DAG-based formal causal analysis indicates its omission.

times. Such analysis allows to assess the quality of measurement, but it does not capture biological variation in nature.

- biological replicate - study samples correspond to independently collected biological samples. The goal is to capture biological variation, and therefore to be able to generalize the results of the experiment to some biological level. If the data is clustered by cage, family, etc., then appropriate multilevel analysis is needed. Also, the more biological variation is covered in the experiment, the better the experiment is generalizable, but the larger sample size is needed to reliably estimate effects of a given size. Importantly, the sensitivity of the experiment goes down linearly with increasing variation in the data, while it goes up proportionally with the square root of increasing sample size. Therefore, it is much more costly to increase sensitivity by increasing N, rather than by decreasing SD.[14] As a rule, one can only extrapolate results to the population, from which a representative sample was drawn for the study (random sampling is a good way of obtaining a representative sample). This level of extrapolation is called **external validity** of the study.
- reproducing the experiment means reanalysing the original data using the original analytic protocols, and getting pretty much exactly the same results.
- replicating the experiment means redoing the whole experiment, starting by collecting new data, and getting reasonably similar results. The idea is to use exactly the original protocols (although increasing the sample size, and maybe improving the analysis in a clearly defined way, may be permissible).[15]

[14] This is the reason why in clinical studies many patient groups that will eventually get the treatment are excluded from the trials that test the efficacy and safety of said treatment. As we are talking about patients, who are older and with more comorbidities, this is a serious problem.

[15] Conceptual replication means that it is not exatly the original experiment, but something conceptually similar, that is done. The usefulness of conceptual replication in verfying original results is uncertain.

Causal inference is in essence a comparison between potential outcomes of what might have occurred under different scenarios. This could be between a factual state and one or more counterfactual states, representing what might have happened, or it could be a comparison among various counterfactuals. Thus, ideally one would compare, how the same individual fares in parallel universes. In one she would get the experimental treatment, in another the control treatment. Our lack of access to parallel universes is **the fundamental problem of causation**. Without the ability to inspect parallel universes the next best thing is **controlled randomized experiment/trial** (RCT), involving administering experimental treatment and control tratment to different groups of individuals, who are randomized in respect to all other variables that might affect the outcome of the experiment. Then (usually) the mean values of the two groups are compared to calculate the effect size.

> In practice, we can never ensure that treatment and control groups are balanced on all relevant pre-treatment characteristics. However, there are statistical approaches that may bring us closer. At the design stage, we can use randomization to ensure that treatment and control groups are balanced in expectation, and we can use blocking to reduce the variation in any imbalance. At the analysis stage, we can adjust for pre-treatment variables to correct for differences between the two groups to reduce bias in our estimate of the sample average treatment effect. We can further adjust for differences between sample and population if our goal is to estimate the population average treatment effect. (Gelman, Hill, and Vehtari, 2020).

If asked, you should be able to answer:

1. What scientific question does your experiment address?
2. What are the reasonable controls (negative and positive controls)? Every field has its own standards, as to which controls are required.
3. Are there any blocking and/or clustering structures in the data that you should/could use in the analysis?
4. How many times have you repeated your experiment? Can you explain your sample size? Specifically, have you done a formal power analysis?
5. Do you have biological and/or technical replicates? What is the level of generalizability of your experiment?
6. Have you excluded any failed experiments from analysis and if so, why? Do you have any missing, non-measured, data? How are you addressing this in analysis?
7. How have you ensured reproducibility of your experiment? (publication of analysis code)
8. How have you ensured replicability of the experiment? (publication and sharing arrangements of materials and methods)
9. Have you pre-specified any aspects of data collection and/or analysis?

1. Irrational thinking

> "the rule that human beings seem to follow is to engage the brain only when all else fails—and usually not even then" Hull (2001).
> "In effect, all animals are under stringent selection pressure to be as stupid as they can get away with." Richerson and Boyd (2005).

Technically, we humans are a largely irrational bunch. This is not to say that we naturally think badly. We are actually doing pretty well, evolutionarily speaking, thinking and making decisions quickly where it matters. But as a trade-off, we cut corners in our thinking process. The result is innate fast and dirty thinking, which can be a poor tool for a scientist. Indeed, rational thinking seems to be highly unintuitive and needs to be learned at school. The situation is so bad that even people who have mastered logic, probability theory and statistics still continue to make the usual thinking errors, and need to sit down and concentrate to do normative thinking. This, alas, is the human condition.

The types of irrationality that affect science are usually connected with aspects of updating of belief, use of irrelevant information, or with overzealous use of causal explanations.

Some typical examples of irrationality relevant in science:

belief updating errors

1.1. underweighting of new data leads to inability to change one's mind.

- preferring old information to new (the first datapoints are used to make up one's mind, then the rest of the data is just for confirmation)
- overweighting of repeating information (How to do a metastudy from studies A, B, and C, where studies B and C, but not A, analyse the same dataset with different methods?)
- overweighting of emotionally charged information (people overestimate the frequency of plane crashes and murder because these things are in the news)

1.2. overweighting of new data leads to capricious changes of opinion

- ignoring base rates (p value only compares new data to null hypothesis). [16]
- ignoring sampling effects and sample size

1.3. incorporating irrelevant information

- confirmation bias gives higher weight to data supporting a favourite theory
- anchoring (irrelevant numbers bias your estimate)

1.4 undervaluing relevant information

- Letting emotional weight of evidence and availability to influence weight of evidence (things that first come to mind seem important; frequency of murder vs. frequency of dying because of hip fractures)

2. *ignoring the inherent randomness of life*

- gamblers fallacy – misunderstanding of randomness
- disregarding sample size (A hospital reported an annual fraction of 0.55 of girls from all births. Is it more likely/less likely/equally likely that this hospital is in a small town than a big town?)

3. *imposition of overly strong causal structures over nature.*

- conspiracy theories establish order through well-defined causes, as do detective stories. For example, in the spring of 2021 15% of UK respondents believed that it was more reliable than not that "the coronavirus is part of a global effort to enforce mandatory vaccination", 11% believed that the epidemic is a jewish conspiracy to further their financial interests, and 9% believed that "the new 5G network may be making us more susceptible to the virus".[17]
- Magical thinking uses subtle causes to understand and manipulate the world (subtle causes come with oversized effects, like a black cat crossing the road). For example, Babylonian science has been divided by modern scholars into five distinct scholarly professions: (i) the "scribe of Enuma Anu Enlil, who was expert in astral phenomena (ii), the "one who inspects (the liver and exta)", i.e., the diviner expert; (iii) the "exorcist", who treated human beings afflicted by divine disfavor via incantations and rituals aimed

[16] Example of base rate neglect: the cab problem: A cab was involved in an accident at night. You are told that 85% of the cabs are green and 15% are blue. A witness reported that the cab was blue. The witness correctly identifies each of the two colors 80% of the time. What is the probability (expressed as a percentage) that the cab involved in the accident was Blue?

Subjects do not have to calculate the answer precisely. The point is to combine the two pieces of information: (i) 15% of the cabs are blue and (ii) 80% witness accuracy. The probability that the cab is blue is 41%. Less than half of subjects give answers between 20% and 70% and most answers are around 80%.

[17] `https://www.theguardian.com/theobserver/commentisfree/2021/may/16/do-people-believe-covid-myths`

at re-establishment of the right relationship between human and divine; (iv) the "physician", who treated the body in the grip of demonic or divine influence (what we call disease); and (v) the "lamentation priest", who was responsible for religious ritual performance (songs of lamentation, also the playing of the kettledrum for the ritual against the evil of a lunar eclipse). This scientific tradition lasted for 2 millenia.[18]

[18] The Cambridge History of Science vol. 1, 2020.

- regression to the mean and placebo effect

In general, people tend to use heuristics, meaning that they substituting simple questions for hard ones. This often leads to errors of judgment, but it also meshes with reductionism in science.

- framing (patient has a 6% chance of dying vs. patient has a 94% chance of surviving)
- overconfidence (CI)

For reference on pitfalls of natural human thinking see

Tversky, A., and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. Science (New York, N.Y.), 185(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124.

Pohl, R. (2004). Cognitive Illusions (pp. 1–451). Psychology Press.

Stanovich, K. E., West, R. F., and Toplak, M. E. (2016). The Rationality Quotient: Toward a Test of Rational Thinking (pp. 1–479). The MIT Press.

2. *Rational thinking under certainty.*

Certainty means that you have all the relevant data, which comes without measurement error, bias, or anything else that might make it unreliable. Then rationality is more or less the same thing as logical thinking. More precisely, propositional logic can be viewed a model of rational argument. Schematically, logical thinking looks like this:

True premises -> the algorithms of logic -> true conclusions.

The algorithms are used to convert premises into conclusions. If this is done correctly, then the truth of the conclusions is guaranteed as a logical necessity.

For example:

- Premiss1: All men are mortal
- Premiss2: Socrates is a man
- Conclusion: Socrates is mortal

The algorithm is given by the truth table, and the logical rule is called the conditional, a.k.a. the implication:

- If A then B (a.k.a. A -> B)
- The truth table is:
-

A	B	if A then B
T	T	T
T	F	F
F	T	T
F	F	T

The conditional is FALSE if and only if A is TRUE and B is FALSE. In all other cases the conclusion is TRUE. A -> B does not mean that B is logically deducible from A. So, in any everyday meaning of the word, A does not imply B. All A -> B means is that A and B have the same truth value. In formal logic every true statement implies every other true statement.

Also note that logical implication is a convenience rule rather than a fundamental rule of logic. The whole of propositional logic can be deduced from any of several sets of simpler rules, including {conjunction and negation}, {NAND – not(A and B)}, {NOR – not(A or B)}. This is why you can build computers that have a single type of logic gate. For example, a NAND gate's output is up when at least one of the inputs is down, and down if both inputs are up. By combining NAND gates one can create any logic function.

There are 3 things to remember:

1. The concept of causality is not encoded in logic. Logic is about co-occurrence of A and B; it cannot even express the thought that A might be a cause of B. Mathematical description of causal relations is its own thing, independent of classical logic and probability theory and statistical inference. We will come to it later.

2. Propositional logic is monotonic, meaning that once a conclusion is achieved, then it cannot in principle be changed. Getting new data after forming one's conclusions makes no difference – once you have a conclusion, this conclusion stands forever.

3. Logic is deductive, like mathematics – it essentially provides risk-free arguments with the downside being that technically no new information can be bought into being via logical argument. Logic is an axiomatic system, like geometry. From axioms one can deduce theorems, a.k.a. rules for doing stuff, but these rules are already embedded in the axioms.

The previous points show that while the rules of logic can be important for the general structure of scientific argument (the most basic of them being that a thing cannot be TRUE and FALSE at the same time), they are clearly not sufficient for explaining, or modelling, or prescribing the scientific argument. This does not mean that a scientific argument is extra-logical or illogical, it merely means that we need an expanded non-monotonic logic that can deal with uncertainty.

We need an expanded of logic that allows to make risky conclusions under uncertain information. This expanded logic must reduce into classical logic, when uncertainty is reduced to certainty. The "risky" part means that the expanded logic must be inductive – it

must be able to generalize from particular data (countable instances of something happening) to general conclusions. Obviously, these general conclusions cannot be certain (uncertainty can never beget certainty), which leads us to the next requirement: the new logic must be able to quantify uncertainty – to give it numerical values.

This expanded logic is called probability theory, and it quantifies uncertainty as probabilities. Probabilities are real numbers between 0 and 1 that conform to certain mathematical rules, which we describe later in chapter ...

An historical aside: tri-valued logic

In 1465 Peter de Rivo of the University of Louvain was asked by a student a thorny question: after Christ had told to St Peter "Thou will deny me thrice", could St Peter, in principle, not denied Christ? If the answer had been "No", then the de Rivo would have committed the heresy of determinism (without free will there would little point in christian doctrine of salvation). On the other hand, if the answer had been "Yes", then that would have made Christ a liar, at least potentially so. In the event, the professors answer was "yes, St Peter had the power to not deny Christ, but this would not have made Christ a liar because his utterance to St Peter was neither true nor false, but has neutral truth value. And the contradictory truth value for a neutral statement is itself neutral. Accordingly, if you claimed that St Peter did not deny Christ, then your claim would be heretical only if it is false (in the sense that St Peter actually denied Christ as predicted). Thus, for a prediction about tomorrow to be true today, it must be unpreventably so."Either there is no present and actual truth in the articles of faith about the future, or what they say is something that even a divine power cannot prevent." This trivalent logic was condemned by the Pope in 1474, not to reemerge until the 20th century. The proof that we can build a self-contained probabilitistic logic where between True and False lie the infinite set of real numbers between 0 and 1, so that this logic is fully compatible with the traditional bivalent logic, had to wait until the middle of 20th century.

3. *Rational thinking under uncertainty.*

3.1. understanding uncertainty

There is a pandemic coming our way, like a lurching sea monster from the deep pit, and we need to model its course. To do this we need, among other things, to fix the model input that is called the Ro (the number of people infected on average by a single infected person). Initially, we will have little data to estimate the Ro on, which makes Ro a *known unknown*. So we estimate it on the data that we have, only to end up with wide error margins, which will be carried over into our prediction of the course of the epidemic. This way we transfer uncertainty from data to model to prediction to interpretation to policy. Surely the policy makers deserve to know that there is a large but quantified level of uncertainty about how many hospital beds are needed within a month or so, as this would allow them to implement policies that err on the safe side. Or, alternatively, we could just pick a couple of point values for the Ro from thin air and run a couple of models that result in predictions with much narrower error margins. Of course, everybody loves certainty. Indeed, this is pretty much what happened in March-April 2020 with a dozen or so covid models in the UK, US, and Estonia, which gave wildly different predictions with non-overlapping quite narrow errors. In some cases these models resulted in policy decisions that clearly failed. This way we have manufactured model-based certainty, by incurring the cost of banishing the residual uncertainty into the murky realm of *unknown unknowns*. We only become aware of the presence of unknown unknowns from absence of success, and as we cannot really understand, much less quantify them, they are really scary.[19] The moral of our story is that the aim of the scientist should be to tame as much of the wild uncertainty that she can, by turning it into quantifiable known unknowns. It is not to convert uncertainty into certainty, which is logically impossible, as much as building a *perpetum mobile* is physically impossible, and often leads to grief.

In any decent model of a biological process there are different types of uncertainty that include

[19] "Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns - the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tends to be the difficult ones." Donald Rumsfeld (US Secretary of Defence, 2002)

(i) the uncertainty arising from biological variation

(ii) the uncertainty arising from sampling error

(iii) the uncertainty arising from the model fitting algorithm

(iv) the uncertainty arising from the model structure (linearity, normality, and whatnot)

(v) the uncertainty arising from what is actualy modeled vs. what we want to know based on this modeling.[20]

[20] It is a thorny question, how fundamentally these uncertainties really differ from each other. Here we take the strong view that ultimately all uncertainties that matter (with the possible exception of quantum mechanics) are epistemic, meaning that they are in our heads. For example, if we could see every atom, every electron in the cell, we could perhaps predict every mutation and lose the randomness from biological variation

Taken together, these motley sources of uncertainty pretty much guarantee that our predictions of complex phenomena are overconfident. This holds for both formal and informal predictions. For example, In April 2020, David Spiegelhalter and his colleagues asked 140 UK experts and >2,000 non-experts for prediction on the number of Covid-19 deaths by the end of the year. Experts median estimate was 30 000, whereas the non-experts was 20 000. The true number was 75 000, being inside only a third of the experts' prediction intervals and in only a tenth of the non-experts'. If people estimate a 95% prediction interval for something and reality falls into this interval in only 30% of the cases, then on average they underestimated uncertainty very greatly; while not only realizing it, but rejecting even the possibility of values similar to the true value as very unlikely. A major problem with formal modeling of uncertainty is that it gives false confidence about uncertainty!

Let's assume, for the sake of argument, that we live in a fully deterministic world where there is no uncertainty concerning the past, present, or future. In this world it is possible, in principle, to predict exactly when a butterfly flips its wings a thousand years from now. The deterministic world is fully described by a set of rigid cause-effect relations, and there is no stochasticity in it. Objectively, in there nothing is uncertain. But still, there are humans in the ointment, and humans tend to be fallible. Even when they measure a quantity with a perfectly calibrated instrument, there is always some measurement error. And that error is enough to introduce uncertainty into our deterministic world. This uncertainty is subjective, or epistemic – it belongs to us, not to the thing that was measured – and it propagates to any conclusions based on these measurements. In such a world, even if in principle we could say that patient X dies within a week, because of this kind of propagated error we can only say that it is plausible that patient A dies within a week, or that there is a risk that patient X dies within a week.

Of course, things can be more or less plausible and the risk can be higher or lower, which means that we need a language that allows numerically quantification of these things. This language is the probability theory. We define probabilities mathematically in chapter . . . , but here suffice it to say that probabilities are real numbers between 0 and 1 (although not every scale 0. . . 1 is probability scale). If we denote the hypothesis "patient X dies within a week" with "A", then P(A) is the probability of the hypothesis A being true (patient X really dying).

- P(A) = 0 means that we are absolutely sure that X stays with the living (but can we ever be sure of such a thing?),
- P(A) = 0.1 means that we are pretty sure that X lives,
- P(A) = 0.5 means that we think that both outcomes are equally likely.
- P(A) = 0.9 means that we are as sure that X dies as we were sure that X lives, when P(A) was 0.1,
- P(A) =1 means that we believe that it is a sure thing that X is dead within the week.

Now, after the week is up the P(A) can only be 0 or 1 (unless we lost track of X), but until that time, or until X dies, the P(A) will move between these extremes, subject to X's condition and our knowledge of it.

Why we need numerical probabilities? The answer is twofold: (i) we can do formal calculations with numerical probabilities, which is basically what statistics is, and (ii) it has proved patently impossible to make different people understand the non-numerical representations like very likely, extremely likely, somewhat likely, the same way. This holds even for professionals.

Now, suppose that we live in a world, where truly stochastic stuff happens (be it in a quantum level or in weather prediction, or wherever). This world is only partially causal, meaning that (i) cause-effect relations can come with built-in randomness (sometimes they work, sometimes they don't, like smoking causing cancer), or (ii) some relations are inherently acausal (completely unpredictable, like radioactive decay) Does the meaning of P(A) change? The answer is, no. In this world we still have measurement uncertainty, which we try to reduce as much we can, and on top of this is the real, irreducible uncertainty, which comes from how the world itself works. They converge, and can never be disentangled, in our final probability P(A), which is still epistemic.

Another example: the weather report says that P(rain) is 60%. What does it mean? It either rains or it doesn't, so P(rain) cannot describe a physical feature of the world. "60% rain" cannot be objective in the sense of "being really there". Usually P(rain) is calculated by running a weather model many times with slightly differing input data and counting the relative fraction of runs where the model predicts rain. If 6 models out of 10 predict rain, then the weatherman says that "P(rain) = 0.6, or there's a 60% chance of rain". Although the 60% was derived from a relative frequency, this does not matter for the interpretation – the probability is still epistemic.

I'm writing this because traditionally it is taught to students that probability is a long-run relative frequency of events that really happen, or at least could happen in principle, and that therefore P(A) is "objective" and "a real thing". It has been understood for a century that this thinking is fundamentally flawed, as it ascribes reality to fictional objects. So probability is a bit like a unicorn: a beautiful fiction. Likewise, there is no shame in interpreting P(A) as a frequency, but this does not make it "objective" in any meaningful sense.

Numerical measures of uncertainty

- Risk, probability - nr_events/(nr_non_events + nr_of_events) a.k.a succeses/tries
- Relative risk or risk ratio (RR) - $\frac{P(A|B=0)}{P(A|B=1)}$. Probability of cancer for smokers divided by probability of cancer for non-smokers.
- Relative risk reduction is how much risk is reduced in an experimental group compared to a control group. (CER – EER)/CER, where CER is control group event rate and EER is experimental group event rate. If 70% of the control group died and 35% of the experimental group died, the relative risk reduction for the new drug would be (0.70 - 0.35)/0.70 = 0.5 or 50%. In other words, the death rate in the experimental group is half that in the control group.
- Absolute risk reduction (risk difference, AAR) is the absolute difference in outcomes between 2 groups. It tells how much the risk (Pr) of something happening decreases if a certain intervention happens. AAR = CER (Control Event Rate) – EER (Experimental Event Rate). 25% of people on medication A have poor outcomes, but 8% of people who receive medication B report bad outcomes. The absolute risk reduction is 25% – 8% = 13 percentage points.
- The hazard ratio (HR) is a comparison between the Pr of events in a treatment group, compared to the Pr of events in a control

group. It's used to see if patients receiving a treatment progress faster (or slower) than those not receiving treatment. HR can be defined as the relative risk (RR) of an event happening at time t. HR = 3 means that three times the number of events are seen in the treatment group at any point in time. In other words, the treatment will cause the patient to progress three times as fast as patients in the control group. HR= 1 means that both groups are experiencing an equal number of events at any point in time. HR= 0.333 tells that the hazard rate in the treatment group is one third of that in the control group. Hazard ratios can be used to: show the RR of a complication in treatment group vs. control group, to show whether a treatment shortens an illness duration, or to show which individuals are more likely to experience an event first. While a HR is similar to RR, it isn't exactly the same. Let's say a clinical trial investigated survival rates for two drugs (A and B). The reported HRs and RRs were both 3. ThecRR tells that the risk of death is 3 times higher with drug A over the entire period of the study (i.e. it's cumulative). The HR tells you that the risk of death is 3 times higher with drug A at any particular point in time. When evaluating hazard ratios, also use another measure such as median survival time, overall survival, or time to progression. These tell about clear benefits (i.e. survival time is increased on average by 180 days).

- Odds = P(event)/P(non_event). $odds = \frac{P}{1-P}$ where P is probability. If P = 0.5 then odds = 1/1 or 1:1 or 1 to 1.

- OR or odds ratio $OR = \frac{odds(cancer\ in\ smokers)}{odds(cancer\ in\ non-smokers)}$. OR is always larger than RR. If P(event) is low (<0.1) then OR is similar enough to RR, otherwise OR can be larger by an order of magnitude or more. OR has no intuitive interpretation (the RR does). OR = 1 means that exposure to A does not affect the odds of property B. OR > 1 means that there is a higher odds of property B happening with exposure to A. Only the OR can be used in case-control studies. Because to calculate the RR, one must know the risk. Risk is a proportion of those exposed with an outcome compared to the total population exposed. This is impossible in a case-control study, in which those who already have the outcome are included without knowing the total population exposed.

3.2. Probability as counting

In a commencement speech at USC in 1994, the famous investor Charlie Munger said,

> "If you don't get this elementary, but mildly unnatural, mathematics of elementary probability into your repertoire, then you go through a long life like a one-legged man in an ass-kicking contest. You're giving a huge advantage to everybody else. One of the advantages of a fellow like Buffett, whom I've worked with all these years, is that he automatically thinks in terms of decision trees and the elementary math of permutations and combinations. . . ."

Lets find out, what this means.

We start with a simple question: "what is the probability of getting a 6, when throwing a fair die?". Surprisingly, it took the gamblers a couple of millennia, until 16th century to find out the answer - specifically, it was first described by Fazio Cardano (1444–1524). The truth is that we can simply throw a die N times and count the k sixes that we get. Then P(six) = k/N. The probability of sixes is nothing more than the relative long-run frequency of sixes over all throws of the die. As the N increases, the estimated P(six) will gradually approach the true probability of sixes. If the die has six sides, which are equally likely to come up, then P(six) = 1/6, as is the probability for getting any number between 1 and 6.

3.2.1. Counting from a series of events

1. Imagine all possible results of a single throw of the die: {1, 2, 3, 4, 5, 6}. This is akin to imagining all possible futures that can ensue from a throw of a die.
2. Count all possible futures, where the result is a six. As there is only one, k = 1.
3. Count all possible futures. As there are 6, N = 6.
4. Find the probability of sixes as P(six) = k/N = 1/6.

This scheme seems almost too trivial to show, but as we shall soon see, it is actually quite subtle, and extremely important for understanding the scientific method.

The further assumptions of this kind of futures counting (besides that all futures are equally likely to occur) are stability of the system (i.e., the properties of the die, the thrower, the table, the laws of physics, etc. do not change during the experiment), the constancy of nature (this is an assumption behind all causal thinking), and exchangeability (i.e., the result does not depend on the order that you record the data).

The advantage of this frequency interpretation of probability is that it is fairly easy to grasp; but more importantly, it seems to be objective, as it describes a process that happens in the world (in this

case, someone throwing a die). Although, one should be aware that the process so described includes the thrower, as well as the die, the table and the laws of nature, to say the least. On the other hand, a serious disadvantage of the frequentist probability is that it can only apply where there is a frequency of events. Clearly, scientific hypotheses, being unique manifestations of human ingenuity, do not have a frequency. Thus, people whose repertoire gets no further than simple frequentist probability may call themselves objective, but they cannot even ask the question, "what is the probability that I am right in calling myself that?"

3.2.1.1. Counting the events in a series of experiments leads to frequentist statistics

Words: frequentist statistics, null hypothesis (H_0), central limit theorem, expected value, confidence interval (CI), credible interval (CI), p value, type I error, standard error of the mean (SEM), significance level (α), False Discovery Rate (FDR)

Interestingly, the people who think they cannot ask the question about probability of a hypothesis, nevertheless try to answer it. For this they use something called frequentist statistics or null hypothesis testing. This is a statistical paradigm created by R.A. Fisher, Jerzy Neyman and Egon Pearson in 1920-1930-ies. The crux of frequentist statistics is that while hypotheses have no frequency, datasets do (albeit, only under certain precisely defined conditions). Imagine an infinite population of measurement values, from where you randomly and independently draw a sample of N values, then calculate a statistic, for example the mean. This sample mean reflects the population mean, but because of sampling error it does not do so precisely.

> The sampling error tends to smaller when N is larger – and it gets smaller proportionally to the square root of N. So, as N grows, it gets progressively harder to push the sampling error downwards.

Now, imagine that we draw k such samples (each with size N) and calculate k means. The central limit theorem of statistics tells us that no matter what the sample data distribution, those means are approximately normally distributed. Note that for the central limit theorem to work, N must be at least 30, dependent on the data distribution. The overall mean of the k means is then the most likely estimate for the population mean, which in turn is the value most likely to be encountered in the population. We illustrate this by first

simulating 1000 random samples, each with N=3, from standard log-normal distribution, and then calculating 1000 means. Not surprisingly, plotting the raw data leads to a nice lognormal distribution (dashed black curve), and plotting of the means (continuous black curve) leads to a fairly similar curve. However, when the 1000 samples get bigger (N=20, red curve; N=60, blue curve), the distribution of their means gets closer to normal.

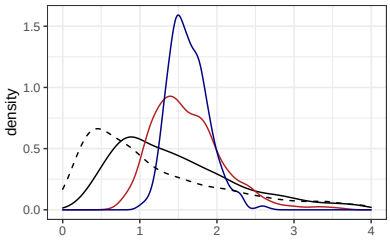


Figure 4: Density plot of distribution of 1000 means each of 1000 log-normally distributed samples. Black curve, N=3; red curve, N=20; blue curve, N=60.

Moreover, from the normal distribution it is easy to find the region that contains X% of the area under the curve, or X% of the probability density. The parameter values that delineate this area define X% confidence interval (CI).

The interpretation of X% CI is that if we draw many samples and calculate as many means, then the true population mean falls inside X% of them. [21] Formally, the CI makes a frequency statement about the error properties of the algorithm that calculates the CI, rather than about your sample mean. Nevertheless, if we have no background information to take into account and if the aforementioned assumptions (especially those of normality, and random and independent sampling) hold, then a confidence interval is numerically equivalent to Bayesian credible interval (also abbreviated as CI), whose interpretation is simply that X% CI means that the true expected value lies within this interval with X/100 probability. So you could view your confidence interval as a quick and dirty approximation to the credible interval, which you really want to have; but only if some strong assumptions hold.

[21] This does not mean that your particular sample mean lies within the X% CI with X% probability.

The confidence interval is not the only, or even the most popular, tool of frequentist statistics. This distinction belongs to the p value.

The logic of p value goes like this.

1. As hypotheses do not have probabilities, but datasets do, we can speak of probability of data under some hypothesis – $P(d|h)$ or "probability of data given the hypothesis is true" – in the hope that this says something about the probability of that hypothesis, or $P(h|d)$.
   For a two-group comparison (experimental values versus control values, say) the easiest way to do that is to define a null hypothesis, or H_0, whereby the means of the two groups are identical [22].

2. Now, if there really is no effect, then by virtue of the sampling error we expect to see an effect anyway (this is called sampling effect or sampling error). And any true effect will be on top of the random sampling effect, which is equally likely to increase or decrease our estimate of the true effect. Therefore it is important to be able to distinguish between sampling effects and possible

[22] that is, on average the experimental treatment has no effect at all or mean(treatment value) – mean(control value) = 0

true effects. There are two possibilities. (*i*) If the true effect is not much larger than the sampling effect, they cannot be disentangled (at least not without collecting new samples). We must then say that the observed effect is consistent with the sampling effect and declare that we could not find evidence for any true non-random effect. (*ii*) If the observed effect is much larger than estimated sampling effect, then we can declare with confidence that the effect is non-random. Alas, non-randomness in no way guarantees that the effect was actually caused by the experimental treatment.

3. We model the sampling error of the H_0 as a normal distribution, whose mean equals zero (the zero effect) and whose standard deviation equals the sample standard deviation divided by the square root of N.[23] The standard deviation of the means (as opposed to original data points) is called the standard error of the mean or SEM. If N is large enough (at least 30), then $mean +/- SEM = 68\%\ CI$, $mean +/- 1.96SEM = 95\%\ CI$ and $mean +/- 3SEM = 99\%\ CI$.

[23] The null hypothesis does not have to be normal, but historically they usually are.

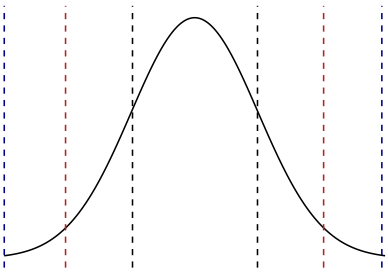


Figure 5: A null hypothesis with SEM, 1.96SEM, and 3SEM shown as black, red and blue dashed lines, respectively.

We use the sample data in estimating SEM of the H_0 because statistical software assumes that the scientific hardware (your brain) has no independent knowledge of the true population variation, and thus the sample values is used in lieu of population values. However, sample estimates of SEM can be highly misleading when N is small. Also, when the data is non-randomly drawn from the population, then even a very large sample is likely to be a poor representation of the population. As in many lab experiments N tends to be very small and it is often hard to even say, what a random sample means, H_0 testing tends not to work well in many branches of biology.

We now calculate the mean of our sample and superimpose it on the normal distribution of H_0. If our sample mean is situated where the H_0 distribution is still high, then it is consistent with the sample effect. On the other hand, if the height of the null distribution is very low, then the sample mean is inconsistent with the sample effect. Quantitatively, the fraction of the H_0 density that is cut of by our sample mean is the one-sided p value.

> p value is the long run relative frequency of obtaining data that is as far or further away from H_0 than your sample data, if the H_0 holds.

The p value was introduced by Roland Fisher in 1923 for use as an informal measure of strength of evidence. If p is small, 0.001, say, we might be tempted to claim that we have strong evidence that the H_0 is false.[24]

[24] This intuition can be very wrong.

A decade or so later Neyman and Pearson added to it something called significance level, often denoted by α, and derived a quite

different use for the p value in the long-term quality control of the experimental setup. In their scheme, which is mathematically impeccable, p value is not a direct comment on the experiment at hand, but rather on the experimental system that generated this experiment.

The goal of employing α is to provide an automatic procedure that makes discoveries, and does so with fixed error rates. By discovery we mean that "an effect is real", or something like it. A discovery dichotomizes our contiuous measurements into statistically not significant non-discoveries and statistically significant discoveries, which can be either true discoveries or false discoveries (false-positives).

If $p < \alpha$, then we can call the p value statistically significant. By convention, α is set either as 0.01, 0.05 (the most often used value), or 0.1. By setting α we keep the long run relative frequency of type I errors at α level.[25] This dichotomization, which comes at a cost of information loss, allows to define the frequency of **type I errors** as the relative frequency by which true H_0 are falsely classified as "statistically significant. Thereby we have a formal procedure that allows us to make discoveries without any more scientific input as data. If this actually worked, being a scientists would be a low-education job.

[25] Thanks to α, we can define the confidence interval through the p value: $X\%$ CI covers a range of parameter values, whose p values are not statistically significant at significance level $\alpha = 1\text{˜}X$. Thus, the CI and the p value carry redundant information.

The Neyman-Pearson procedure is fine, statistically speaking. But its use in science carries great dangers. When scientists employ it with α set to 0.05, they implicitly agree to publish on average every 20th true H_0 as a statistically significant false-positive finding. The level of harm that ensues depends on the true frequency of true effects. This is quantified as the False Discovery Rate (FDR), where FDR = false significant effects / all significant effects. If scientists test predominantly true effects (most experiments end up in showing a true effect of experimental treatment), then FDR is low and the use of statistical significance is not an obvious folly. On the other hand, if scientists are taking risks and testing hypotheses, most of which are false (most H_0s are true), then the Neyman-Pearson procedure leads to a terribly misleading picture of science as a whole. As philosophers seem to agree that testing risky hypotheses is what good scientists should be doing, therein obviously lies a problem.

3.2.2. Counting in the multiverse

Luckily for us, after 16th century with its borderline medieval thinking came the 17th and the scientific evolution. The thing that really unleashed scientific rationality was a question of gambling. Consider a game where 2 players equally pool their money to a pot that goes to the winner. The game is simple: they cast dice and after each round, whoever gets the largest sum, gets a point (draws don't count). The game goes on until someone collects 9 points and

wins the pot. But what happens when the game is suddenly interrupted after, say, player A has 7 points (and two to go) and player B has 8 points (and one to go)? Is there a just way of dividing the pot? This question bugged the best minds of Europe for centuries, when it was pretty universally thought that there isn't a solution, until in 1654 Blaise Pascal and Pierre de Fermat solved it. While earlier attempts centered on getting an answer from the games actually played (the 8 and 7 points in our example), which led to contradiction, the insight of Fermat was that one could do a counterfactual analysis by pretending that the game goes on, albeit not in our world, but in an expanding array of parallel worlds.

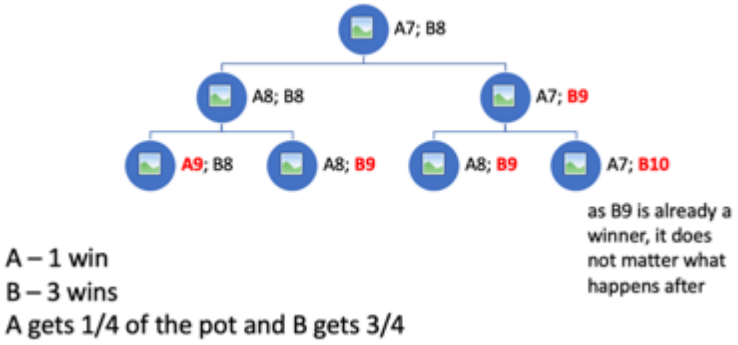


Figure 6: The game continues in the multiverse

Here the starting point (A7B8) is in our universe. Then the 1st imaginary round splits it into two imaginary ones (A8B8 and A7B9), one of which leads to a win by player B. But on the next round player B does something that he would presumably never do in life: although there is no way for him to lose, he nevertheless keeps on playing and consequently in the second round the worlds split again, now into four. At this level every game ends in somebody winning and we can count the relative frequency of A-s wins (1) and B-s wins (3). Thus it makes sense to give ¼-th of the pot to A and ¾-th to B. This is the same as to say that P(A wins) = ¼ and P(B wins) = ¾. Probability theory is nothing but counting across the multiverse. Well, actually it is mostly about inventing algorithms that speed up this counting, as the multiverse, expanding exponentially, soon becomes non-conductive to hand-counting.

At a more conceptual level, we now have a frequency probability for a unique event. But in the multiverse everything has a frequency, and thus a probability. Since the summer of 1654, even scientific theories have probabilities. So we can bring to bear the mathematics of probability theory that is based on counting these frequencies.

3.3. From rules of combinatorics to probability distribution

3.3.1. The Addition Rule

Let E_1 and E_2 be mutually exclusive events. Let event E describe the situation where either E_1 or E_2 will occur. The number of times event E will occur can be given by the expression:

$$n(E) = n(E_1) + n(E_2)$$

where

- $n(E)$ = Number of outcomes of event E
- $n(E_1)$ = Number of outcomes of event E_1
- $n(E_2)$ = Number of outcomes of event E_2

3.3.2. The Multiplication Rule

Suppose that E_1 can result in any one of $n(E_1)$ possible outcomes; and for each outcome of the event E_1, there are $n(E_2)$ possible outcomes of event E_2. Together there will be $n(E_1)n(E_2)$ possible outcomes of the two events. That is, if event E means that both E_1 and E_2 must occur, then

$$n(E) = n(E_1) \times n(E_2)$$

If you have a choice of n_1 elements for the 1st position, a choice of n_2 elements for the 2nd position, and so on up to n_k, then the number of possible combinations of sequences of k elements is $n_1 \times n_2 \times \ldots \times n_k$.

3.3.3. Permutations

How many different combinations from 3 elements ABC can we get? The answer is 6: {ABC, ACB, BCA, BAC, CAB, CBC}. This geeralizes as $n_1 = 3$, $n_2 = 2$, $n_3 = 1$ -> $n_1 \times n_2 \times n_3 = 6$, or by using n factorial: $n! = 3! = 3 \times 2 \times 1 = 6$.

n factorial is defined as the product of all the integers from 1 to n.

$$n! = (n)(n-1)(n-2)...(3)(2)(1)$$

n distinct objects can be arranged (permuted) in n! ways.

3.3.4. Variations (the order matters and repetitions are not allowed)

n – nr of elements k – the length of the sequence

$$V_n^k = n(n-1)\dots(n-k+1) = n!/(n-k)!$$

8 athletes compete, how many different possibilities for the first three places? $n_1 = 8$, $n_2 = 7$, $n_3 = 6$

8 x 7 x 6 = 336
(8 - 3)! = 5!
8!/5! = 336

3.3.5. Combinations (the order is not important)

How many possible combinations of k elements is it logically possible to get of n distinct elements, or $\binom{n}{k}$?

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

There is a single way of obtaining 0 elements out of n possibilities

$$\binom{n}{0} = 1$$

and there is a single way of obtaining n elements out of n possibilities

$$\binom{n}{n} = 1$$

The number of different sets of 4 letters which can be chosen from the alphabet is $\binom{26}{4} = \frac{26!}{4!(26-4!)}$

3.3.6. From counting to binomial probability distribution

Q: What is the probability $P_n(k)$ that event A occurs exactly k times out of n tries?

1. we start by assuming that all k outcomes come up in an uninterrupted sequence. We look at the full series of n tries. Then by the multiplication rule:

$$P(A_1\dots A_k \neg A_k + 1 \dots \neg A_n) = P(A_1)\dots P(A_k)P(\neg A_k + 1)\dots P(\neg A_n) = p^k(1-p)^{n-k}$$

We would get the same results no matter where the k outcomes are in the sequence. If the sample space only has two members: $\{0,1\}$, a.k.a. if k is limited to only 0 or 1, and if n = 1, then this becomes Bernoulli probability mass distribution

$$f(k,p) = p^k(1-p)^{n-k}$$

Example: if p = 0.1 and k=0, then $f(0,0.1) = 0.1^0(0.9)^1 = 1 \times 0.9$; and $f(1,0.1) = 0.1^1 * 0.9^0 = 0.1 \times 1$

This is, however, not the answer to our question, where both n and k can be any integer equal or larger than zero.

2. If k can be > 1, then we also need to take into account that the nr of combinations of length n, where event A comes up k times, is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

3. By multiplying 1. And 2. [Pr of a single series times the number of combinations with k events] we get

$$P_n(k) = \binom{n}{k} p^k(1-p)^{n-k}$$

This is the binomial distribution.

Example: What is the Pr of getting 6 heads out of 10 coin tosses?

$$P_{10}(6) = \binom{10}{6}(1/2)^6(1/2)^4 = (10!/(4!6!))(1/2)^6(1/2)^4 = 0.205$$

Table: Full binomial probability distribution for $P_{10}(5)$

k	0 & 10	1 & 9	2 & 8	3 & 7	4 & 6	5	sum
Pn	0.001	0.010	0.044	0.117	0.205	0.246	1.000

We can plot this probability distribution as a symmetric graph (it is symmetric as long as A and notA are equally probable).

3.4. Probability distribution

As probability measures epistemic uncertainty, the role of the probability distribution is to provide the full description of uncertainty over the sample space. The formal output of any statistical analysis is a probability distribution.

> A probability distribution lists for every member of sample space the probability of this member.

The sample space (Ω) is the full set of mutually exclusive and exhaustive hypotheses/parameter values that comprize all logically possible futures (the multiverse).[26]

Importantly, we can slice the Ω cake in infinitely many ways, and the probability distribution that we end up with depends on the slicing. Whatever may be said of the downstream analysis, slicing the Ω is always a scientific, rather than logical, exercise. A robot may well be able to calculate a probability distribution, but this will not turn it into a scientist!

[26] Ω could be {H_1 – my wife is cheating on me; H_2 – my wife is not cheating on me}, or {H_1 – my wife is cheating on me with my best friend; H_2 – my wife is cheating on me, but not with my best friend; H_3 – my wife is not cheating on me}.

3.4.1. Binomial disrtibution

For a coin throwing experiment the omega is $\{heads,\ tails\}$, for measurements of a continuous variable Ω is the infinite set of real numbers, and so on.

What sort of information does a probability distribution contain? Say, we have a fair coin which we toss 10 times and count the number of heads. We model this experiment as a Bernoulli process, leading to binomial probability distribution. The sample space X is then some number of heads between 0 and 10, a.k.a. $X = \{0, 1, \ldots, 10\ heads\}$.

Any sequence of experiments conforming to these requirements is a **Bernoulli process**.

- n independent experiments (tries), each with two possible outcomes ($\Omega = A \wedge \neg A$). [27]
- If A_i is the outcome A in i-th experiment, the $P(A_i)$ is constant over the experiments
- $P(\neg A) = 1 - P(A)$

[27] Sample space (Ω) is a list of all possible outcomes.

```
data <- tibble(x = 0:10, y = dbinom(x = x, size = 10, prob = 0.5))

ggplot(data = data, aes(x = x, y = y)) + geom_col() + theme_bw()
```

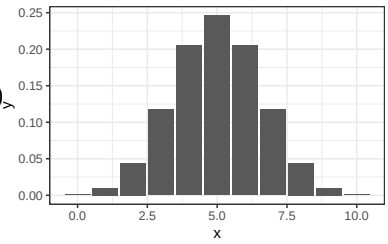


Figure 7: Binomial probability distribution models tossing of a coin. x is sample space, y denotes probability of each member of Omega.

The most likely outcome of this experiment, X = 5, is at the highest point of the binomial distribution. And consequently, the higher the distribution over some value of X, the higher the probability that we will get as many heads from our experiment. We can look at this as a probability ratio – if the probability density over some value $X = x_j$ is two times higher than over some other value $X = x_k$, then the result x_j is two times likelier to happen than the result x_k (but note that both results may be a lot less likely than the most likely result). Thus, even before starting the experiment, we can ask the following kinds of questions, solely based on the probability distribution:

1. What is the most likely number of heads that we will get? Technically this is the expected value of X, or EX. The expected value of sample space X is the sum of each individual value x_i times the probability p_i of that value being the true value. It is also the value $X = x$, which is the center of mass of the probability distribution.

$$EX = \sum x_i p_i$$

or in R:

```
(EX <- sum(data$x * data$y))
```

```
## [1] 5
```

Interpreting the EX is a thorny problem. Perhaps the most honest way to do this would be to imagine that every individual value, from which the EX was calculated, is in a bag in euro currency. Then, if you pay EX euros to extract a random value, this transaction will most probably allow you to stay even. For most purposes, this does not sound like a scientifically relevant interpretation. Importantly, the EX is not the most likely single value you might draw (this being the mode), nor is the value from which half the values are smaller/larger (this being the median).[28] The EX is more popular than the mode or the median because historically it was often impossible to calculate CI-s and p values from the latter. In modern statistics this is no longer so. Thus you should choose your measure of central tendency/typical value based on scientific preferences.

[28] For symmetrical probability distributions, incl. the normal distribution, the EX = mode = median.

2. What is the standard deviation of the probability distribution? We start by defining variance as $VX = \sum(x_i - EX)^2 p$ where $x_i - E(X)$ is the deviation of i-th value from the expected value of X.

$$SD(X) = \sqrt{VX}$$

```
sqrt(sum((data$x - EX)^2 * data$y))
```

```
## [1] 1.581139
```

The range $EX + / - SD(X)$ usually covers about 2/3 of the values at the center of the distribution. For binomial distribution the variance is $V[X] = Np(1-p)$, and if Np is an integer, then mean = median = mode. Standard error for a proportion p is $\sqrt{\frac{p(1-p)}{N}}$.

3. What is the probability that the true value is greater than/smaller than/lies within an interval of X? If, for example, 2% of the probability density lies beyond (is greater than) some value x_i of X, then we say that the probability that $X > x_i$, or $P(X > x_i) = 0.02$.

To get the probability that we get 7 or more heads out of 10 coin tosses we sum the probabilities for every outcome with 7 or more heads:

```
d1 <- data %>% filter(x >= 7)
sum(d1$y)
```

```
## [1] 0.171875
```

4. What interval of X covers some given fraction of the probability density? For example, if 95% of the probability density lies between values $X = x_j$ and $X = x_k$, then we say that the true value of x lies within the interval x_j and x_k with 95% probability. This is then the 95% credible interval.

When I make 10 tosses, then how many heads will I get with 50% probability?

```
a <- quantile(data$y, prob = 0.5)
data %>% filter(y >= a)
```

```
## # A tibble: 7 x 2
##       x      y
##   <int>  <dbl>
## 1     2 0.0439
## 2     3 0.117
## 3     4 0.205
## 4     5 0.246
## 5     6 0.205
## 6     7 0.117
## 7     8 0.0439
```

So with probability 0.5 I get 2 to 8 heads.

3.4.2 Poisson distribution

If you have a bernoulli sequence, whose number of events (successes) is much smaller than the number of experiments, then the binomial distribution can be converted into Poisson distribution, which only takes in the number events. Poisson distribution has a single parameter λ, which is the expected value (the number of events), and its standard deviation is defined as $\sqrt{\lambda}$. Its formula is

$$f(x) = exp(-\lambda)\frac{\lambda^x}{x!}$$

If we have observed 5 events in a time interval (5 deaths in a week, say), then the probability that next week we will have 3 deaths is

```
exp(-5) * (5^3/factorial(3))
```

```
## [1] 0.1403739
```

or

```
dpois(3, 5)
```

```
## [1] 0.1403739
```

We can plot a Poisson probability distribution like this

```
x <- 0:15
y <- dpois(x, 5)
ggplot(data = NULL, aes(x, y)) + geom_col() + theme_bw()
```

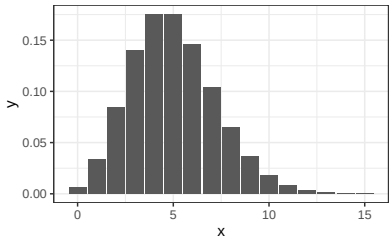


Figure 8: Poisson distribution for lambda = 5

3.4.3. Normal and log-normal distributions

The binomial distribution is the most common probabilistic model for experiments and other phenomena with two possible outcomes. As its assumptions include stability of independent and random events, it is not suitable for all natural mechanisms that generate binary outcomes. So mathematiciens have created a host of other distributions suitable for binary or categorical outcomes, incl. the Piosson distribution, Negative Binomial distribution, the hypergeometric distribution (assumes random draws without replacement, unlike the binomial), etc.

Another common class of outcomes is real numbers, which we call continuous outcomes. The most commonly used probability distribution for modeling continuous outcomes is the normal distribution. For example, if you are running most classical statistical models, like a t test, principal component analysis, correlation, or linear regression, you are probabaly using the normal probability model under the hood. This is so mainly for historical computability reasons - the mathematics of the normal distrtibution is convenient for calculating important statistics with pen and paper. In biology the most useful probability distribution is actually the log-normal distribution.

We start by simulating a process that generates normal/lognormal data.

Lets suppose that we have 12 separate genes influencing a continuous trait (i.e. growth rate). The effects of these genes are independent of each other (The allele status of gene 1 does not have any influence on the effect of allele status of gene 2 on growth rate, and so on). Then the effects are additive. Each realization of an organisms growth rate comes from a random mixture of these genes (the effects of individual alleles are randomly picked from 1 to 2)

A single growth rate:

```
# runif(12, 1, 100) generates 12 random numbers between 1 and 100
sum(runif(12, 1, 100))
```

```
## [1] 687.814
```

What happens, if we simulate 10,000 growth rates?

```
growth <- replicate(10000, sum(runif(n = 12, min = 1, max = 2)))
ggplot(data = NULL, aes(growth)) + geom_density() + theme_bw()
```

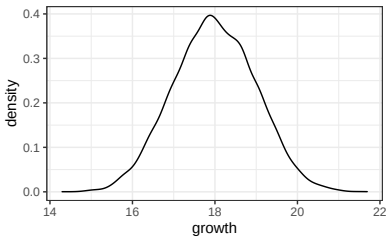


Figure 9: independent effects result in normal data

This is close enough to a normal distribution. Lets model these data through an actual normal distribution. The parameter, which determines the exact shape of a binomial distribution, is probability of success (p). Similarly, the exact shape of the normal distribution is determined by two parameters, the μ, which determines the location of the peak, and the σ, which determines the scale of the distribution (roughly speaking, the width of the thing).[29] We calculate these parameters directly from the simulated data, as is done by most classical statistical techniques. By fixing the values of μ and σ we have *fitted* the model. We can see that the modeled distribution is very close to data, which means that it is now easy to simulate new data from the normal model and, generally, to study the model as a stand-in for the physical world.

[29] For the normal distribution, but not for many other distributions with location/scale parametrization, $\mu = EX = mean$ and $\sigma = SD(x)$.

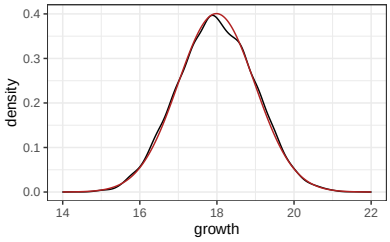


Figure 10: Superimposed data distribution of simulated growth rates and growth rates modeled by normal distribution (red line)

What, if we do not assume independence of effects? Then we assume that the status of gene 1 has an effect on how much the statuses of genes 2, 3, ... 12 affect the outcome (growth rate). The the simplest model of interacting effects involves taking their product. And this results in lognormal distribution of growth rates.

The thing with log-normally distributed data is that by taking the logarithm of such data (or by plotting them in log scale) we get normally distributed data.

The best test of whether your data is lognormal is to take the logarithm of these data. In the log-scale these data should look normal.

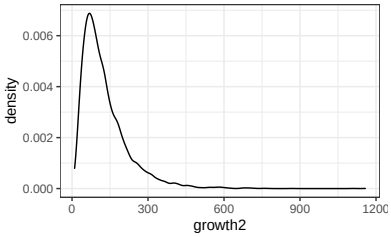


Figure 11: Interacting effects result in log-normal data

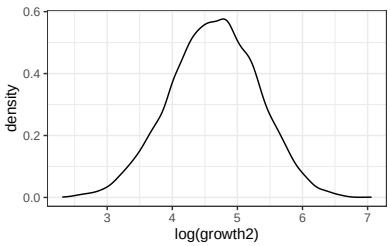
If we think that our data comes from a process that generates approximately normal data (statistically speaking, we are talking about the population, not the sample), then we should model this by normal distribution. The full normal model is described by the function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(-(\frac{(x-\mu)^2}{2\sigma^2})$$

where x and μ are real numbers between -Inf and Inf, and $\sigma > 0$.

If we set mu to zero and sigma to one (this parametrization is called the *standard normal distribution*), we can calculate it like this:

```
x <- seq(-10, 10, length.out = 1000)
mu <- 0
sigma <- 1
y <- (1/(sqrt(2 * pi) * sigma)) * exp(-((x - mu)^2/(2 * sigma^2)))
ggplot(data = NULL, aes(x = x, y = y)) + geom_line() + theme_bw()
```



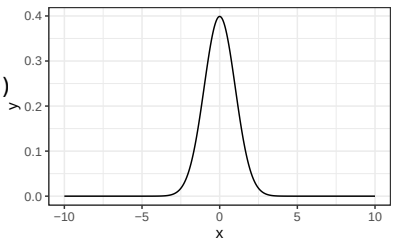


Figure 12: Standard normal distribution

In R we have this function as dnorm()

```
y <- dnorm(x = x, mean = 0, sd = 1)
ggplot(data = NULL, aes(x = x, y = y)) + geom_line() + theme_bw()
```

function rnorm() will draw random numbers from a normal distribution

```
rnorm(3, mean = 100, sd = 10)
```

```
## [1]  87.36132  78.76522 105.60800
```

function qnorm() will show quantiles of normal distribution. For example, if p is a vector of probabilities, then

```
qnorm(p = c(0.05, 0.95), mean = 100, sd = 10)
```

```
## [1]  83.55146 116.44854
```

tells us that if we draw random numbers from this distribution, then we can expect 5% of these numbers to be smaller than 83.55 and 95% to be smaller than 116.45. Thus 83.55 is the 5th percentile (quantile) and 116.45 is the 95th percentile (quantile) of this distribution. This also tells us that 90% of the datapoints (probability density) will be in the range: 83.55 to 116.45.

function pnorm() gives the probability of obtaining a value $< q$ from a normal distribution.

```
pnorm(q = 83, mean = 100, sd = 10)
```

```
## [1] 0.04456546
```

So the probability of obtaining a sample value less than 83 is about 4%.

and the probability of obtaining a value between 85 and 100 is

```
pnorm(q = 100, mean = 100, sd = 10) - pnorm(q = 85, mean = 100, sd = 10)
```

```
## [1] 0.4331928
```

Mathematically the lognormal model is described by this function:

$$f(x) = \frac{1}{x\beta\sqrt{2\pi}} exp(-\frac{1}{2}(log(x/\alpha)/\beta)^2)$$

kus $x > 0$

$$E(X) = \alpha \times exp(\beta^2/2)$$

Again, here we have a model with two parameters, but the EX and SD(X) no longer equal a model parameter.

$$SD(X) = \alpha^2 \times exp(\beta^2) \times (exp(\beta^2) - 1)$$

```
x <- seq(0, 100, length.out = 1000)
y <- dlnorm(x, meanlog = log(100), sdlog = log(10))
ggplot(data = NULL, aes(x = x, y = y)) + geom_line() + theme_bw()
```

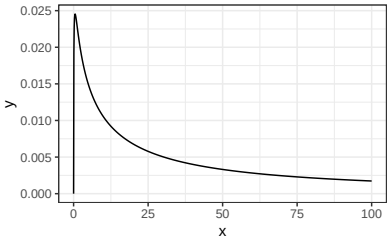


Figure 13: the log-normal model

Note that the R function `dlnorm()` takes its parameters in logarithmic form.

```
qlnorm(p = c(0.05, 0.95), meanlog = log(100), sdlog = log(10))
```

```
## [1]    2.265408 4414.216471
```

```
plnorm(q = c(2.265, 4414), meanlog = log(100), sdlog = log(10))
```

```
## [1] 0.04999194 0.94999780
```

A probability distribution f(x) shows for each value of X, how likely it is to get this value, in relation to other values of X.

For continuous X we talk about probability density (functions), for discrete X we talk about probability mass (functions).

Not all probability distributions have defined expected value and/or SD. They may have other properties that make them usable models.

3.5. A philosophical aside

A probability distribution can arise in two ways. We just used the binomial distribution as a model for a coin tossing experiment. We assumed that we know how many times we plan toss the coin (N = 10) and, more importantly, that the coin is fair (prob = 0.5). Thus we fully understand the experiment and its results, albeit probabilistic, are also fully predictable. This is the deductive use of probability theory that, while making the implications of the coin tossing game excplicit, generates no new knowledge, which is not already present in the rules of the game. If we new the rules of the game, then general properties of the data that might happen are already implicitly present in those rules. We cannot say exactly, which data we will see, but we know exactly the probability of getting a head, from which we deduced the probability of getting k heads (or "k or more heads", etc.) out of n tosses.

In classical philosophy there are two ways ways of understanding the world. One is knowldege (*episteme*), which is by definition true. Another is belief (*doxa*), which comes to us in degrees. Knowledge is the domain of mathematicians and philosophers, while belief is what everybody else has to survive on. Knowledge we can get deductively from axioms and beliefs we get inductively from the phyisical world. In the coin tossing simulation experiment, where we knew the data

generating system, we calculated the probability distribution to get knowledge. But in the real world we never get to have knowledge in this sense, which means that we need to use probabilities differently.

This fundamentally different belief-generating usage of probability theory uses the results of N experiments (data) to guess the rules of the game. In particular, we would like to estimate the probability of getting heads.[30] This means creating new knowledge which is not implicit in data. As we cannot create knowledge in this way, we have to make do with creating new beliefs. These new beliefs still come in the form of a probability distribution. To make this inductive leap we must go from the probability $P(data|rules)$ to $P(rules|data)$. For this we use the Bayes theorem, which we will explain in the next chapter.

[30] Of course, we could ask for the probability that the coin is fair, but we already know this probability to be zero, as no coin is completely fair.

3.6. Probability theory and the Bayes theorem

Words: proposition, conditional probability, prior probability, principle of total information, sample space

We define rationality narrowly as the best way of converting evidence into beliefs, and beliefs into decisions. Probability theory deals with the first part of this definition. Philosophically speaking, one could think of probability theory as a model of rational thinking, rather than the real thing, thus admitting that there might well be more to rationality than is captured by it, while stressing that probability theory does capture an important part of what it means to be rational.

How would you describe the thinking of a narrowly rational robot? What would we want to be the basic rules of reasoning or desiderata of such a robot, on which we could then formalize into a mathematical description of thinking under uncertainty. In other words, can we present a full list of norms of common sense, from which we can derive a rational thinking machine? These rules should be such that a rational person would not want to consciously violate any of them. Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.

The robot reasons about unambiguous Aristotelian propositions, which have truth values (but we don't need to actually know, whether a proposition is true or false, it suffices that it in principle has this truth-property). So it cannot help with much of ordinary human communication, to which unchanging truth values cannot be assigned.

The robot reasons by assigning degrees of plausibilities to propositions, based on evidence given to it. Each piece of new evidence leads to shifting of these plausibilities. Hence the 1st desideratum:

1. Degrees of plausibility are represented by real numbers. We represent a greater plausibility with a greater number.

2. Qualitative correspondence with common sense - if the robot gets information that supports hypothesis A, then the plausibility of A increases, and *vice versa*.

3. Consistency.

3.1. If a conclusion can be reasoned in many ways, then each must lead to the same result.

3.2. The principle of total information – all relevant information must be taken into account when shifting plausibilities. Also note that while causal physical influences propagate only forward in time, logical information works in both directions, so it does not matter when the evidence was collected in relation to the phenomenon it describes.

3.3. Equivalent states of knowledge are presented by equivalent plausibility numbers.

Surprisingly, these rules are enough to ensure a rational brain. Mathematical probability theory is fully consistent with these desiderata and requires no more than them.

Degrees of plausibilities are quantified by probabilities and probability distribution are gives a full description of uncertainty.

> The fundamental principle of all probabilistic inference: to judge the truth of proposition A, calculate the probability that A is true, conditional of all the evidence on hand: $P(A|E_1E_2\dots)$.

3.6.1. The Kolmogorov axioms

Notation:

$P(A)$ probability of A. A stands for proposition, event, hypothesis, parameter value, etc.

$\wedge$ - AND, $P(A \wedge B)$ means "probability that both A and B are true"

$\vee$ - OR, $P(A \vee B)$ means "probability that either A or B, or both A and B are true"

$\neg$ - NOT, $P(\neg A)$ means "probability of not A" or "probability that A is FALSE".

$A|B$ - conditional, $P(A|B)$ means "probability of A given B" or "probability that A is true if we assume that B is true"[31]

[31] B does not have to be actually true, we just assume for the sake of argument that it is.

Prior probability $P(Hypothesis)$ is a purely logical term that incorporates all relevant evidence that is not present in data. The separation of evidence into "data" and "prior" is a matter of convenience of calculation, nothing more.

Likelihood $P(data|Hypothesis)$ models how likely is the sample data for each member of the sample space. If $\Omega = \{H_1, H_2\}$, then the

likelihood space has 2 members as well: $P(data|H_1)$ is the probability of seeing sample data, if it happens that H_1 is true, and $P(data|H_2)$ is the probability of seeing sample data, if it happens that H_2 is true. If the likelihood ratio $LR = \frac{P(data|H_1)}{P(data|H_2)} = 1$, then the data is irrelevant. LR = 10 means that data supports H_1 ten times more than H_2.

LR is also called *evidence*.

The fact that our sample data provides a lot more evidence for H_1 over H_2 does not in itself mean that we should think it more likely that H_1 is true than H_2, or any other logically possible hypothesis. Likelihoods do not sum to one over the sample space. Therefore, likelihood is not a true probability.

Posterior probability $P(Hypothesis|data\ \&\ prior)$ is the product of prior and likelihood over each member of the sample space, normalized over the posterior probability distribution so that probabilities sum to one. Posterior distribution is the output of Bayesian inference. This is the probability distribution that holds our beliefs about uncertainty concerning parameter values.

Axioms:

1. $P(A) \geq 0$
2. $P(\Omega) = 1$[32]
3. $P(A \vee B) = P(A) + P(B)$, if A and B are independent.
4. $P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$[33]

The definition for conditional probability arises naturally from set theory: The number of elements in sample space S is n(S), and S contains two subspaces A and B, which have an overlap (intersection) $A \wedge B$. Calculating $P(A|B)$ is the same as calculating the number of elements contained in the intersection of A and B, relative to the number of elements contained in the whole subspace B: $P(A|B) = \frac{n(A \cap B)}{n(B)}$. By dividing both numerator and denominator with n(S) we convert absolute numbers into fractions, a.k.a. probabilities:

$$P(A|B) = \frac{n(A \cap B)/n(S)}{n(B)/n(S)} = \frac{P(A \cap B)}{P(B)}$$

3.6.2. Some derivations

5. $0 \leq P(A) \leq 1$ - probabilities are bounded by 0 and 1.
6. $P(\neg A) = 1 - P(A)$

[32] The probability of a sure thing is one, which means that the probabilities under a probability distribution sum to one.

[33] This is the definition of conditional probability. If you find it unintuitive, try this: in the urn there are 3 spheres and 3 cubes. Two of the spheres are blue and one is red. What is the probability that a random object drawn from the box is a red sphere? Solution: $P(sphere) = 1/2$; $P(red \mid sphere) = 1/3$; $P(sphere \wedge red) = P(sphere)P(red\ |sphere) = 1/2 \times 1/3 = 1/6$.

7. Monotonicity: if B is a subset of A ($B \subset A$), then $P(B) \leq P(A)$. If B can be deduced from A: $[A \vee (B \wedge \neg A) \Leftrightarrow B]$, then $P(B) \leq P(A)$.

8. $P(A \wedge B) \leq P(A)$; $P(A) \leq P(A \vee B)$

9. If B can be deduced from A and $P(A) > 0$, then $P(B \mid A) = 1$ and $P(\neg B \mid A) = 0$.[34]

10. Logically equivalent propositions/hypotheses have the same probability: if $A \Leftrightarrow B$, then $P(A) = P(B)$

11. Def: A and B are independent (not correlated) if and only if $P(A \mid B) = P(A)$

12. If A and B are independent, then

$$P(A \wedge B) = P(A)P(B)$$

13. If A and B are not independent, then

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

and for 3 events:

$$P(A \vee B \vee C) = P(A) + P(B) + P(C) - P(A \wedge B) - P(B \wedge C) - P(A \wedge C) + P(A \wedge B \wedge C)$$

[34] If evidence e can be duduced from hypothesis H (H predicts e) and if $P(H) > 0$ and $P(e) < 1$, then $P(H \mid e) > P(H)$, or e increases the probability of H.

3.6.3. The Bayes theorem

We can derive the Bayes theorem straight from the definition of conditional probability (4.)

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

,

and equivalently

$$P(B|A) = \frac{P(B \wedge A)}{P(A)}$$

.

As

$$P(A \wedge B) = P(B \wedge A)$$

$$P(A|B)P(B) = P(B|A)P(A)$$

and

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

This is the Bayes Theorem. Now, lets substitute A with Hypothesis H and B with data d.

$$P(H|d) = \frac{P(H)P(d|H)}{P(d)}$$

Here $P(H \mid d)$ is the posterior, $P(H)$ is the prior probability of the hypothesis, $P(d \mid H)$ is the likelihood, or the probability of data given the hypothesis, and $P(d)$ is the total probability of the data, summed over all the hypotheses in the sample space. P(d) is the normalization member, which guarantees that the posterior probabilities sum to one over the sample space.

The single most important thing to remember is that posterior is proportional to the product of prior and likelihood.

What does $P(d)$ really mean? If $\Omega = \{H, \neg H\}$, then

$$P(d) = P(d \wedge H) + P(d \wedge \neg H) = P(H)P(d|H) + P(\neg H)P(d|\neg H)$$

where the final step comes from the definition of conditional probability.

Example (with real US data from the spring of 2020): We have an antibody test for COVID-19. If 1000 covid-19 positive patients are tested, on average 840 of them get a true-positive result (sensitivity of the test is 84%). If 1000 non-infected people are tested, 5 get a false-positive result (specificity of the test is 1 - 0.005 = 0.995, or 99.5%). We think that about 1% of the population has been exposed to the virus. Now, if a random person gets a positive result, what is the probability that this person is infected?

- hypothesis space has 2 members: H_1 - infected, H_2 - not infected
- priors: $P(H_1) = 0.01$, $P(H_2) = 1 - P(H_1) = 0.99$
- likelihoods $P(+|H_1) = 0.8$, $P(+|H_2) = 0.05$

```
Pr_H1 = 0.01
Pr_H2 = 1 - Pr_H1
L_H1 = 0.84
L_H2 = 0.005
(Pr_infected = (Pr_H1 * L_H1)/(Pr_H1 * L_H1 + Pr_H2 * L_H2))
```

```
## [1] 0.6292135
```

There is a more intuitive solution. Imagine 1000 random persons, 10 of whom are infected and 990 are not. Now from the infected 10

x 0.84 = 8.4 test positive and from the uninfected 990 x 0.005 = 5 test positive. The probability of being infected, if getting a positive test result, is then 8.4/(8.4+5) = 0.63.

NYT 24.08.2020:

> Getting an antibody test to see if you had Covid-19 months ago is pointless. Many tests are inaccurate, some look for the wrong antibodies and even the right antibodies fade away, said the Infectious Diseases Society of America, which issued the new guidelines. Antibody testing generally should be used only for population surveys, not for diagnosing illness in individuals, the panel said. Even for that purpose, only tests that are correctly positive more than 96 percent of time and correctly negative at least 99.5 percent of the time should be used, according to the guidelines. Very few of the dozens of tests that the panel looked at met those standards. None can be done at home or immediately in doctors offices, and the best are assays known as Elisa or chemiluminescence immunoassay. With two exceptions, antibody tests should not be used to diagnose individual infections, the society said. When a patient has all of the symptoms of Covid-19, including X-ray evidence of pneumonia, but still comes up negative on repeated diagnostic PCR tests for the virus, an antibody test may be useful. The tests can also be used for diagnosis when a doctor suspects a child has multisystem inflammatory syndrome, a rare but serious complication of Covid-19 in children. Because it is not known how long after the initial infection this inflammation begins, doctors should do both a PCR test and an antibody test, the guidelines said.

What if we want to estimate the true value of a continuous parameter (for example the probability that a patient dies of COVID_19, or the mean length of hospitalization with COVID-19)? In such cases the hypothesis space consists of infinite number of elements (real numbers from 0 to 1, and integers from 0 to Inf, respectively). This means that we must specify an infinite number of likelihoods and an infinite number of prior probabilities, before we can run the calculation! Luckily we can do this using continuous functions, which by definition specify infinite number of points. Then the posterior will be a continuous function as well.

More generally, for a sample space with n members (n mutually exclusive and exhaustive hypotheses) $\Omega = \{H_1, H_2, ...H_n\}$ and

$$P(d) = P(H_1)P(d|H_1) + P(H_2)P(d|H_2) + ... + P(H_n)P(d|H_n)$$

This is the formula for total, a.k.a. marginal, probability.
So the Bayes theorem becomes

$$P(H_1|d) = \frac{P(H_1)P(d|H_1)}{P(H_1)P(d|H_1) + P(H_2)P(d|H_2) + ... + P(H_n)P(d|H_n)}$$

Now we have a formula that allows us, in logically the best way, to combine evidence that is present in sample data (as modeled in the likelihood) with data-independent beliefs (as modeled in the prior), and end up with a posterior probability of hypothesis, given data and prior. In other words, we can use the Bayes theorem to combine disparate strands of information into a single posterior that (usually) comes in the form of probability distribution. There is no magic involved, just basic probability theory. The Bayes theorem is not a truth machine – it just combines information that we put into it. If the information in the prior or in the likelihood is grossly wrong, there is no guarantee that even in the long run the posterior converges on truth (although under most realistic scenarios it does just that). Importantly, Bayes theorem works iteratively and there is no lower limit of data that you can put into it: N=1 is fine.[35]. Also, it is easy to incorporate new data, as you can simply redefine your posterior from the previous iteration as the new prior, add new data as likelihood, and calculate the new posterior, which before long will become a prior. Thus Bayesian inference recapitulates the piecemeal nature of scientific inference. Importantly, unlike with frequentist statistics, there are no stopping rules, meaning that you can safely calculate the posterior after receiving each new datum.[36] In addition, as Bayesians calculate the probabilities of hypotheses/parameter values directly, the have no need for the sampling distribution of data, which lets them off the hook with the random sample. Bayesians look at the exact data, not some imaginary random sample from an infinite statistical population. As frequentist statistics needs its samples to be independent and identically distributed random variables (*iid*, also see point 11.), for Bayesian analysis the data need less restrictively to be merely exchangeable, meaning that the order by which individual datums are put through the analysis should not have any effect on the results.

[35] Frequentist statistics does not work like this. There you need to have some minimum N for the algorithms to work as advertised.

[36] In frequentist statistics you need stopping rules (predetermined sample size), so you would not simply call a "significant effect" as soon as the randomly fluctuating p value first drops below the significance level. In Bayesian statistics there is no significance level and no dichotomization of continuous evidence, so stopping rules would be meaningless.

Bayesian inference comes straight from probability theory, while frequentist inference is a collection of ad hoc procedures, which require more stringent conditions to work properly.

4. *The basics of Bayesian inference*

In frequentist statistics the main workflow is something like this:

1. Decide on the stopping rules and statistical tests that will be used, write a pre-registered analysis protocol.
2. Do a power analysis to determine optimal sample size N for finding significant effects
3. Collect the sample of size N (no more, no less). There can be no peeking at data, unless pre-specified in step 1, until the full sample is collected. The sample needs to be analysed together. Ideally the experiment is randomized and blinded, so the researcher does not know, which is the experimental group and which is the control group.
4. Run the pre-specified statistical test on pre-specified variables, which have been normalized and transformed according to the pre-specified protocol.
5. Check the test assumptions (normality, linearity, constant variance, etc.) by running additional tests on your data and on the fitted model.
6. If these assumptions are satisfied, interpret the p value as significant or not.

In general, a frequentist analysis should be completely pre-registered to avoid multiple testing problems that arise largely (but not solely) because of dichotomizing continuous evidence by the statistical significance criterion.[37] This avoidance of "researcher degrees of freedom", which can destroy the long-run quality control properties of frequentist tests, makes the analysis very rigid, and often equally frigid, or incapable of reacting to different-than-expected results. It is a serious mistake to suppose that frequentist requirements to statistical inference, which come from the need to fix the long-run frequency of type 1 errors, reflect how real scientists do real scientific inference.

[37] While multiple testing, narrowly understood, is not a problem for the Bayesian, false alarms are. The currently favored Bayesian mitigation of false alarms is multilevel modeling.

In Bayesian statistics we do not care about stopping rules and instead of a power analysis we might simulate realistic-looking data and try to analyse these, to see whether the models and algorithms to fit them work well enough for these kind of data. This will also double as a power analysis as one can find an optimal sample size that gives sensible estimates. Also, we do not pre-specify, which models we will use on our data, as we will build several bespoke models depending on the data that we eventually get, so we can compare the models and their fits, and try to learn from them as a group. Instead of formally checking model assumptions, we rather generate simulated data from the fitted models and compare with actual data, to see, which aspects of our data are well represented in the model structure, and which ones are not. This approach is a major benefit of the Bayesian approach. As a whole, Bayesian modeling is much more flexible, both in terms of model structures and in how we work with the models, in that it tries to mirror the scientific process, where we gradually learn from data, and from our mistakes, to build better experiments and better ways of analysing their results.

A simplified Bayesian procedure:

1. model data as likelihood, by using an appropriate probability distribution that reflects the data generating mechanism.
2. model additional information as prior, by using a probability distribution that best allows to describe your prior beliefs.
3. use a Bayesian procedure to get a posterior. These days, instead of trying to calculate from the Bayes theorem directly, which can be overwhelming computationally, various MCMC approximations are used.
4. compare the model fits, and the posteriors that we got from different models
5. As needed, modify your models to achieve better fit and more realistic model-generated data. Then repeat the cycle, until you are satisfied.
6. interpret the posteriors in scientific context, make counterfactual predictions from the fitted model.

It should be noted that at the end of the day both Bayesian and frequentist statisticians want the same thing. They want to use the fits of statistical models in formulating scientific conclusions. The numerical outputs of statistical models themselves are not it. They have to be interpreted in non-mathematical and out-of-the-model contexts. However, by artifice and by tradition the two great traditions of statistics go about their business in quite different ways.

In frequentist statistics it is usually the raw effect sizes and the p values that get interpreted. In Bayesian statistics, they are the shrunk effect sizes (we will learn soon, what this means) and the assorted credible intervals. In frequentist inference usually a single model is interpreted per effect, while in Bayesian inference often several models are interpreted together. In frequentist inference the statistically non-significant effects are often not interpreted at all (this is not what frequentist theory proposes, but rather an unfortunate historical accident), in Bayesian inference all effects are interpreted together (remember that the *law of total information* is a law of logic; you do not want to go around, breaking these!).

4.1. Fitting the Binomial data model

Mortality of a disease is 50% and we have 3 patients. How many of them are likely to die? We have 2 pieces of data (mortality rate p = 0.5 and N = 3) and a 4-member sample space (0 dead, 1 dead, 2 dead, and 3 dead). We start by enumerating all logically possible scenarios (d - dead, a - alive)

d d d
d a d
d d a
a d a
a a d
a a a
a d d
d a a

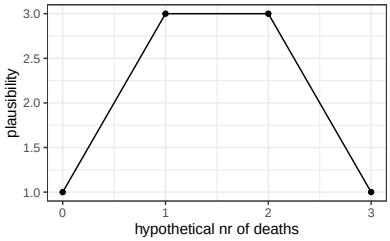
Kui P(dead) = 0.5, then by simple counting we get that:

- 0 dead - 1,
- 1 dead - 3,
- 2 dead - 3,
- 3 dead - 1

Thus we have 1 + 3 + 3 + 1 = 8 possible futures that divide between the 4 hypotheses. This means that for each member of the parameter space we know the probability of its realization (P(1 death) = 3/8, etc.). It is this knowledge we now turn into a likelihood function.

```
# Parameter space: all possible futures
x <- seq(from = 0, to = 3)
```

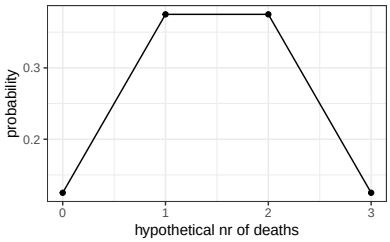
```
# Likelihoods for each x value, or P(deaths I x)
y <- c(1, 3, 3, 1)

ggplot(data = NULL, aes(x, y)) +
geom_point() + geom_line() +
xlab("hypothetical nr of deaths") +
ylab("plausibility") + theme_bw()
```



One death and two deaths are equilikely and one death is three times as likely as no deaths (or three deaths). Likelihood simply tells for each number of deaths, how likely is this outcome, given the mortality figure that we state.

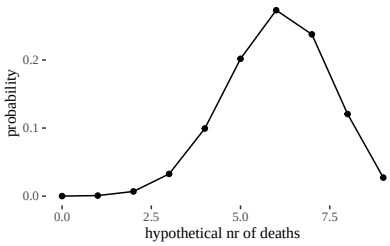
Now we demonstrate the same using the binomial distribution model. The only difference is that now we get on the y-axis normalized probabilities.

```
y <- dbinom(x, 3, 0.5)

ggplot(data = NULL, aes(x, y)) +
geom_point() + geom_line() +
xlab("hypothetical nr of deaths") +
ylab("probability") + theme_bw()
```



How many dead when we have 9 patients and mortality rate is 67 percent?

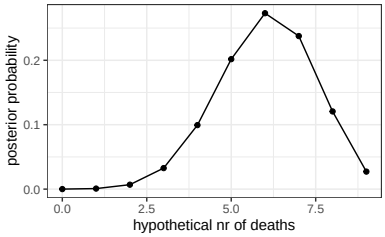
```
x <- seq(from = 0, to = 9)
y <- dbinom(x, 9, 0.67)

ggplot(data = NULL, aes(x, y)) +
geom_point() +
geom_line() +
xlab("hypothetical nr of deaths") +
ylab("probability") +
ggthemes::theme_tufte()
```



Next we add to the likelihood a flat prior and use the Bayes theorem.

```
x <- seq(from = 0, to = 9)  #parameter space: nr of deaths
prior <- rep(1, 10)  # flat prior

# Compute likelihood at each value in grid
likelihood <- dbinom(x, size = 9, prob = 0.67)

# Compute product of likelihood and prior
```

```
unstd.posterior <- likelihood * prior

# Normalize the posterior, so that it sums to 1
posterior <- unstd.posterior/sum(unstd.posterior)

ggplot(data = NULL, aes(x, posterior)) +
geom_point() + geom_line() +
xlab("hypothetical nr of deaths") +
ylab("posterior probability") + theme_bw()
```



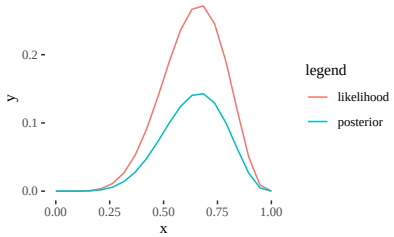
Now its time for a shift of perspective. We want to estimate the true mortality rate.

- data: 6 dead out of 9 patients.
- parameter to estimate: the mortality rate (p).
- sample space: real numbers between 0 and 1.

We can still use binomial likelihood. dbinom() has 3 arguments, nr of events, nr of tries and probability of events. Again, we have two of them as data and one as the parameter whose value we want to estimate. We use the flat prior again. As there are Inf number of members in Ω, we cheat a little and calculate for the grid of 20 evenly spaced parameter values.

```
# grid: mortality at 20 evenly spaced probabilities from 0 to 1
x <- seq(from = 0 , to = 1, length.out = 20)

# prior
prior <- rep(1 , 20)

# likelihood at each value in grid
likelihood <- dbinom(6, size = 9 , prob = x)

posterior <- likelihood * prior / sum(likelihood * prior)

a <- tibble(x=rep(x=x, 2),

            y= c(likelihood, posterior),

            legend= rep(c("likelihood", "posterior"), each=20))

ggplot(data=a) +

  geom_line(aes(x, y, color=legend))+
```

```
  ggthemes::theme_tufte()
```

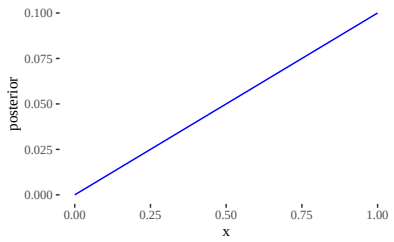


Our posterior sums to 1, but thanks to flat prior its shape is the same as the likelihood.

Lets do this datapoint-by datapoint, so that n = 1.

First datapoint is that the 1st patient died.

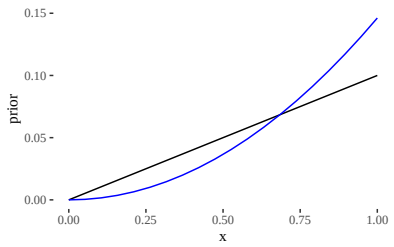
```
likelihood <- dbinom(1, size = 1, prob = x)
posterior <- likelihood * prior/sum(likelihood * prior)
ggplot(data = NULL) +
geom_line(aes(x, posterior), color = "blue") +
ggthemes::theme_tufte()
```



Zero mortality is now logically impossible and 100 percent mortality is the hypothesis best supported by the data.

Second datapoint is that the 2nd patient died. Our previous posterior is now the prior.

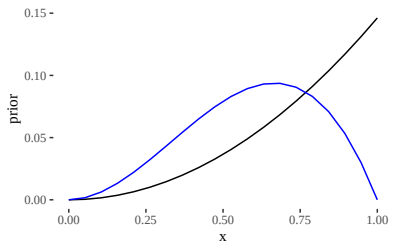
```
prior <- posterior
likelihood <- dbinom(1, size = 1, prob = x)
posterior1 <- likelihood * prior/sum(likelihood * prior)
ggplot(data = NULL) +
geom_line(aes(x, prior)) +
geom_line(aes(x, posterior1), color = "blue") +
ggthemes::theme_tufte()
```



The posterior is no longer a straight line. 100 percent mortality is still the most likely mortality estimate.

Third datapoint is that the 3d patient survived.

```
prior <- posterior1
likelihood <- dbinom(0, size = 1, prob = x)
posterior2 <- likelihood * prior/sum(likelihood * prior)

ggplot(data = NULL) +
geom_line(aes(x, prior)) +
geom_line(aes(x, posterior2), color = "blue") +
ggthemes::theme_tufte()
```



Now zero mortality and 100 percent mortality are both logically impossible. The most likely value for mortality is 2/3 or 75 percent.

4.2. Normal data model

Richard McElreath:

> There are an infinite number of possible Gaussian distributions. Some have small means. Others have large means. Some are wide, with a large σ. Others are narrow. We want our Bayesian machine to consider every possible distribution, each defined by a combination of μ and σ, and rank them by posterior plausibility. Posterior plausibility provides a measure of the logical compatibility of each possible distribution with the data and model. ... The parameters to be estimated are both μ and σ, so we need a prior $Pr(\mu, \sigma)$, the joint prior probability for all parameters. In most cases, priors are specified independently for each parameter, which amounts to assuming $Pr(\mu, \sigma) = Pr(\mu)Pr(\sigma)$.

How does one make a US president? Lets start by generating some heights of presidential candidates. We assume that human growth is a deterministic process that depends on a complex mix of many genetic and environmental factors. But we really do not understand this process well enough to directly use it to generate prospective heights of future presidents. We think that human height is approximately normally distributed (conditional on sex, age, etc.), so we could model presidential heights as a normal distribution where each possible height is given a probability of occurring. Now, we could generate new heights from this stochastic model proportionally to these probabilities. But before we do this, we have to fix the shape of the normal distribution so that we know the relevant probabilities. This can be done by fixing the values of two parameters of the normal distribution, μ and σ. Luckily we have data on the heights of the last 12 presidents:

```
heights <- tibble(value = c(183, 192, 182, 183, 177, 185, 188, 188, 182, 185, 188,
    182))
```

As the parameter μ is identical to the expected value, i.e. the mean of the normal distribution and σ to standard deviation (SD), we can simply calculate those and then use these numbers to generate new heights.

Data mean is:

```
(Mean <- mean(heights$value))
```

```
## [1] 184.5833
```

and SD is:

```
(SD <- sd(heights$value))
```

```
## [1] 3.964807
```

Predict 4000 new heights:

```
set.seed(1)
rnorm(4000, Mean, SD) %>% sd()
```

```
## [1] 4.10711
```

This procedure is not ideal, though, as it does not take into account the fact that our estimates of μ and σ, calculated from limited data, are not precisely true for the full population of US presidents, including the future ones. So on top of the randomness in predicting new presidents that comes from our imperfect knowledge of the mechanisms that generate human heights we must model the uncertainty that comes from imperfect data. How can we do this? When modelling this 2nd order epistemic uncertainty we will to end up not with point estimates of μ and σ, but probability distributions for μ and σ values. This way we are not predicting new heights from a single fixed normal distribution, but rather from many different distributions, where each is parametrized by a plausible combination of μ and σ values (each of these combinations is supported by the data and by our prior knowledge of plausible presidential heights).

```
pres_m <- brm(data = heights, formula = value ~ 1, file = "models/pres_m")
pres_posterior_sample <- posterior_samples(pres_m)
set.seed(1)
rnorm(4000, pres_posterior_sample$b_Intercept, pres_posterior_sample$sigma) %>% sd()
```

```
## [1] 4.546833
```

Now the standard deviation of predicted presidents heights is slightly larger (by 0.44 cm), reflecting the extra uncertainty. Note that the model pools both kinds of uncertainties so that we cannot re-separate them in later analysis.

Next we fit this model the old-fashioned way, using the Bayes theorem with grid approximation.

Basic Bayesian inference goes like this:

1) What is the sample space of true mean heights? From -Inf to Inf. Silly, but convenient, if we want to use the normal data model (likelihood).

2) How likely it is for each possible height to be the true average? The answer to this question is the likelihood, which depends on our sample data and on the model we choose to represent this data with. Here we model the likelihood as a normal distribution, whose mean equals the sample mean and whose standard deviation equals the sample sd/sqrt(N), or SEM (standard error of the mean).

3) Next step is quantifying prior beliefs, independent of sample data, about the mean height. Our prior for the mean is also normal distribution with $\mu = 175.4$ (the mean height of US males), sd = 3

(picked so, that a mean height of 181 cm for the presidents would still be fairly possible). And the prior for σ is also normal with $\mu = 5.2$ (the SD for the heights of the general US male population) and sd=1 (so that true SD values between about 3 and 7 cm would be possible enough). This is a rather informative prior.

4) Then we calculate the posterior.[38]

[38] This is Bayesian method in a nutshell: at each parameter value (x) we multiply the likelihood (y_lik) with the prior (y_pr), thus combining our prior knowledge with the data and the data model. Knowledge (prior and data) in – knowledge (posterior) out.

Some preliminary points:

- Likelihood models data as a probability distribution.
- Likelihood does not have to be normal.
- Prior is your belief about the parameter value, modeled as a probability distribution.
- Prior does not have to be normal.
- Likelihood and prior(s) do not have to be represented by the same distribution.
- Posterior is narrower than prior and likelihood. This reflects the combining of information in the likelihood and prior. Bayes theorem is provably the most efficient way of doing this.

Now, lets go on. Next item in the menu is the grid method for posterior in 2D. Instead of multiplying likelihoods, we will add log-likelihoods (which is mathematically the same thing).

The model is:

$$height_i \sim normal(\mu, \sigma)$$

$$\mu \sim normal(175.4, 5)$$

$$\sigma \sim normal(5.2, 1)$$

Where the 1. line is the likelihood, containing two parameters, and the next 2 lines are priors for each parameter.

i is index for heights, which means that the model assumes that each individual height comes from the same normal distribution.

prior predictive simulation

Here we calculate the posterior solely on the prior. As we add no likelihood, the data has no influence on the posterior – only the prior does.

```
mu_prior1 <- rnorm(10000, 175.4, 5)
mu_prior2 <- rnorm(10000, 180, 50)

sigma_prior1 <- rnorm(10000, 5.2, 1)
sigma_prior2 <- runif(10000, 0, 100)

y1 <- rnorm(10000, mu_prior1, sigma_prior1)
y2 <- rnorm(10000, mu_prior2, sigma_prior2)

ggplot(data = NULL, aes(x = y1)) + geom_density() +
geom_density(aes(x = y2), color = "red") +
xlab("mean presidents height") + theme_bw()
```

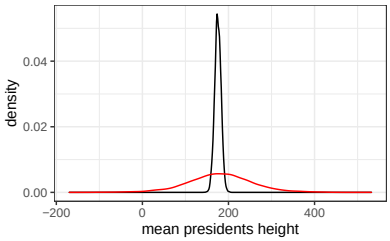


Figure 14: A prior predictive check.

We use two different prior combinations: one is informative (the black line in the figure) and another one is weakly informative (red), meaning that it carries no useful biological information, but nevertheless gives a lot more weight to postive values and values < 4 meters (thats right - this prior assumes that 4-meter mean height is possible). Currently, weak priors are all the rage, because they make model fitting by MCMC methods easier, while requiring minimal scientific input. Lets hope that this situation is about to change in favour of strong scientifically informed priors!

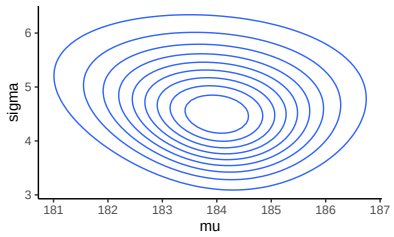
Calculating the 2D posterior

The posterior is a joint distribution over all parameters in the model. If we have 2 parameters, then we have a 2D posterior. If we have 200 parameters, then we have 200D posterior. We can always collapse the posterior to a desired 1D view for inspection (see below).

```
# d_grid contains all possible combinations (10 000) of mu and sigma values.
n <- 100
d_grid <- crossing(mu = seq(from = 180, to = 190, length.out = n),

#crossing() expands the table to cover all possible combinations.

                   sigma = seq(from = 1, to = 8, length.out = n))

#grid_function generates a density value from a normal function,
#whose x is a vector of presidents height and mu and sigma are plucked from d_grid.
#sum() in the log-scale is prod() in the linear scale. In each row of d_grid
#we multiply likelihoods (actually, sum log-likelihoods) for each data point, using
#different combinations of mu and sigma to model the normal distribution.

grid_function <- function(mu, sigma) {
```

```
  dnorm(heights$value, mean = mu, sd = sigma, log = TRUE) %>% sum()
}

d_grid <- d_grid %>%
  mutate(log_likelihood = map2_dbl(mu, sigma, grid_function),
# map2_dbl() takes each mu and sigma pair in d_grid (row-wise) and puts it into the grid_function
         prior_mu    = dnorm(mu, mean = 175.4, sd  = 5, log = TRUE),
#for each value of mu (one in each row) we calculate normal density from prior distr.
         prior_sigma = dnorm(sigma, mean = 5.2, sd = 1, log = TRUE),
#for each value of sigma we calculate normal density from prior distr. for sigma
         product = log_likelihood + prior_mu + prior_sigma,
#sum() in the log-scale is product in the linear scale.
         probability = exp(product - max(product)))
#product - max(product) mirrors into negative scale, which after exp() becomes probability scale.
#probability vector contains the posterior probabilities across mu and sigma.

d_grid %>% ggplot(aes(x = mu, y = sigma, z = probability)) +
  geom_contour() + theme_classic()
```

This 2D posterior is drawn as a 3D surface, where the third dimension is posterior probability. [39] In the center of the contour is the peak of the posterior distribution. This is, by the way, a bivariate normal distribution.

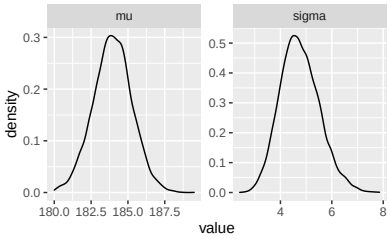


[39] geom_contour() visualizes 3D surfaces in 2D. To specify a valid surface, the data must contain x, y, and z coordinates, and each unique combination of x and y can appear once.

Sampling from the posterior – 1D posteriors for mu and sigma.

I am sampling randomly while weighting on probability, so that mu values with higher posterior probability are more likely to end up in the sample. From such random samples of mu and sigma (the posterior samples) we can plot the posteriors:

```
d_grid_samples <- d_grid %>% sample_n(size = 10000, replace = TRUE, weight = probability)

d_grid_samples %>%
pivot_longer(mu:sigma) %>%
ggplot(aes(x = value)) +
geom_density() +
facet_wrap(~name, scales = "free")
```



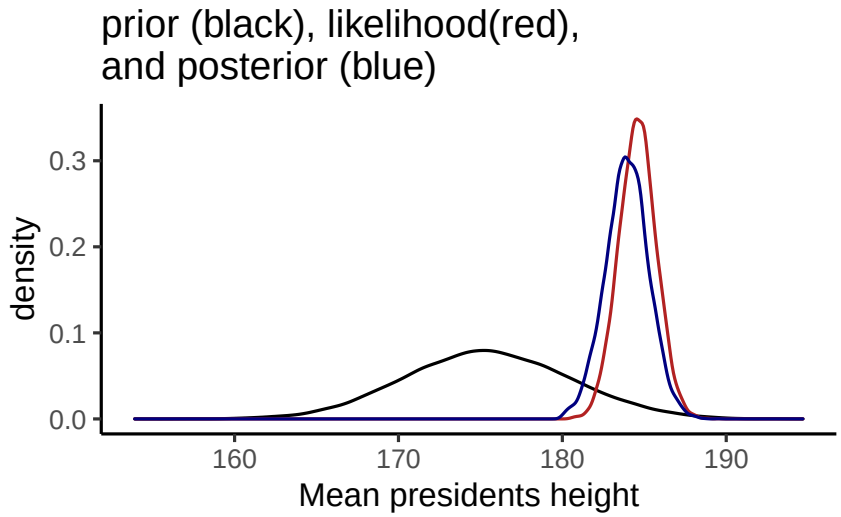
summarising the posterior (95% HDI)

```
posterior_summary(d_grid_samples[, 1:2])
```

```
##         Estimate Est.Error      Q2.5      Q97.5
## mu    183.951485 1.3195671 181.31313 186.565657
## sigma   4.755323 0.7611191   3.40404   6.373737
```

Triplot for mu

```
prior_mu <- rnorm(10000, mean = 175.4, sd = 5)
lik_mu <- rnorm(10000, mean = mean(heights$value), sd = sd(heights$value)/sqrt(nrow(heights)))
post_mu <- d_grid_samples$mu

ggplot(data = NULL, aes(prior_mu)) +
geom_density() +
geom_density(aes(lik_mu), color = "firebrick") +
geom_density(aes(post_mu), color = "navyblue") +
xlab("Mean presidents height") +
ggtitle("prior (black), likelihood(red), \nand posterior (blue)") +
theme_classic()
```



Even with only 11 datapoints the prior is so much wider than likelihood that it has minimal influence on the posterior.

Some things to notice

1) As the priors are true probabilities that sum to one, and the likelihoods are not, we can run our models on either priors and data, or on priors alone, but never on data alone. Running a model on priors alone can be a good idea, if you want to know, how much your data influences the model fit. When you have little data and strong priors, this is the thing to do.

2) Each fitted model with k parameters gives you a k-dimensional posterior (this is slightly more complicated with multi-level models), which we flatten into k side-by-side variables in the posterior samples table. This means that we have to treat each row of this table as a single unit, when we use them in further calculations. In each row there is one plausible combination of k parameter values – if you mix the values between rows, this might no longer hold.[40]

3) As we fitted the full model (both mean and sd), we can generate *in silico* data from it, using the full posterior samples, and compare these generated data with the actual data that was used to fit the model. By this we mean that all Bayesian models are *generative*. This comparison is called a *posterior predictive check* and it is a very convenient method to get a first impression on how well your mathematical model recaptures the physical data generating mechanism in the nature. PP check is a major advantage of Bayesian modeling, as the usual frequentist models that estimate the mean, do not estimate the sd. Such frequentist models are not generative.

4) The more data you have, the narrower the likelihood; the more scientific information you put into the prior, the narrower the prior (usually, but not always). Whichever of the two is narrower, will influence the posterior more. The posterior will always be narrower than likelihood and prior. A lot of data swamps the prior, making it unimportant in respect of the posterior. Conversely, with a small sample the prior can have a major influence on the posterior.

5) In Bayesian estimation, posterior is everything. It contains all of the information about the estimates that is contained in the fitted model. There is nothing in the model about the parameter values that is not covered in the posterior. Every fitted model gives you a posterior for every parameter in the model. This is it. All you have to do, is to study those posteriors, and to make calculations with them to get new estimates. For example, we will soon learn, how to get the estimate for a group mean by adding another group mean and the effect size. These calculations are

[40] With the normal likelihood it will actually hold, because the variation of μ and σ are independent of each other, but this is not so with most other likelihoods.

done at the level of posterior samples, the idea being that you do not want to lose information about estimation uncertainty. Also, with many likelihoods you can calculate straight from the posterior either the mean aor the median (and often the mode), all with appropriate credible intervals. This is not something you can easily do in frequentist statistics.

5 Basic linear regression

Lets start by thinking once more about modeling a normally distributed continuous variable, y, with the normal model:

$$y \sim normal(\mu, \sigma)$$

We will get the mean y ($mean = \mu$) and data level sd ($SD = \sigma$) from this model. But what if we want to model the mean y and/or the sd in different levels of another variable, x?

Suppose that we have measured on each object of measurement its height and we also know its sex. We code the male sex with 0 and female sex with 1, so the sex is now technically a continuous variable, for modeling purposes. We want to model the mean height for sex = 0 and for sex = 1 subjects. The easiest way to do this would be to partition the data by sex and run 2 separate models.

$$h_{sex=0} \sim normal(\mu, \sigma)$$

$$h_{sex=1} \sim normal(\mu, \sigma)$$

We will get separate estimates from separate models, which know nothing of each other, and from these, we can get effect sizes with full posteriors.[41]

[41] posterior_mu_model1 - posterior_mu_model2 gives the posterior for the effect size, i.e. by how much on average are males taller than females.

This is wasteful, however, because we could use the information on females heights on males heights, and *vice versa*. The model, by which we do this, is called *linear regression*. We regress y on x. The regression model differs from the simple $y \sim normal()$ model in that as the simple model estimates the unconditional (a.k.a. marginal) probability of P(y), the regression model estimates a conditional probability $P(y|X)$, where X stands for one or more predictor variables. We estimate the mean value of y given pre-specified exact values of X.

In principle, we should strive for big models that use all our data. However, for practical reasons, we can also divide the data and run

separate models. Statistics is a practical thing. If a full model is too hard to specify, to understand, or to fit, use several smaller models instead!

A regression model can be built in different ways. For instance, we could say to the model that (i) on different levels of sex, heights are allowed to have different mean values, but they must have identical sd-s ; or (ii) that heights have to have identical means between sexes, but may have different sd-s; or (iii) that both means and sd-s are allowed to vary by sex. The most common model, by far, would be (i), so lets start from that one.

- What we model is the y-variable (height).[42]
- What we model y on is the x-variable (sex).[43]
- The Model consists of two parts, the deterministic part and the stochastic part: (i) the process model shows the precise deterministic relationship between a mean y value and a precise x value and (ii) the likelihood shows, how we model stochastically the variation in the y-variable (but not in the x-variable). Linear models can only have a choice between a normal and a students t likelihood. Other likelihoods, like binomial, Poisson, lognormal, etc., define GLMs (Generalized Linear Models), of which more in later chapters. A linear model is also linear in the process model, defined as any equation that can be reformulated as an addition (for example, if X = 2 and Z = 3, then 2 x 3 = 6 can be reformulated as 2 + 2 + 2 = 6, so $y = xz$ would be a linear process model, as would be $y = x + x^2 + x^3$).

[42] Also known as dependent variable, predicted variable.

[43] Also known as independent variable, predictor.

The linear process model for the mean (remember, we want the mean to vary, but the sd to stay put in different levels of x) is then $\mu = \alpha + \beta x$, where α is the intercept and β is the slope. This is nothing more complicated than linear equation, which allows to place a straight line through 2D Cartesian space. It binds the x-variable (sex) into μ, but in doing so, we have redefined the parameter μ through two new parameters: intercept (α) and slope (β).[44] So, instead of μ, we will be fitting these two.

[44] It is also common to denote intercept as β_0 and slope as β_1.

$$y \sim normal(\mu, \sigma)$$

$$\mu = \alpha + \beta x$$

or, equivalently, $y \sim normal(\alpha + \beta x, \sigma)$

Either way, we now need three priors, one each for a, b, and sigma.

$$\alpha \sim normal(,)$$

$$\beta \sim normal(,)$$

$$\sigma \sim student_t(,,)$$

Next we do a simple simulation of two groups, each N = 100 and sd = 50, mu1 = 100 and mu2 = 110. The plotting of these data shows a moderate difference of the means, relative to sd.

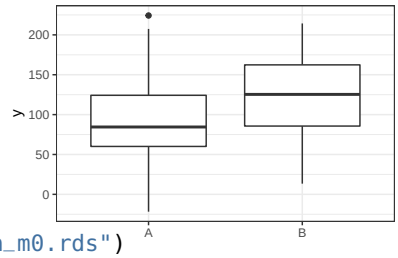


Figure 15: Two-goup simulated data.

We fit the model in brms:

```
simple_lin_m0 <- brm(y ~ x, data = df1, file = "models/simple_lin_m0.rds")
```

```
##             Estimate Est.Error     Q2.5     Q97.5
## b_Intercept 89.84760  7.200801 75.47945 104.28763
## b_xB        33.00253 10.433655 11.96027  53.69062
## sigma       51.31485  3.772500 44.65745  59.51646
```

- Intercept denotes α, which in our case is the mean for group 1 (where x = 0)
- xB denotes β, which is the effect size, a.k.a. difference between means of group 1 and group 2.

The mean of group 2 is then $\alpha + \beta$ and the β coefficient is our estimate for the difference between the groups.[45]

[45] You can do this by simply adding the posterior samples of the two coefficients.

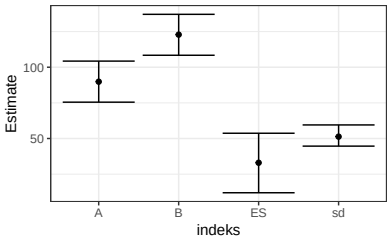


Figure 16: A plot of our estimates for the mean of each groups (A and B), the effect size, and the individual-level variation

So far we used a categorical x-variable or predictor, sex, which we recoded as a continuos variable from which we have 2 values (0 and 1). We did this solely in the interest of the mathematics of line fitting. Thanks to this trick we could ask, what is the height of someone, whose sex = 500? There is no problem getting an estimate the from the fitted model. All we have to do is to substitute the fitted values into intercept and slope coefficients. We will use the full posterior samples, so we do the calculation 4000 times (this is how many posterior samples we have) with slightly different estimates for intercept and slopes. As our Bayesian model considers all of these pairs plausible, our best option is to calcuate them all and plot the full posterior as an histogram. This posterior gives us the full information on uncertainty around the prediction.

```
y_500 <- post$b_Intercept + post$b_xB * 500
```

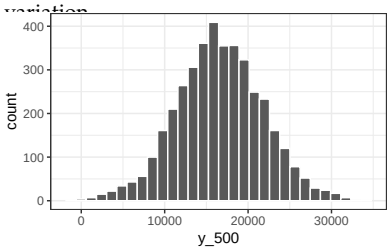


Figure 17: Posterior for an absurd prediction from linear model.

So the posterior depicted in the histogram tells us that the predicted height for a subject, whose sex = 500, is about 200-fold larger than the height of males and females. This is a fine prediction matemathically, but in the world, it makes no sense in more than one level.

However, if our predictor really has more meaningful levels than 0 and 1, then this is exactly how we fit and interpret the linear regression. The meaning of Intercept and slope remains exactly the same:

- The intercept is the estimated mean value of y, if x = 0,
- The slope gives the predicted change in the mean value of y, if we increase the x value by one unit.

As this is the *linear* regression, it makes no difference, what the actual x value is, every one-unit change in x is accompanied with a slope-unit change in the predicted mean y. The fitting of the model is mathematically the same, and so is its interpretation!

Now we simulate a regression for 2 continuous vars (N = 100, mean = 0, sd = 1):

We run the model and plot out 50 random fits, using 50 random draws from the posterior.

```
simple_lin_m1 <- brm(data = xy_data, y~x,

                     prior = prior(normal(0, 1), class=b),


                     file = "models/simple_lin_m1.rds")
```

All regression lines go through both mean(y) and mean(x) – both are 0 in our sim. The data range is about 3-4 units: see how the fit gradually becomes uncertain beyond that. Whether or not it makes sense to predict mean y values beyond the x data range is a scientific question – the relations in nature are rarely linear beyond a short range, if that. We are not doing linear regression because we think that the world is linear, we do it because it is convenient and the models are easy to interpret from coefficients alone. The surprising thing is, though, that it often works. This is the metaphysical mystery behind linear regression: it tends to work.

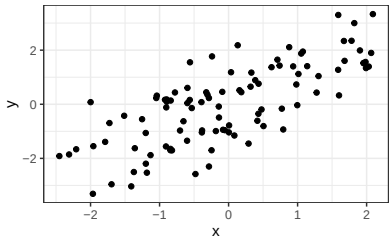


Figure 18: Simulated variables to test linear regression for continuous predictor.

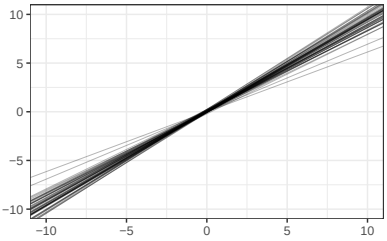


Figure 19: The model is fitted with data. We see 50 random predictions of best fit based on 50 random draws from the posterior.

Adding predictors

We add additional predictors when we believe (i) that all predictors $(x_1, ..., x_i)$ influence the predicted variable (y), and (ii) that those influences are independent of each other. Independence leads to an additive process model.

$$y \sim normal(\mu, \sigma)$$

$$\mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i$$

and priors for α, β_1, β_2, and σ.

What happens if we expand our regression model by including more predictors?

Now the slope of x_1 tells us, how much mean y changes, if we change x_1 by 1 unit and x_2 by 0 units (keep x_2 constant). Equivalently the slope of x_2 tells us, how much mean y changes, if we change x_2 by 1 unit and keep x_1 constant.

There is another way of interpreting multiple regression betas: β_1 tells us, what extra information on predicting the y value we get from x_1, if we already know how x_2 influences y. And conversely, what extra value in predicting y do we get from x_2, if we already know how x_1 influences y. For instance, we can predict someones height from the length of his right foot, but if we already know the prediction from the left foot, the right foot gives us (almost) nothing. And since including it in the model leaves very little covariation with height that is not also covariation of the left foot with height (and vice versa), then most of the predictive information that is included in the R^2, is not included in the respective betas (which consequently have extremely wide CI-s). This is collinearity.

some points:

- there is no reason to think that if we compare models $\mu = \alpha + \beta x_1$ and $\mu = \alpha + \beta_1 x_1 + \beta_2 x_2$, then β and β_1 would be similar, or even have the same sign. The first is the total effect of x_1 on y, and the second is, in causal interpretation, the direct effect of x_1 on y.
- if you have several predictors in your regression, then for easier comparison of the betas, it makes sense to standardize them. Then the predictors that have betas farther from zero, have more effect on y.
- If you have only continuous predictors, use this `x_st = (x - mean(x))/sd(x)`. Then all your predictors have mean = 0 and sd = 1. Then the intercept is the value of y, when all predictors are at their mean values (which is always 0), and each slope gives the change in mean y, if the relevant predictor changes by one standard deviation, and all other predictors remain unchanged.
- If you have both continuous and binary predictors, then you will get approximate comparability of betas with the following transformation: `x_st2 = (x - mean(x))/(2*sd(x))`. Now the meaning of the slope changes - it is the change in mean y if x changes by 2 standard deviations (that is from the 2.5th quantile to the 97.5th quantile, which is a change that covers almost the range of x variation). As before, all other X-s are kept constant, and intercept is connected to mean values of all X-s.

- sometimes transforming a predictor helps to linearize the y vs x relationship.

The thing with adding predictors to linear regression process model, is that the additive model structure specifies parallel slopes. To see, what this means, lets simulate some data. We have a categorical variable z with 2 levels, "A" and "B", denoting experimental conditions A and B. We have a x-variable (akin to a time variable), which covers the range 0 to 2 with 50 equally spaced points. And we have a continuous measurement variable y, which in experiment "A" goes up as x increases, and in experiment B goes down with equal rate.

```
## # A tibble: 6 x 3
##   z          x       y
##   <chr>  <dbl>   <dbl>
## 1 A      0      -0.145
## 2 A      0.0408  0.159
## 3 A      0.0816  0.0636
## 4 A      0.122   0.226
## 5 A      0.163  -0.113
## 6 A      0.204   0.130
```

What happens if we model these data by regression equation $y = \alpha + \beta_1 x + \beta_2 z$?

This model completely fails to capture the simulated data! The slopes of the A and B experiments are restricted to be parallel (fixed) by the additive model structure. The regression lines cross the $mean(y|x = A)$ and $mean(y|x = B)$ points, which in this simulation is the same value (1). This model fit loses in its predixtion the X-shaped data – the model thinks that A and B conditions lead to pretty much identical outcomes!

We can save the day by introducing an interaction term, which is actually a multiplication term. Like this $y = \alpha + \beta_1 x + \beta_2 z + \beta_3 xz$.

Now the model fit perfectly recaptures the simulated data. Its general intercept = 0, its deflection for z=B is 2, its slope for x is 1, and its deflection for x=B is -2. So the fitted model is $y = 0 + 1x + 2z - 2xz$

these are the fitted model coefficients.

```
posterior_summary(m_sim_interaction)[-6, ]
```

```
##                Estimate  Est.Error       Q2.5      Q97.5
## b_Intercept -0.0110400 0.07945562 -0.1682108  0.1379032
## b_x          0.9610941 0.06813031  0.8323058  1.0946147
## b_zB         2.0473411 0.11085603  1.8273933  2.2652138
```

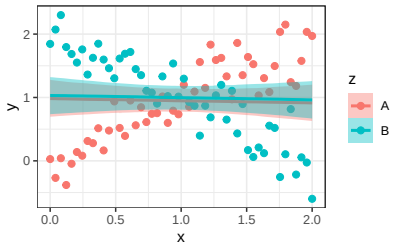


Figure 20: An additive model y = a + b1 * x + b2 * z, where z is a two-level categorical variable leads to fixed parallel slopes.

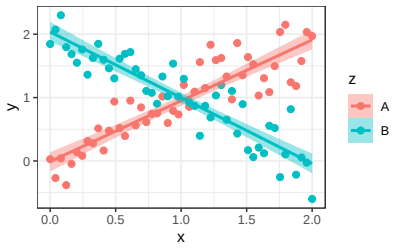


Figure 21: An interaction model y = a + b1 * x + b2 * z + b3 * x * z leads to free slopes.

```
## b_x:zB      -1.9991338 0.09518095 -2.1831346 -1.8108486
## sigma        0.2908296 0.02123843  0.2525395  0.3349225
```

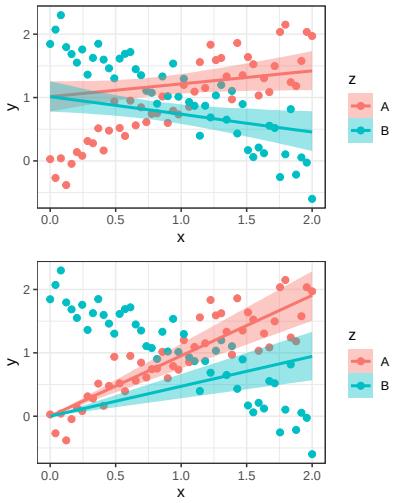
What happens if z is a continuous variable? The interaction model still works and gives a separate x slope for every z value, and conversely, it gives a seprate z slope for each x value.[46] This means that the model expects that the relationship between x and y, e.g. how much a unit increase in x predicts the mean y value at this x value, to depend on the value of z. And vice versa, it symmetrically expects that the relationship between z and y, e.g. how much a unit increase in z predicts the mean y value at this z value, to depend on the value of x. The model neither knows nor cares, which variable, x or z, is of main interest to you, or which (if any) of them causally affects y. But it does know (assume) that the variablity of x is identical at all levels of z, and vice versa. Also, it provides no shrinkage.

What happens, when we omit the additive part of the model like this: $y = \alpha + \beta xz$?

Now the fitted intercept is at the mean values of $x|z = A$ and $x|z = B$, which in both cases is 1. The slopes are 0.2 and -0.3 for A and B, respectively.

And an even simpler model $y = \beta xz$, where intercept = 0. Here the slopes are 1 and 0.5, for A and B, respectively.

[46] It must be noted, though, that continuous interaction models require more data to accurately estimate the interaction term (the difference between slopes) than simple additive models.



6. Multilevel models.

In the first instance, multilevel models are used to model data that falls naturally into groups. In these cases we can partition variation (i) into data level variation and (ii) into between-groups variation. The data level variation is captured by SD, as in ordinary single-level regression, and the between-groups variation is really the SD of group means. So if we have a 100 test results from 10 schools, then this variation is calculated as standard deviation from 10 estimated school mean scores (and the data level variation is from the 100 tests). The beauty of this approach is in that when the model if fitted information flows from the data into all levels of the model (individual level and the school level) and also moves between the levels. So the parameters for each school are informed both by data coming from this school, and data coming from all other schools. This way we use data more efficiently than in ordiary regression.

6.1. ANOVA-like models for fighting false alarms

We have a table containing i rows, where for each row corresponds (1) a single pupils math score, (2) his/hers school name, (3) sex and (4) as a school-level variable the number of pupils in that school. There are j schools in this four-column table and each pupil has a single score from a single school. For brevity lets name the columns as follows:

y - score

sch - school name (index) as a factor

We start, again, by simply modeling y as a normal distribution. This model says that the scores are randomly drawn from a normal distribution of scores, whose parameters are determined on the actual scores.

The model looks like this

$$y_i \sim N(\mu, \sigma)$$

$$\mu = \alpha$$

or written more concisely in a single line:

$$y_i \sim N(\alpha, \sigma)$$

Here we simply redefine μ as α or intercept.[47] This is mathematically meaningless, but will come handy later when we start adding predictors to the model. N stands for *Normal distribution*, plus priors for the 2 parameters (μ and σ). In the R modelling language it can be written as $y \sim 1$. This model knows nothing about pupils belonging to different schools and its the *complete pooling model*.

[47] In the model description equal sign (=) means deterministic redefinition, tilde (∼) means stochastic relation.

The opposite, *no-pooling model*, models the mean of each schools score completely independently from other schools (but it may or may not model the pupil level SD as identical across the scools, a.k.a. as full data average):

$$within\ school\ j : y \sim N(\alpha_j, \sigma_y)$$

This mathematical description means that there are as many models as there are schools (thus j models), each with its own μ, which is renamed as α or intercept in the regression language. α_j means that there are j different alpha coefficients, each standing in for a mean score of a different school. There is also a single data level SD of all scores for all schools, σ_y, which is the same thing as the σ in the previous model.

In R language this model is $y \sim 0 + sch$.

And the version where both mean and SD are modelled completely independently looks like this:

$$y_i \sim N(\alpha_j, \sigma_{y[j]})$$

or $bf(y \sim 0 + sch, sigma \sim 0 + sch)$ (This version works in brms).

If there are plenty of data for each school, then this sort of model would be good for assessing the mean test result for each school.

But what, if some schools have only a few datapoints? Then there we will see some schools, where only a few students took the test and where the results might not be representative of the actual quality of the school. We are wary of extreme results from such schools, but we do not want to ignore them either. What we can do instead, is to build a model where estimates of the mean scores for all schools are shrunk towards the mean score of all schools. We do this by adding another 2nd level regression model.

$$y_i \sim N(\alpha_j, \sigma_y)$$
$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha)$$

in brms this 2-level model reads as $y \sim 0 + (1|sch)$

The second row of this model says that the school-specific mean scores themselves come from a normally distributed infinite population of mean scores.

This model has j + 3 parameters:

- j α-s - j mean scores, one for each school,
- σ_y - SD of the distribution of scores,
- σ_α - distribution of the school-specific mean scores, not individual scores,
- μ_α - the mean of the j α coefficients of the individual schools, which is again not the mean of scores but the mean of the mean scores.

We can also write the previous model like this:

$$y_i \sim N(\alpha + \alpha_j, \sigma_y)$$
$$\alpha_j \sim N(0, \sigma_\alpha)$$

Here we get the common mean score over all schools as α and deflections from this for each individual schools (α_j), which should be added together, to get the estimate for j-th school. In brms this model is $y \sim 1 + (1|sch)$. This form of the model, while mathematically equivalent, is a bit easier to fit, and arguably to work with.

To recap:

- In the no-pooling model, the α_j-s correspond to j school means.
- With complete pooling the α_j-s are all fixed at a common α.
- In the multilevel model α_j-s are assigned a probability distribution, whose 2 parameters are estimated directly from data. The distribution pulls the estimates of α_j toward the mean level μ_α, but not all the way.
- If $\sigma_\alpha \to \infty$, there is no pooling;
- If $\sigma_\alpha \to 0$, the estimates are pulled to zero, giving complete pooling.

The last 2 points mean that single-level models can be viewed as special cases of multilevel models. Thus, even if we only have 2 groups, or many of the groups have very small N, the multi-level model is at least as good as the corresponding single-level model!

One way to interpret the variation between schools, σ_{alpha}, is to through the ratio, σ_α / σ_y. If the ratio is small, then most of the variation in scores is at the pupil level and the schools are very similar to

each others. Otherwise, different schools have different mean scores. The goal of the model is to look beyond the samples of scores from each school and into the true mean ability of each schools students, as if each school contained an infinite number of test takers whose average ability is the issue.

Lets suppose that the fitted data level SD, σ_y, is 0.74 and the SD for the mean scores, σ_α, is 0.33. Then we can see that to get from 0.74 to 0.33, we have to divide 0.74 with square root of 5. This means that the SD of mean scores is the same as the SD of 5 individual score measurements, and that the amount of information in this 2nd level distribution of the means is the same as that in 5 measurements within a school. Thus, a school with <5 scores, puts more information into the group-level model than into the schools own model (in the sense of providing a higher-SD estimate of the schools mean score). As a result, the mean score estimate in an N<5 school is closer to the complete-pooling estimate, and closer to the no-pooling estimate when N>5.

We can express this very same 2-level model in an alternative way by putting priors on priors, that is, by using meta-priors. With the meta-priors we assume that the scores of each school have their own distribution that partially overlaps with the other schools, and that these school-specific distributions of test scores come from a single overarching meta-distribution, defined by the meta-prior.

If we use the meta prior, then we need to turn our usual school-specific prior, over which the meta-prior arches, into an adaptive prior. For this we take the school-specific prior $\alpha_j \sim normal(0, .)$, where the dot stands for some concrete sigma value that you care to put there, and substitute this value with another parameter, σ_{sch}, that needs to be fitted. So the adaptive prior contains an additional parameter, which in turn needs a meta-prior to be fitted.

During fitting the meta-prior facilitates using school-specific information laterally, across the schools, leading to partial pooling of information and shrinkage. Information flows from an individual school up to the meta-prior, and down again to other schools.

$$y_i \sim N(\mu, \sigma)$$

$$\mu = \alpha + \alpha_j$$
$$\alpha \sim N(., .)$$

$$\alpha_j \sim normal(0, \sigma_{sch})$$

[48] adaptive prior for each school

$$\sigma_{school} \sim cauchy(.)$$

[49] meta-prior

$$\sigma \sim exponential()$$

[50] prior for the student-level SD

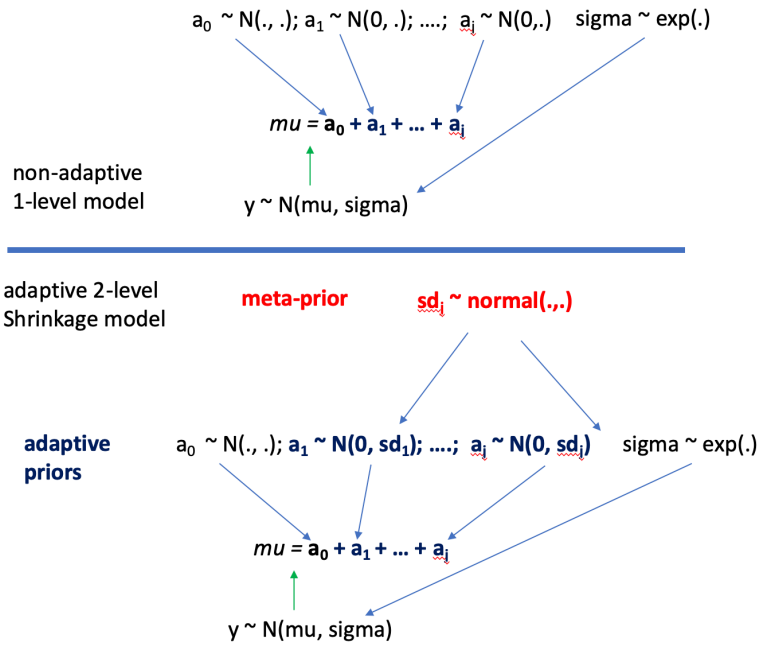


Figure 22: Schemes of the non-adaptive 1-level model and of the adaptive 2-level model.

It is easy to expand this model by adding another level. If we suppose that the meta-prior denotes school district, we cold simply turn it into another adaptive prior by putting another 2nd-level parameter into it, and then adding another meta prior (3d level), to denote the city. Then we have each school under 2 levels: schools belong to school districts, and school districts belong to the city.

When to use multi-level models? The answer is that using them should be your default position whenever data generating process contains grouping structures that have distinct variations. Such groups can be defined by schools, cities, countries, enzyme batches, study centers, experimenters.

> If you think that there is something in common between the groups, but also something different, consider using a multi-level model. If you

think that there is everything in common (the variation of each group is identical to all others), then ignoring the groups and running a single 1-level model should not change the result.

Simple single-level models are often bad for hierarchical data: with few parameters, they cannot fit large datasets well, whereas with many parameters, they tend to overfit. Multilevel models can have enough parameters to fit the data well, while using a population distribution helps to avoid overfitting. It is often sensible to fit hierarchical models with more parameters than there are data points.

When j is small, it is difficult to estimate the σ_α and multilevel modeling often adds little beyond no-pooling models. When σ_α cannot be estimated well, it tends to be overestimated, and so lead to no pooling. With 1 or 2 groups multilevel modeling reduces to classical regression. Even 2 observations per group is enough to fit a multilevel model. It is even acceptable to have 1 observation in many of the groups. When groups have few observations, their means won't be estimated precisely, but they can still provide information that allows estimation of the coefficients and variance parameters.

Multi-level models carry 6 main advantages:

1) a multi-level model gives more honest posteriors (CIs) for higher-level estimates (in our example, the estimated city-wide mean test score). These posteriors tend to be wider than the ones calculated from a single-level model.

2) A multi-level model allows to estimate at each level – you get an estimate for each school, for each district, and for the whole city.

3) A multi-level model defends against false alarms by shrinkage of the estimates. This is usually more efficient and also better interpretation-wise, than correcting for multiple testing by adjusting p values/CI-s – but leaving the actual effect sizes at their raw data level.

4) A multi-level model more accurately reflects clumpy data generating mechanisms in the wild. This makes it easier to give scientific interpretations to it.

5) A multi-level model allows to find batch effects and other anomalies of experimentation.

6) A multi-level model allows to generate new *in silico* data at every level, including generating new groups.

The multi-level model we introduced is a kind of Bayesian one-way ANOVA analogue (only ANOVA does not provide shrinkage). This model is much easier to expand than ANOVA models. We can

substitute the normal likelihood with biomial likelihood, for example to do a similar shrinkage model for fractions, and/or we can easily turn this into a full linear, or generalized linear, or even non-linear regression model (see next chapters). All of these models provide shrinkage and take into account the clumpiness in your data.

When not to use multi-level models? Avoid them if your groups do not come from a single overarching super-group. For example, if you have several control groups and a single experimental group, then putting them together under a single meta-prior can lead to too much shrinkage for the experimental group. But if you have several control groups and several experimental groups, then its perfectly all right to put both kinds of groups under their own meta-prior.

To test our intuition, lets build an artificial dataset of 10 schools that have identical "true" means of test results, each = 100, drawn from a normal distribution with SD=25. The 10 schools have different numbers of test-takers: 1 in school 1, 2 in school 2, 3 in school 3, . . . , 10 in school 10.

Here are the first 6 rows

value	ID
120.00000	1
100.46865	2
95.39369	2
65.71674	3
85.02081	3
107.36363	3

In the next graphs we are going to show not the full means, but school-specific estimates of deflections from the overall mean (over all the schools).

Our first model treats all schools completely independently. There is no information transfer between schools. We use reasonable priors (not very strong, but not very weak either) for mean and SD. This considerably narrows our estimates in schools where N is very small.

In school 1, where a single student took the test, thanks to the prior we still have a reasonably informative estimate for that school – we believe that this school could be somewhere from 25% below average to about 50% above average. With a single student it would be strange, if our estimate would be much more precise.

Next we present a model, where school means are completely independent, but school SD-s are completely pooled.

The pooling of variation information narrows our CIs a bit, but not that much.

And now a 2-level model, which partially pools information on the

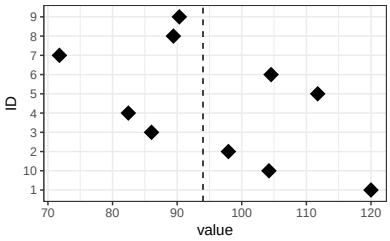


Figure 23: The empirical sample means of the 10 schools. All samples have randomly drawn from the same population (mean = 100, SD = 25). The global sample mean, at 94, is shown as dotted line

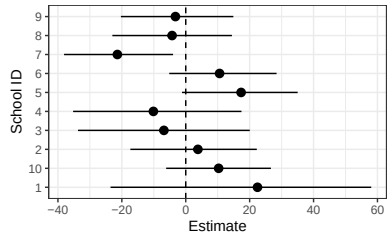


Figure 24: The estimates for mean test result with 95 percent CIs for each school, modeled completely independently.

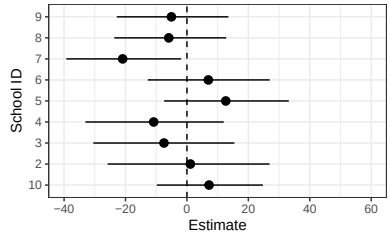
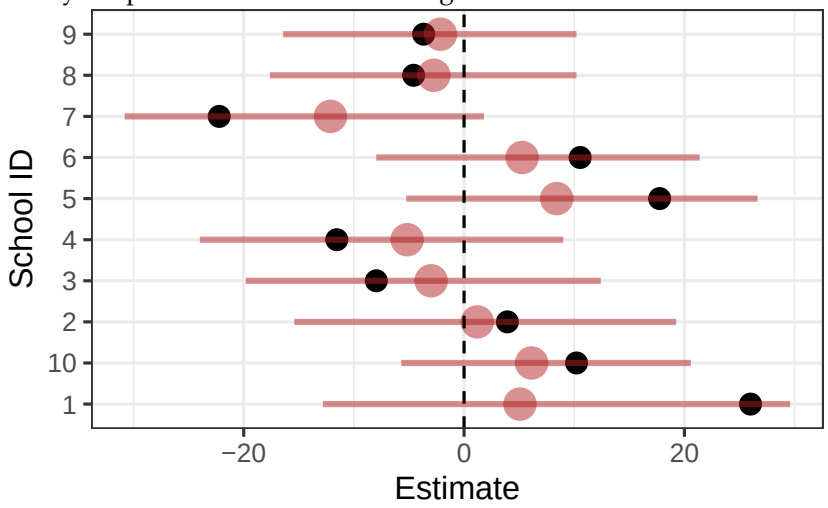


Figure 25: Classical linear regression estimates the means independently, while completely pooling information on variation across the schools.

means, leading to shrinkage towards the global mean.

Here the CI-s are clearly narrower.

Finally we plot out the actual shrinkage for each school



The school 1, which had a single datapoint, clearly gives the most shrinkage, while its CI-s are now very close to other schools with more data. School 2, which only has 2 datapoints, happens to be close to overall mean at the raw data level, and therefore gives the least shrinkage. (Shrinkage is shown by the red points with their estimation uncertainty cptured by 95% CI.)

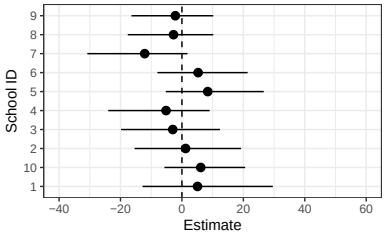


Figure 26: The 2-level model results in shrunk estimates.

6.2. *Multilevel regression models: free intercepts*

Now we add two predictors to our schools example: X - is students sex (male = 1 or female = 0) and S - is the school type (gymnasium or not). S is a group-level predictor, so there is only one S value per school, which is repeated in the table for all pupils in this school.

We start with converting the simple regression y = a + b1x + b2s + b3sch + error into a 2-level model. (In brms this model is $y \sim x + s + (1|sch)$.)

$$Within\ school_j : y_i \sim N(\alpha_j + \beta x_i, \sigma_y)$$

$$\alpha_j \sim N(\gamma_0 + \gamma_1 s_j, \sigma_\alpha)$$

The student-level predictor is in the 1st level and the school-level predictor is in the 2nd level of the model. There are j intercepts but only one slope for both predictors, which means that the model says that whatever the schools mean score, it will have the same difference between the sexes and the same dependence on the size of the school. We will relax this strong assumption in chapter 6.3. The 2nd level model models the J school means as dependent on the intercept

and slope of the school size regressor. The multilevel model combines the J local models in two ways: first, the local slopes are the same in all J models. Second, the J intercepts are connected through the group-level model. The meaning of σ_y and σ_α is the same as before (see chapter 6.1.), except the variance is estimated after conditioning for various predictors. Group-level predictors reduce the unexplained group-level variation and thus the group-level SD of σ_α, thus leading to more shrinkage and to more precise estimates of the group means, especially for groups with small N. Adding predictors at the individual and group levels typically reduces the unexplained variance at each level (adding a group-level predictor can also increase the unexplained variance).

6.3. Prediction from multilevel models

predict the test score for a schools males (x = 1) in j = 6. Conditional on the model parameters, the predicted value has a mean of $\alpha_6 + \beta$ and α sd of σ_y.

$$y|\theta \sim N(\alpha_6 + \beta x, \sigma_y)$$

where θ represents the vector of model parameters. to predict the score for a male student, but in an unmeasured school, we generate a new school-level error term, α, which we sample from its $N(\gamma_0 + \gamma_1 u_j, \sigma_\alpha)$ distribution. The prediction will be less certain than for a student in a measured school, since we lack information about α. From the fitted model, the within-school (residual) sd is σ_y, and the between-county sd is σ_α. The score for a new student in a measured school can be measured to an accuracy of about +/- σ_y. The score for a new student in an unmeasured school can be predicted to an accuracy of $\sqrt{\sigma_y + \sigma_\alpha}$. The equation $\frac{\sqrt{\sigma_y+\sigma_\alpha}}{\sigma_y} - 1$ gives the percentage by which a predictive interval for a new student in an unmeasured school is wider than that of a new student in a measured school.

6.4. Free intercepts and slopes

First the model without school-level variable S – $y \sim x + (x|sch)$ in brms syntax.

$$y_i \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y)$$
$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta \end{pmatrix})$$

where ρ models between-group correlation. This means that we model the school specific intercepts and slopes together so that there

is also partial pooling of information between intercepts and slopes, and we estimate the correlation between them accross schools.

We now expand this 2nd part of the model to include the school-level predictor S.

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N(\begin{pmatrix} {\gamma_0}^{\alpha} + {\gamma_1}^{\alpha} u_j \\ {\gamma_0}^{\beta} + {\gamma_1}^{\beta} u_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta \end{pmatrix})$$

Its brms version reads $y \sim x + s + x : s + (x|sch)$, so we have an interaction between the school type and sex of the pupil. Varying slopes can be considered as interactions between group indicators and an individual-level predictor.

Sometimes it might be reasonable to allow the slope, but not the intercept, to vary. When J separate experiments are performed on samples from a common population, with each experiment randomly assigning a control condition to half its subjects and a treatment to the other half, while the "controls" are the same for each experiment but the "treatments" vary. Then we remove the intercept from the group of coefficients that vary by group $y \sim T + (T - 1|sch)$ where T is the treatment variable $0, 1$.

Now we get a different treatment effect for each school And here is model with an individual-level predictor x: $y \sim x + T + (T + x : T - 1|sch)$, where the treatment effect and its interaction with x vary by school

There is nothing wrong with a high correlation between the αs and βs. As with interaction models in classical regression, centering the predictor(s) may not remove these correlations, but the correlation is easier to interpret.

Appendix: Statistical Best Practices

The following should be taken as recommendations, at your own responsibility. Statistics is complicated and most often there is no single best way of doing it. Neither is there a single correct answer.

Data structure

For analysis in R often the tidy format is best for plotting data in ggplot.

Tidy data: each row is an object of measurement (what is measured) and each column is a single variable that is measured. Different types of objects of measurement (mouse, human) go into separate tables. If there are several measurements per object of measurement (time points, etc), then there are several rows for every object of measurement, and a separate ID variable.

```
tribble(~ID, ~age, ~cholesterol, ~weight, ~sex, "A", 67, 180, 87, "M", "A", 68, 167,
    87, "M", "B", 56, 162, 72, "F", "B", 57, 176, 72, "F") %>% kable()
```

ID	age	cholesterol	weight	sex
A	67	180	87	M
A	68	167	87	M
B	56	162	72	F
B	57	176	72	F

In this table we have 2 objects of measurement (2 people), A and B, both have given 2 cholesterol measurements a year apart, plus there are weight data that have been measured once and data on the sex (male or female). This is tidy data, which is optimized for visualization with ggplot2.

For biological data, other formats are often required for compatibility with specialized analysis packages.

EDA - exploratory data analysis

The goal is to check for general features of data (range, variation, distribution), as well as surprising and alarming features, like impossi-

ble values and other outliers. But also check possible co-dependencies in data by looking at co-variation between 2 or more variables. So EDA checks the data quality (and by this, the quality of the experiment and/or data collection), but it also tries to form a first impression about anything scientifically interesting in the data. EDA aslo gives hints about possible routes for data pre-processing for formal modeling.

Always start your analysis by plotting out the data, and simple dependencies. Never start by formal modeling of the data.

Visulizations of one-dimensional continuous data in the order or increasing information loss:

1. strip chart plots out single datapoints in one-dimensional data (a single variable). No information loss.

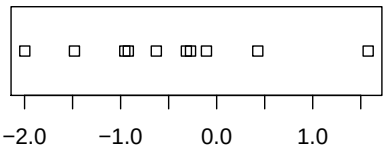


Figure 27: A strip chart of normal data, n=10

2. histogram plots out single-dimensional data with loss of information. The shape of the histogram depends on arbitrarily set bin width. Therefore, it is often a good idea to try different binwidths and see, what happens.

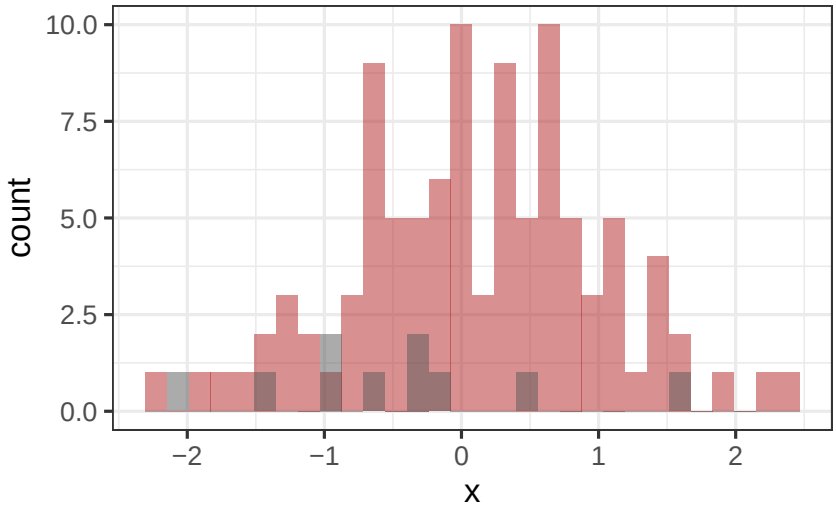


Figure 28: two histograms of normal data, n=10 and n=100 (red)

3. density plot is a smoothed histogram whose y-axis is normalized into probability scale. The area under the density plot equals one. Violin plot is another form of density plot.

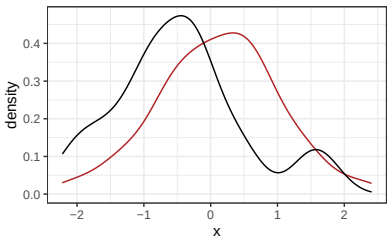


Figure 29: density plots of same data as before

For comparisons of several continuous data distributions:

1. Individual data points with medians (preferable) and/or means, shown as bars. This should be your default representation when

neither the sample size n or the number of comparison k is not large. For n < 10 and k < 10, this is nearly always the best option. There is no information loss! Also, it is possible to superimpose standard deviation/standard errors/confidence intervals.

2. For at least moderately large n (>30, say) and small k (2-4, say), superimposed histograms or density plots are a good choice.

3. For at least moderately large n and moderate k (10-30, say), violin plots do the trick

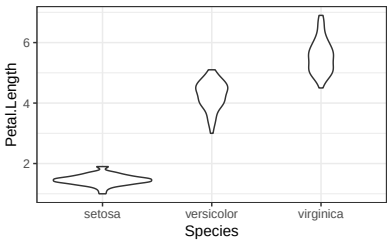


Figure 30: violin plot of iris data

4. For at least moderately large n and fairly large k (20 - 50, say), consider box plots. Box plot presents the median, 25% and 75% quantiles and (typically) whiskers that strech for 1.5 times interquartile range (until the last datapoint that falls inside 1.5xIQR). Datapoints outside 1.5 x IQR are shown separately as outliers. Box plots lead to largest information loss, but can be a good option, if one wants to show many distributions (many variables) side-by-side.

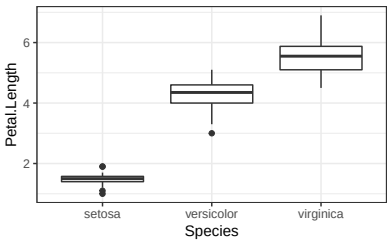


Figure 31: box plot of iris data

5. For at least moderately large n and large k (>30), consider joy plots, a.k.a. ridge plots.

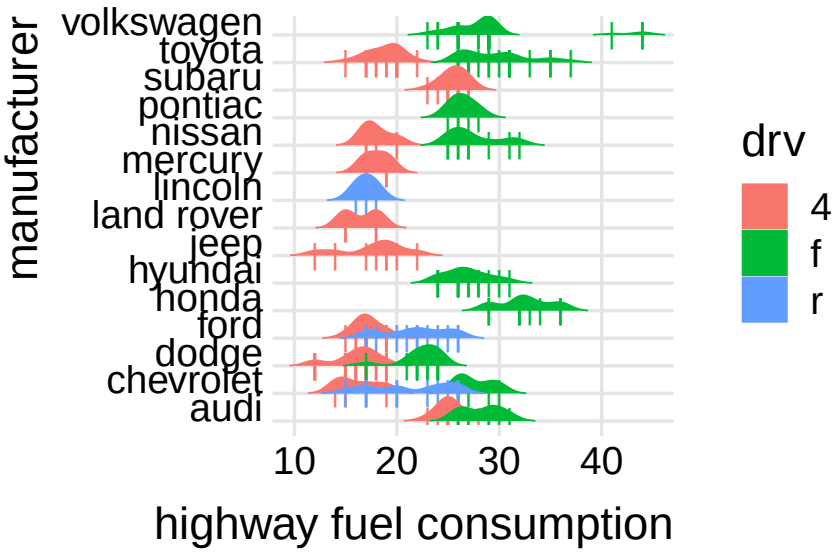


Figure 32: A joy plot of fuel consumption of cars of different producers and drives. Individual cars are shown as ticks.

For examining co-variation in continuous data:

1. xy scatter plots, plus linear regression fit. To this one can add dimensions by stratifying by color of the points and regression lines

(either for categorical or, less preferably, continuous stratifying variables), by shape of the points and regression lines (for categorical stratifying variables). In Fig . . . we see a 5-dimensional data representation in two-dimensional space. As you can see, some dimensions/variables are easier to grasp visually than others, depending on the geometric representation of a variable.

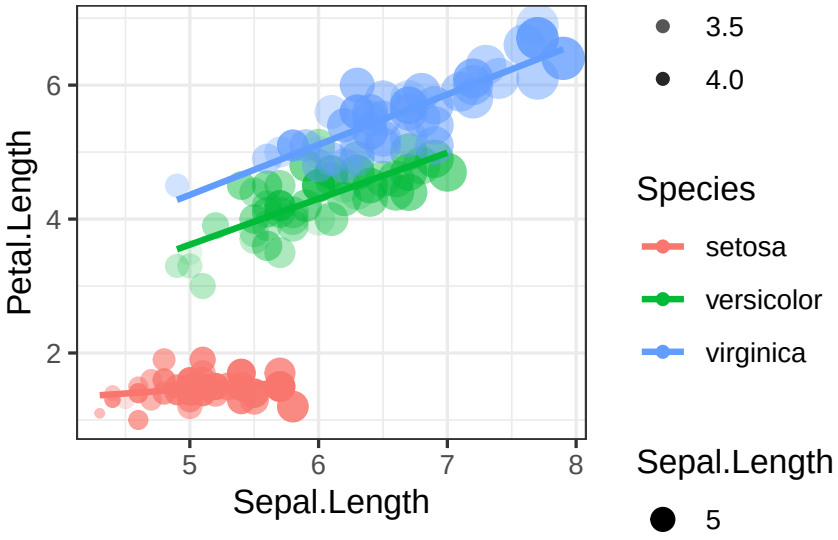


Figure 33: scatter plot of 2 measured variables of 3 Iris species. Each point corresponds to an individual plant. Point size indicates sepal lenght of the plant and opacity indicates sepal width.

To plot more than about 4 dimensions of data, consider using dimension reduction techniques, like principal component analysis (PCA) and its relatives (tSNE for large datasets with non-linear dependencies, for example). When you do use them, understand that they are harder to interpret and can be seriously misleading. Dimension reduction requires great care and detailed knowledge about the assumptions behind each technique.

A psychological note: Human cognition is good in accurately perceiving differences in bar lengths that are anchored at a common starting point. We are lousy in perceiving differences in areas, in angles, and in color saturation/intensity. So avoid, as much as you can, comparisons through circle diameter or area (in addition, in published work these are usually not differentiated to the reader).

What to do and what not to do in plotting data for publication.

While EDA you are doing for your own consumption, data is also plotted for publications, and there the goal is not to find new stories, but to present the ones that you have already made up. Such figures are presented to further the storytelling. This means that graphs are no longer neutral, but try to convince the reader of the validity of a

made-up story. Like an official photograph of a sex symbol, graphs can be deceptive!

1. Do not use pie charts to show proportions of a whole. In fact, never use pie charts at all. To visualize proportions, use bar plots in the range 0...1.

2. For continuous variables, consider avoiding bar plots. Just plotting the means as dots or line segments is often visually better. Then its also easier to include individual data points

3. When the x-axis is continuous, prefer line plots. If intermediate points on the x axis between the measured points are scienifically possible, use the line graph, rather than bar plots.

4. Be careful with plots with two distinct y-axes. They are controversial and can often deceive. Use of two y-axes is not controversial, when the two axes are transformations of each other (like Fahrenheit and Celsius).

5. Remember that by choosing the ranges of y axis (but also the x-axis) you can make the effects seem larger or smaller to human perception. Even the aspect ratio of the graph can make a great difference in terms of perceived slopes. Use this power responsibly!

6. Do not superimpose straight lines over non-linear associations. Use loess-lines instead (like the geom_smooth function with default settings).

In general, when making publication-quality graphs pay attention to how much of the ink in the graph is actually showing data (like lines, dots, *et al*.), versus other elements of the graph (like borders, shading, axes with their ticks and whatnot). You want this ratio to be rather large. Avoid focusing readers to elements of the graph that are not directly needed for understanding the data. For this reason a sucessful black-and-white graph is better than one in grayscale, which in turn is better than a colour graph. In the end, your goal is to make the reader understand, what you want to say with a graph, in a shortest possible time. Of course, this does not mean that graphs cannot be complicated and take a long time to ingest. They can, if you do not have a better choice.

Data summaries

If you are graphing individual datapoints, then summarising the data by the median (which is also graphed) is the safest bet, because the mean (arithmetic mean) becomes hard to interpret for

non-symmetrical data distributions. The median always has the same interpretation as the data point, from which half the data is larger and half is smaller.

The most likely datapoint is not the median, but the mode. The mode is often harder to estimate with any precision than the median or the mean. The mean is more efficient than the mode or the median, in the sense that one can build narrower confidence intervals around the mean than the median. But this does not change the fact that median often makes more intuitive sense. In fact it is hard to imagine a situation in biology (although not in economy) where the mean makes better conceptual sense than the median.

For log-normally distributed data (and a lot of biological data are log-normal) the geometric mean carries the same meaning in the original scale as the arithmetic mean in the logarithmic scale. The geometric mean should be accompanied with multiplicative standard errors.

> For description of central tendency use mean for symmetrical distributions, but the median for most datasets you come accross.

For summarising data variation the most common measure is standard deviation (SD). SD is defined conditionally on data distribution. For most distributional models used in biology one can calculate the interval, which encompasses 68% of the area under the peak. 1.96 SD-s cover 95% of the area under the peak. This interval is the SD. So, for normal, lognormal and many other distributions there are different algorithms that give you the SD. The usual equation for SD that is thaught at shool applies for the normal distribution.

A close relative of the SD is the **S**tandard **E**rror of the **M**ean (SEM). When we take a random sample sized n from a population then 68% of its individual datapoints fall within the interval $mean +/- SD$. But when we take many such samples with size n, and calculate the mean for each one of them, then 68% of these means fall within the interval SD divided by square root of n (centered on the mean of the means, or the combined mean of all the samples). This new measure $\sqrt{n}$ is the SEM, which is a measure of our uncertainty about our estimate of the mean.[51] If n is large, then 1SEM = 68% CI, 1.96SEM = 95% CI and 3SEM = 99% CI. In biology, n is ususally not large enough for this simple conversion; also note that, even with large n, presenting on your figures 68% CI-s is probably not what you really want to do.

[51] SEM is sometimes called the "SD of the mean".

An alternative to SD is the MAD - the median absolute deviation. If SD is normally calculated as

$$SD = \sqrt{\frac{\sum(x_i - \hat{x})^2}{n-1}}$$

then mad is

$$MAD = \frac{\sum |x_i - \hat{x}|}{n}$$

The MAD is both more intuitive and less sensitive to outliers (note the lack of the squaring operator). However, it does not have the neat mathematical properties of the SD (remeber the SEM and the CI).

As a convention, mean and SD should be used together, as median and MAD should be used together.

Another useful summary of data variation is the interquartile range (IQR). This is usually geven as 5 numbers: the minimum value – 25th percentile – median – 75th percentile – maximum value. So the IQR really covers half of the data around the median (this is exactly what the box in a boxplot does).

Beware of summarising data via min and max, as these depend on sample size. The more data you have, the further min and max tend to creep from the median. For the same reason, beware of estimating population distribution from the shape of the data distribution, as even highly non-symmetric distributions can appear near-normal with moderate sample sizes. Statistical normality tests should therefore be avoided (especially as with large n-s they always provide small p values, suggesting non-normality even with approximately normal data).

Confidence intervals

- SD-s show data level variation. As a rule, you should not put SD on a graph with data points shown (because it does not add much). Rather use it in text like this: mean (SD).
- SEM indicates uncertainty around your estimate of the mean, but it should not be put on a figure or directly interpreted.
- 95% CI means (in its Bayesian interpretation) that you are 95% certain that the true population mean lies within this interval.[52]

  A p value of 0.05 means that the 95% frequentist CI just barely touches the zero effect (assuming that H_0 is defined as a zero effect). Likewise, p = 0.11 means that the 89% CI touches the zero effect.

- If you want to show uncertainty around your estimate of the mean, use CI for the mean. If you want to show uncertainty around your estimate of the effect size (mean1 - mean2), then show the CI for the effect, not two CI-s for each mean!
- There are continuous CI-s that look like this.

[52] Frequentist CI have a different, and arcane, interpretation, but they are nevertheless often numerically similar to the Bayesian CI (and when not, Bayesian CI usually wins).

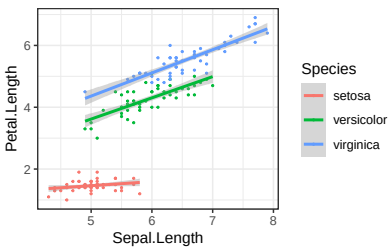


Figure 34: Continuous CI-s or confidence regions

- Do not use discrete CI-s where you can model continiuous ones, as the discrete ones are less efficient and harder to interpret.

You can estimate the difference of slopes with 95% credible intervals like this:

```
m_setosa <- brm(Petal.Length ~ Sepal.Length, data = iris %>% filter(Species == "setosa"),
    file = "models/m_setosa")
m_versicolor <- brm(Petal.Length ~ Sepal.Length, data = iris %>% filter(Species ==
    "versicolor"), file = "models/m_versicolor")
posterior_setosa <- posterior_samples(m_setosa)
posterior_versicolor <- posterior_samples(m_versicolor)
posterior_summary(posterior_setosa$b_Sepal.Length - posterior_versicolor$b_Sepal.Length)
```

```
##        Estimate Est.Error       Q2.5      Q97.5
## [1,] -0.5547548 0.1124594 -0.7798023 -0.3396564
```

So any change in sepal length is predicted to have a smaller effect on petal length in *I. setosa* than it does in *I. versicolor*. This difference is somwhere between 0.3 and 0.8 units. This is quite a large difference in slopes (their absolute values being in the same range).

```
posterior_summary(posterior_setosa$b_Sepal.Length)
```

```
##       Estimate  Est.Error        Q2.5     Q97.5
## [1,] 0.1325976 0.07122992 -0.00966488 0.2738349
```

```
posterior_summary(posterior_versicolor$b_Sepal.Length)
```

```
##       Estimate  Est.Error      Q2.5     Q97.5
## [1,] 0.6873524 0.08905853 0.5110829 0.8595665
```

P values and statistical significance

> P value is the probability of encountering your sample data or more extreme data, if we suppose that there really is no effect.

P value is in the probability scale (0…1) and it tries in effect to answer the question of how likely it is to see data like yours, assuming that there is actually no true effect. A large p value can be seen as a warning that your sample effect migth well be a random sampling error, which is not indicative of a "real" effect. Sample error is not the only possible explanation for a large p value, but as it is so very pertinent to interpreting ones results scientifically, it can rarely be ignored. A small p value is different in that while it can also mean several things, its much harder to convert into action.[53] P value says nothing about the likelihood that the effect, which you see, might be

[53] It can mean a large sample size, a low variation, a large true effect, or a large bias.

bias and it says pretty much nothing about the actual size of your effect. Most importantly, p value does not give the probability that there is no effect (that the null hypothesis is true), and neither gives it the probability that some scientifically interesting hypothesis is true or false. Thus the direct interpretative use of raw p value is quite limited.

Statistical significance is a concept that uses p values as input and actually does convert small p values into action. >If your p value is less than some arbitrary value, like 0.05, and if you then claim that your result is "statistical significant", then in the long run 5% of experiments with zero true effects are classified as "significant".

The interpretation of "significance" is (i) over the long run, rather than about a specific experiment, and (ii) it depends on the fraction of true null experiments (the ones with zero effect) from all experiments that were done.

The multiple testing problem – If you do over some time or experimental system or field of study k independent tests of k independent samples, and call all tests, whose p<0.05 "significant", then $p_{corrected} = p/k$. This is the Bonferroni correction, which fixes the familiy-wise type I error rate at the level p. Presenting any results whose p/k < 0.05 as "significant" has the same meaning as presenting a p value from a single test, whose p < 0.05 as "significant". Thus we fix the familiy-wise type I error rate at the level p. Unfortunately, this correction leads to a serious loss of sensitivity (it increases the frequency of type II errors). To mitigate this problem, a host of compromise procedures, like sequential Bonferroni, has been developed. These corrections result in a higher corrected p value, which no longer fixes the type I error level at 0.05, but lead to a smaller drop in sensitivity. Maybe the most serious aspect of the multiple testing problem is that there is no consensus about what constitutes a family of tests. It could be multiple tests done within a single experiment (like measuring the level of 20 000 mRNAs from a sample of 3 biological samples), or all experiment done by you over your lifetime, or all experiments done by various groups within the confines of a single well-defined experimental system, or whatever. Perhaps the best recommendation that can be given is to record and publish all calculated p values. But even this is far from enough, as it has been sucessfully argued that not only the tests that were done, but also those that would have been done, had data been different, count in the family k.

Multiple testing problem presents a very strong argument for abandoning the very concept of "statistical significance" as unworkable in scientific practice. The alternative approach entails Bayesian estimation of sample means and effect sizes, accompanied by mea-

sures of estimation uncertainty, which are often presented in the form of 95% credible intervals. Within this approach false alarms are dealt by the prior structures and by using multi-level modeling.[54]

Currently, the almost only sensible use of p values would be experiments that do at least hundreds of parallel measurements of each sample (that is, omics experiments in biomedicine). Then we can use the p values as input for checking the quality of the experiment (by p value distribution), to calculate the fraction of true null effects (using the shape of the p histogram), to calculate individual q values, and even the prospective power of the experiment. The idea here is not to actually interpret p values, but to use them as input to further calculations, which might result in better interpretable statistics.

> if a p value indicates P(data I Ho), then a corresponding q value indicates P(Ho I data).[55]

[54] There are also frequentist confidence intervals, which are mathematically consistent with p values (a 95% CI can be defined as a range of non-significant p values, which treat the sample mean as the null hypothesis). These intervals should also corrected for multiple testing, so in this respect they have no advantage over the p value.

[55] A q value estimates the probability that the effect is exactly zero, if we see sample data that is similar to your data. Its still a mouthfull, but is at least getting closer to what one might want from a statistic. Each q value is calculated, using the Bayes theorem, from the corresponding p value and the estimated fraction of true nulls, which in turn is calculated from all the p values in the experiment (we need at least hundreds of them).