Berlin Exchange
medicine

Inferential statistics I –
Hypothesis testing in the basic form of conditional probability / Bayes' rule in R

Absolute Beginner's Stat-o-Sphere

by Steffen Schwerdtfeger

Medical Student, Charité Universitätsmedizin, Berlin
BEM Editor – Data and Statistics / Medical Humanities

Berlin Exchange Medicine e. V.

URL: https://journal.medicine.berlinexchange.de/pub/hypotest

1 Introduction – Structure of the series

Welcome to our first tutorial series within our editorial collection “Stat-o-Sphere”. **You have just entered the absolute beginner’s sphere, so we hope your stay will not come with any turbulent inconveniences and we wish you a pleasant and insightful trip through our tutorial**. As this is our first editorial series within this collection, we decided to give a thorough overview over two concepts that we believe are most important for understanding scientific reasoning as such and that are also often heavily misunderstood – namely: the concept of **hypothesis testing** (this tutorial) and **statistical modelling** (starting in part two of this series). Knowing at least these two concepts in detail is especially for those of you important that seek to get a *stable* heuristic overview over what statistical inference is all about, without digging into every possible method.

As mentioned in the introduction of our tutorial collection, we are trying to provide you with *slow-paced* tutorials that potentially entail the *conceptual / intuitive*, the *mathematical* and the *computer scientific* perspective (programming) on statistical methods. In this tutorial series, we will provide an introduction into inferential statistics on all of the above ‘levels of abstraction’ at one place.

Not every area might be of your immediate interest. This is also the reason why we established **a short summary at the end of every chapter** that can also be used as a chapter overview for the impatient reader. In the future, we will also add chapters going through the statistics of open data papers, which will provide you with a wider range of examples with different levels of complexity and further insight to the application of statistical methods in the wild, so to speak.

Corresponding to a modular educational expansion of actual scientific work, published via our student journal, we will also add condensed chapters that will focus on aspects that are not only important for the application, but mostly for the *interpretation* of the results – e.g. within review processes. This will not be the case for this article though, as we are here introducing a rather general concept of mathematics, present in the field of statistics in various ways.

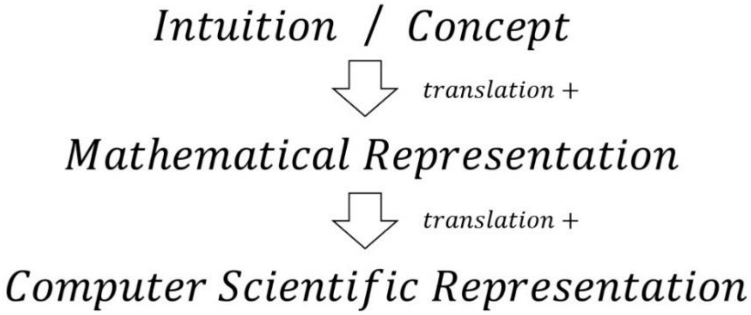

Intuition / Concept
translation +
Mathematical Representation
translation +
Computer Scientific Representation

Fig. 1 Going through all three of the perspectives involves translation most of the time, as they are more or less just three languages representing the same. Note that **when we speak of mathematics**, we will actually mostly refer to logic and less to numerical calculations, for which we have programs such as R to play around

with – which simplifies things a lot and especially fundamentally differs from what you have done in school most of the time: calculating via cognition and with a limitation of tools and time.

This tutorial is therefore structured more or less hierarchically, moving from the intuition to the mathematics and eventually to the code in R, using functions such as **lm()** for linear models – all coming together in the interpretation of the output results (most important for review). **Note that the first part of the series introduces R only to do calculations that can be done by hand or with a calculator as well!** So in case you are scared to get into programming languages, fear no more. We are trying to introduce R as what it essentially is: a very sophisticated calculator.

You may have prior knowledge and find some of the math rather boring at the beginning. As the following tutorials are also introducing coding as such, we chose to start from an "absolute" beginner's level. It also comes with the advantage that this tutorial can theoretically also be understood and mastered by scholars to some degree (we may test for that in the future). We also want to give those a chance to get a full recap of the mathematics that didn't start studying soon after school, or that are not that much incorporated in the matter for any other reason – especially those concerning a lack of interest or aversion. In the end, statistics paves the path of *every* medical inference nowadays and it does so for very good and even *intuitive* reasons, as we will see. It is easy to click through a computer program nowadays, in order to obtain mathematical values as a kind of power of speech within society, marking one's argumentation as being evident. But therefore it is also dangerous when those who do have actually only little understanding what they do and what the consequences of false or untransparent information can be.

Statistical modelling will be the core of our first series of tutorials on inferential statistics, discussed on the basic example of a *linear regression model* (eventually moving to other methods in the future, such as linear mixed effects models). Hypothesis testing in the mathematical sense will be important especially when evaluating the results of our linear regression, i.e., the full output that we have obtained using the programming language R (especially when discussing the *p-value*, obtained via a t-test as a specific form of hypothesis testing on linear regression models).

However, *hypothesis testing in its conceptual and mathematical sense is what statistical modelling is for in the first place*, so we will start our journey into the *Stat-o-sphere* by going through its steps. This will give us a clear view on what a p-value *actually* is in terms of probability theory as well as conceptually, and will show us the rather banal difference between the *frequentist* and *Bayesian interpretation* of conditional probability. There is also a vast variety of mathematical algorithms, called "tests", that all result and reflect on p-values of a model, so we decided that it is best to *not* start with a specific test for a specific type of model, but with explaining the concept of hypothesis testing itself and how the p(robability)-value represents itself not as any, but as specific *conditional probability*.

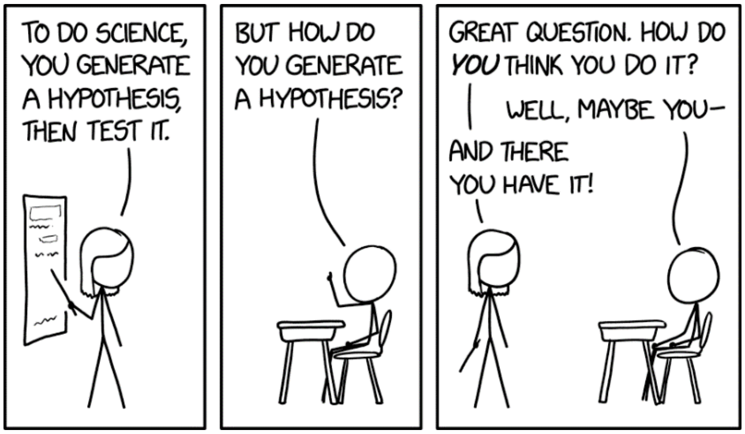

TO DO SCIENCE, YOU GENERATE A HYPOTHESIS, THEN TEST IT.
BUT HOW DO YOU GENERATE A HYPOTHESIS?
GREAT QUESTION. HOW DO YOU THINK YOU DO IT?
WELL, MAYBE YOU–
AND THERE YOU HAVE IT!

Fig. 2 From the webcomic 'xkcd', titled "Hypothesis testing", which is also referred to as abductive inference (xkcd: Hypothesis Generation). Any inference we make as humans can be referred to as abductive inference: making a guess on the world (given prior information) and adjusting it, given new information on the world in order to update a model. Statistical inference can be understood as a way to extend the method by making its steps transparent via mathematics: a mathematical representation of an argument on the world.

We believe that discussing conditional probability as a start is the most economic approach to statistics and data science in general, as it provides a conceptually consistent overview over a vast variety of methods involved in statistical reasoning in general, apart from the p-value (to name a few: positive predictive value, even thermodynamics, information theory (Shannon), AIC, BIC, 'machine learning' (e.g., diffusion models), computational neuroscience ("Bayesian Brain Hypothesis") etc.). Another reason is that hypothesis testing understood as conditional probability is actually really simple and especially surprisingly intuitive *and doesn`t need any mathematical background at all to be understood*.

However, it appears as if most tutorials leave out the topic of conditional probability in the first place and jump right into using concepts such as variance, t-value, confidence intervals etc., so our approach will hopefully close an important gap for those seeking further insight and a *clear intuition* on each of the concepts by themselves and in relation to each other (which also involves a linear model, which we will go through first too, before we get to a commonly used t-test etc.). In other words: the p-value is a conceptual composition, which however still begins with conditional probability / Bayes' rule.

If all of this still sounds a lot: The webcomic above shows that **hypothesis testing involves nothing you don't know or wouldn't do already** and we hope that our tutorial will leave you as surprised as we are, when realizing how easy the steps of testing a hypothesis can be represented by mathematical terms – without much of an effort and without losing any of its conceptual intuition and magic behind it.

1 Hypothesis testing

In general, every scientific study, every experiment, every process of 'testing' involves going through the following three steps in some way or another (even in descriptive statistics we explore data in relation to what we expect, the full argumentation is just not represented via mathematics):

1. formulating a **(prior) hypothesis** – that "what-ever something" is the case
2. **gathering (new) data** that is *related* to our hypothesis
3. **evaluating the results** (testing the hypothesis) and adjusting the prior hypothesis in order to better predict new data (**updating the hypothesis**)

So far this may not reflect on all the formulas and processes that pre-informed readers may expect, when performing statistical analysis. Nevertheless, it makes up the core of every statistical analysis in some way or another. Let's go through them in detail:

1.1 The prior hypothesis

In the beginning of every scientific evaluation there is a claim, i.e., a ***prior* hypothesis**. A prior hypothesis can be understood as a belief about the world, *regardless* any present or future experiences, or in other words: *before an experience was made that could prove or disprove the hypothesis* (the prior hypothesis can also be looked at as the *sum of all past experiences* on a hypothesis). An "experience" in statistics is often called an event. In general, an "experience" or an event is termed *data* in statistics, which represents events in the form of the outcome of measurements of any kind – measurements that have not yet been made to this point of inquiry.

Fig. 3 Ms. Miranda, hypothesizing in her shelter – they say, she knows-it-owl. A hypothesis for itself can be considered prior (a priori), when the data to evaluate the hypothesis with is *not* given yet. In this case the data would be a clear visual proof, as the 'whispering of snow' that Ms. Miranda perceives is too ambiguous and does not bring enough confidence to her mind. Original photo by Kevin Mueller.

1.2 Gathering evident data

The next step in any scientific investigation is a process of gathering experiences, observing events – **gathering *data***. This can be done in various ways. **However, there are certain constraints to what can be considered *data*, which we want to give some special attention here:**

The most important constrain – with which readers may be familiar with to some degree – is the constrain of events being ***evident***. The term *evidence*, as present in **'evidence-based medicine'**, is a science theoretic term and originated in its modern form from concepts such as phenomenology and science theory (e.g., Karl Popper) and became an institutional matter in the 1980s (also following further developments in statistics and scripted approaches to therapy (evidence-based guidelines etc.)). Roughly, *data* is considered to be *evident*, when it can potentially be experienced by *any human inference*, regardless or *independent* of their (prior) beliefs. Formally this can be expressed as *'intersubjectivity of human experience on an independent event'*, such as a ball falling to the ground, which is an event that as such is *not influenced* by my thoughts, or what I wish to happen (are *independent*). Such an event can therefore potentially be perceived by others, again: independent of my thoughts, beliefs, intentions (therefore 'intersubjective', and in order to address the 'tools' used to infer: 'human inference').

Another way evidence as an attribute is commonly expressed is by saying "*data is given"*. The phrase "*data is given*" is etymologically redundant, as *data* originated in the Latin language and also means "that what is given" – and stands in contrast to *what is set in advance*, i.e., our (hypo)theta, the prior hypothesis (the term *hypothesis* originated in Greek and means "to place under" – or *set before* in the sense of our temporal hierarchy of the three steps of scientific inquiry).

That's right, we are here for the evidence. Y'know what I'm sayin'?
Yeah, really nice... By the way - do you have proof for that by any chance?
. . .
My beliefs are sooo good – much better than yours. Believe me, it's so easy!

Fig. 4 Owls in disbelief – tirelessly in search for evidence based kn-owl-edge. Original photo by Tom Rogers.

The above terminological use of the term 'evident' may appear confusing, considering the everyday use of the term evidence: Note that the practice of saying a study to have shown evidence in a belief or hypothesis often implicitly jumps from the prior constrain on data acquisition to the interpretation of the outcome of the hypothesis testing in the mathematical sense. Both involve evidence, either as a constrain, or as a possible interpretation of statistical results. Though this is still partially overlapping with the concept of the *significance of a model* in an unfortunate way, as the significance is not the only marker for 'evidence' of any kind, as we will see. As the pandemic and developments in the recent years have often blurred the view on science massively, we believe that it is important to note here that ***evidence-based medicine*** is not just an advocacy on how to interpret the outcome of a (statistical) hypothesis test correctly (the significance of a model), but also a discourse on the evidence of inquiry as such. This approach to infer "on the world" therefore *explicitly stands in contrast* to mere belief systems of any kind, which may argue that *a belief for itself* leads to explicit knowledge of the world of some kind – e.g., the power, the effect, the impact, the existence of something – *regardless* any ("evident / given") *data* that could prove the conditional relation between the *hypothesis* and the *data*. Such a violation of the steps of scientific inquiry can in some way intuitively be understood as *tilting the temporal course of scientific inference* in the sense that *a prior hypothesis claims to already be the result of the evaluation of data inquiry*, and may even claim to be the *data itself*. **There are other constraints**, such as the validity of a test (does a test really test what it is supposed to test for) – however, for now these can be looked at as just further reflections on the same issue: *gathering and inferring on evident data* in a wider sense.

Fig. 5 Ms. Miranda, on the way to gather some data – wise enough to always challenge her beliefs. Original photo by Alfred Kenneally.

In comparison to our first step, consisting of our prior hypothesis only, **the second step of "gathering data" can now be looked at as an actual or *present relation* between the *data* and our (hypo)*theta***. In probability theory this relation can be looked at logically as a *conjunction*, i.e., an "overlapping" between our hypothesis (*theta*) and the obtained *data*, where the overlapping indicates that both are *true*. True? This may be a little abstract, but the attribute *true* just says that when our prior hypothesis says "it is raining tomorrow" and the *data* shows that "it rained" the next day, then the event "it is raining" is *true* in both cases of *theta* and *data*. For *theta* anything we want can be pre-set as *true* – not so for *data,* as *data is given* in the sense that we have to gather it, make experiences, observe an event.

Those with prior knowledge may recall that the so-called *null hypothesis* refers to a case where the hypothesis is set to be *false* (\overline{theta} or {theta}^{c}, where the "c" stands for complement). Data on the other hand usually refers to a difference in the mean via a t-test (again, entialed in the third part of this series, after we have discussed linear regression models; in nuce though a t-test can be thought of as evaluating a signal to noise ratio, given a threshold (confidence interval)).

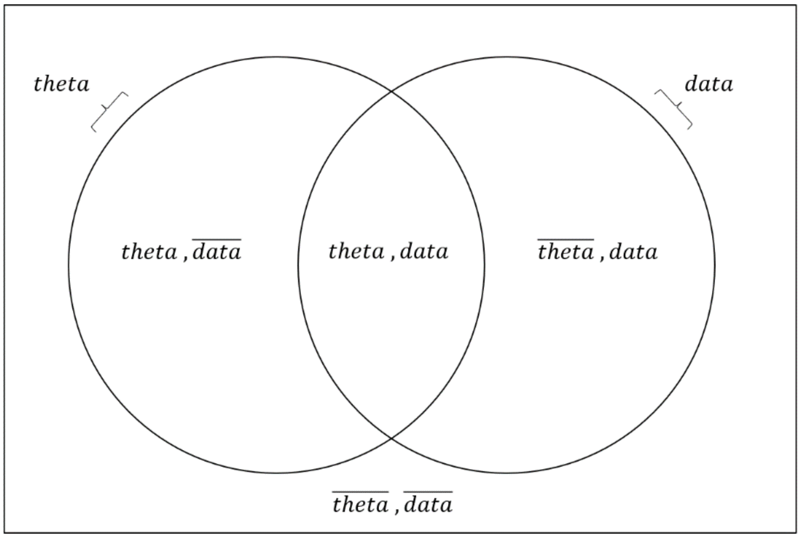

$theta$
$data$
$theta\ ,\overline{data}$
$theta\ ,data$
$\overline{theta}\ ,data$
$\overline{theta}\ ,\overline{data}$

Fig. 6 A so-called *Venn diagram*. The circles each represent *theta* and *data* for itself as 'areas' (compare set theory for details). The overlapping indicates that both theta and data is true. The comma "," in "theta , data" is here considered a mathematical symbol and is spoken "and".

However, in probability theory the conjunction between data and theta is also called a *joint (probability)*. Later we will add actual probabilities to the Venn diagram, suggesting that for any of the possible 'joint combinations', there is a certain probability ranging from 0 to 1 assigned to it. Technically this joint probability is what can be called a *probabilistic model* – a model of the relation of *theta* and *data*. However, our linear model in the next part of this tutorial series will mathematically ***not be*** exactly the same, as it is "not made out of" probability values, but out of the values of a measurement (e.g., tea drank within time). This is where hypothesis tests in the mathematical sense come in play, as a t-test can be seen as a method to obtain values (such as the t-value) that can be used to obtain a p-value for a linear model. In other words: there are ways to "look at" or evaluate the results of a linear model under the consideration of probabilistic relations.

We will get back to all of this in detail soon. For now, just hold on to the idea of an overlapping of our hypothesis and data that fits to it in the sense that the assumptions hold in both "areas" – *theta* and *data* both being *true* or 'the case'.

1.3 Evaluating and adjusting the hypothesis to fit the data

The consequence of gathering *data* – or "making experiences" – is usually that **we evaluate and eventually adjust our hypothesis to (better) fit the actual experience**, the actual *data*. We have to say "usually" as it is unfortunately not a "self-evident" practice, given a high tendency to produce positive results only in science ('publication bias', resulting in a lot of studies not even getting close to the third step of statistical inference, which is unfortunate in a lot of ways).

The result of our evaluation will eventually become our *new prior* hypothesis. The better our hypothesis, the better we are to predict future events. Note that adjusting or updating the hypothesis in the mathematical sense concerns the probability of a hypothesis, as we will see in the next chapter, *not* inventing a completely new thesis (e.g., no changes in the variables or so, which could be referred to as HARKing (**H**ypothesizing **A**fter the **R**esults are **K**nown), which is essentially faking a course of inference). However, in a wider sense, updating a hypothesis in terms of changing our beliefs still represents what we do as a consequence of gathering and evaluating data in the long run: we change/update our model of the world, gathering new insights over time (Bayes' rule therefore represents "learning", as we will see).

Fig. 7 Ms. Miranda has published her results maximum *open access*, so her insights were spread fast in the community. Ms. Miranda was right with her hypothesis. This is not the case with every hypothesis we belief to be true, keep in mind. For further insights into open science as a default practice, we recommend our position paper on Open Research by Default, written by Raphael Leuner, Clara F. Weber, and Jonas Stampka. Original photo by Ray Hennessy.

To give you another simple non-feathery example of what updating a hypothesis means, given a *negative* result: if we were to experience a ball to *never* fall down by itself, we would adjust or "doubt" on our hypothesis of gravitation in the long run.

Now that we have roughly gone through all the steps of hypothesis testing, let us look at them from a mathematical perspective.

We will leave a summary at the end of every chapter, to give you an overview of what we have learned so far:

> We have learned…
>
> … that the general structure of **hypothesis testing** consists of **three elements**:
>
> a) A **prior hypothesis** *theta*, regardless any (new) data to prove the theta – or understood as in relation to the *sum* of all *past* data, all *past* experiences, all *past* events.
> b) The process of **gathering data** and **forming a *relation*** or overlapping **between theta and data.**
> c) This relation is then evaluated and **we subsequently update our prior hypothesis to better fit the data** – to better fit the world we are trying to understand.

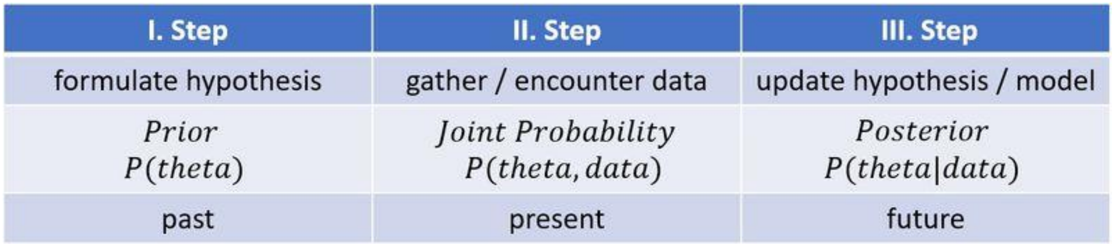
2 Conditional probability and Bayes' theorem

Translating our discussion above into mathematics is fairly easy. There are only two *minor* features we have to include into our intuition on hypothesis testing to make it work smoothly.

1. **Categorical variables:** The binary distinction *true* and *false* can be looked at as categorical distinction. Other categories are also possible, such as *heads* or *tails*. In the latter cases it depends on what you chose first as $theta$, either heads or tails, to translate the coin into a binary. However, there is no restriction to binary categories. The categories that were chosen are for themselves contingent, but of course still depend on pre-set decisions *we make* (e.g., the number of categories). Below we will mostly work with binary outcome options.
2. **Probability values:** Our confidence in a hypothesis will be represented in the form of probability values. Before we were just working with the categorical distinction of *true* and *false (which could be represented by a 1 and a 0 only)*. Each of them will now just get assigned a probability *between* 0 and 1, making inference a lot more dynamic. This is importantly *not the same* as the binary distinction *true* and *false*, as these are just *categories*, such that there could be a 0.1 probability assigned to *theta*. However, for the addition of probabilities to our variables to be recognizable, our variables *theta* and *data* will now be *symbolically marked* by a *P* for 'probability' standing in front of them – such that our prior hypothesis, i.e., our hypothesis regardless any (new) data, will be written as $P(theta)$.

To get an intuition on probabilities in general: we could theoretically assign everything we experience a probability, especially as a very 'prior relation' we have with the world in order to orientate ourselves within it is time and space (fitting well to inference in the sense of learning and inference as a form of making better guesses (predictions) or 'bets on reality' over time). Only if something were to never change, it would be assigned a 100% chance to be the case and such an estimate would involve infinite observation, if someone wants to be certain over a wide timespan as well. So one general way of looking at probability is as a representation of a frequency of events, e.g., in a certain context of time for example.

Now that we are set, we can simply go through the same three steps again, including the mathematical symbols used to address what we have gone through conceptually already. It is also important to highlight the simple and intuitive temporal hierarchy, corresponding the steps.



I. Step	II. Step	III. Step
formulate hypothesis	gather / encounter data	update hypothesis / model
$Prior$ $P(theta)$	$Joint\ Probability$ $P(theta, data)$	$Posterior$ $P(theta\|data)$
past	present	future

I guess the mathematical denotation for the prior and the joint probability does not add much to what we have gone through so far, just the third step, the *posterior probability* should be symbolically new to us. Outspoken the posterior probability is read as "the probability of theta, given or under the *condition* of data", where the sign "|" means *given* or *under the condition of.* **Therefore the name: conditional probability**. The temporal course from theta to data (our condition) may appear inverted to you, and this is partially true! The posterior probability actually refers to the data being a '*prior condition'* to the hypothesis this time, as the ***posterior*** probability reflects *the hypothesis,* ***after*** *we have observed data*. Before we argued that we *start with a hypothesis* and *then gather data*, which was the transition from the first to the second step, not the second to the third step, as we will see. The posterior argues for what we will believe to be true in the **future** (as a new prior for further inference), the joint reflects the **present** overlapping / encounter with data and the prior the **past** (the sum of all past or possible observations, also see 'sum rule' later below).

Let us go *one step back* and fully forget about the posterior for now, in order to focus on what a joint probability actually is. So far, we referred to the second step as an overlapping or a *joint* between $theta$ and $data.$ Both, the joint probability and the conditional probability reflect a relation between two variables. However, the **difference between a conditional probability**, such as the posterior, **and the joint probability though is simple**: A joint probability is considered a probability where the specific course of conditions *is not yet defined or decided*, such that either "$theta$ as prior condition for $data$" or "$data$ as prior condition for $theta$" could be obtained from evaluating the joint (note that this refers to a prior only in terms of our temporal hierarchy, as the prior is usually denoted $P(theta)$).

The joint probability – at least the way it is denoted above – therefore reflects the probability of encountering theta *and* data in general, *not* under a particular condition of our temporal course of inference, so to speak. In another words, the course of inference is ***potentially bi-directional***: we could retrieve $P(theta|data)$ or the inverted $P(data|theta)$ when evaluating the same joint probability for itself. The conditional probability we end up with depends on the variable we chose to start with – either with $P(theta)$ or $P(data)$. Other than that, the steps are the same, also consistent with our 'temporal hierarchy', namely: starting with a "single probability" that is always "prior" in a wider sense, then moving to the joint, ending in a conditional probability. The order of variables is therefore not important when denoting joint probabilities, such that $P(theta\ , data)$ and $P(data\ , theta)$ are the same (again, the comma is in this case a mathematical symbol spoken "and" and refers to the overlapping, the conjunction between data and theta).

There is still another way of representing and especially calculating the joint probability **– and this is where Bayes' theorem comes in play**:

Note that Bayes' rule and conditional probability are essentially the same. However, classic conditional probability does not reflect the joint probability by itself as the result of a *weighted conditional probability* that can also be represented via a decision tree, as we will see below (this is – as much as I know – the only real difference, apart from the historic remarks referring to Thomas Bayes, which we will not get into here). Bayes' rule just extends conditional probability on that matter, so to speak. Note that we will in general not work with the complement of data, as making no observations or making observation other than the ones we actually make so to speak is not what we can logically achieve when doing inference (there is another more mathematical answer to this, also concerning set theory and the Venn diagram, but we will not go into further details in this tutorial).

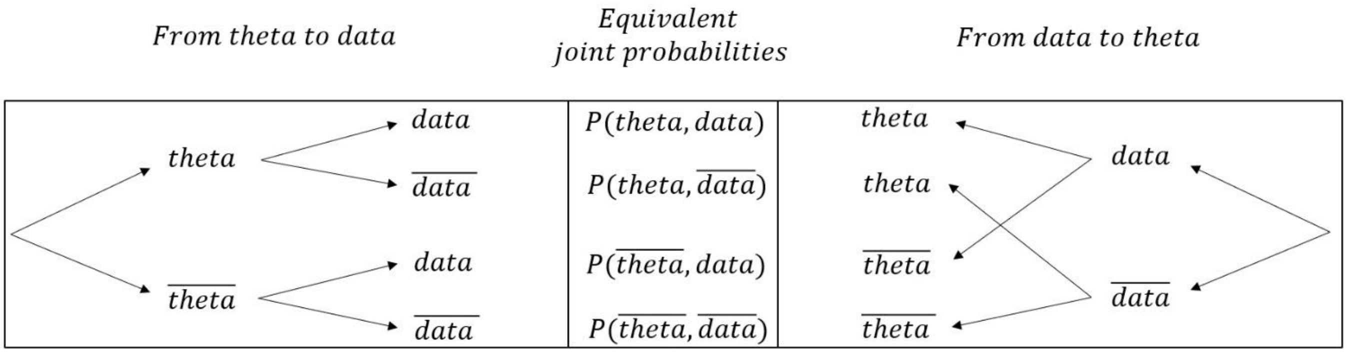


Fig. 8 Simplified decision tree, leading to joint probabilities. The figure just visualized what we learned so far: the order of the inquiry of variables is not reflected in the above *denotation* of joint probabilities. However, the decision trees still imply that there are ways to represent or ***decompose the joint*** as the result of *one specific course of inference* only (e.g., starting with *theta*, as in our basic intuition on hypothesis testing).

To get a closer look at **one specific way of obtaining the joint probability**, let us zoom in into one particular *chain of decisions* made.

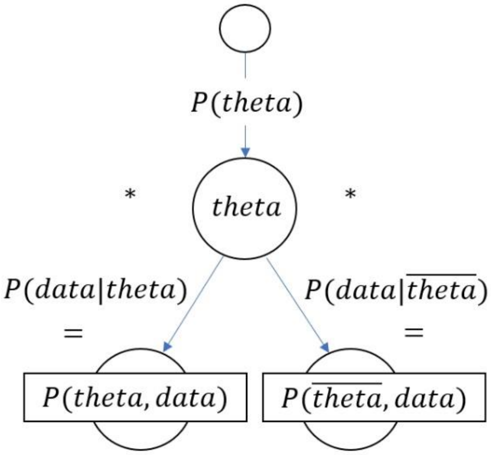


Fig. 9 This graph reflects the same as the upper left part in the previous figure above. The graph also adds the mathematical representation of our particular *chain of decisions*. The decision tree essentially refers to what is called the *chain* or *product rule of probability*. The above also entails a newcomer, $P(data|theta)$, which is also referred to as the ***likelihood,*** and is – as we know – just the inverse condition of our posterior probability $P(theta|data)$, i.e., the data under the condition of or *after* we formed a hypothesis.

With the chain rule we have introduced two things: a mathematical algorithm to calculate the joint probability from one course of inference only, and the mentioned likelihood $P(data|theta)$ – another *conditional* probability.

To make sense of the likelihood: Remember Ms. Miranda, when she was on her way to make new observations? This is what the **likelihood** is actually about: gathering data, *under the condition* of theta, **moving from step I to step II**. The evaluation of the joint eventually resulted in the **posterior**, representing the inverse condition, **moving from step II to step III** – the probability of theta, *after* we observed data.

What Bayes' rule does mathematically is what we literally just have gone through conceptually: the process of hypothesis testing more or less inverts the likelihood to become the posterior. "More or less" as this is numerically only the case under special prior conditions, as we will see, but it is still true: Bayes' rule is method to invert *the experience under the condition of a hypothesis* to become *the hypothesis under the condition of (new) experiences* made (which eventually becomes the future prior and therefore influences or changes the next joint probability formed with the new prior, i.e., the model is updated).

To wrap this up: The chain rule above shows that the joint probability not only consists of the likelihood, moving from step I to step II, but also of the prior itself, which can importantly be understood as a *weight* in the form of an expectation on the likelihood. The likelihood again represents the observation, the experiences we make, our *data, given our hypothesis* (or the prob. of data *after* we formed a hypothesis). But what does *weighting* actually mean? The concept is actually simple: something can be weighted to have a higher probability of occurrence in general (over time) as something else, nevertheless the observation (regardless any data!). In other words: a model can entail a prior "confidence" in a hypothesis which may be higher than the prior confidence into its *complement* and eventually influences how an event is evaluated *a posteriori (after making an experience)*. If this sounds abstract, think of 'classic conditioning' in psychology as a classic example, where a conditioned weight, i.e., a prior expectation, influences the behavior/inference of a subject (build up on weighted inference) in relation to certain events (the behavior is therefore different to unconditioned subjects over time; in other words: classic Bayes' rule represents a "learning algorithm").

Note up front that a probability of .5 for theta and .5 for its complement refers to a special kind of "balanced prior" weight, which is also referred to as "uniform" prior – we will get there soon.

In order to fully make sense of all of the above, we can now finally discuss Bayes' rule via actual mathematical formulas. Let us start with the fact that there are two approaches to obtain one and the same joint probability, since:

$$P(theta|data) * P(data) = \ P(theta, data) = P(data|theta) * P(theta)$$

On the far-left we see the posterior multiplied or weighted by $P(data)$. On the far right we see the likelihood weighted by the prior. Both multiplications lead to the same joint $P(theta, data)$.

Above we can see how Bayes' rule can be understood as a way to obtain a joint probability *from* (weighted) conditional probabilities. However, Bayes' rule is essentially about a specific conditional probability: the posterior. Let us first take the simplest route to Bayes' theorem just by taking basic rules of equations into consideration. In order to obtain the *posterior* probability from the formula above, we could just divide the whole line of equation by $P(data)$ to get rid of the term on the far-left side of the equation:

$$P(theta|data)P(data) = \ P(theta, data) = P(data|theta)P(theta) \quad => \ \div P(data)$$

$$P(theta|data)\cancel{P(data)} = \ \cancel{P(theta, data)} = \frac{P(data \mid theta)P(theta)}{P(data)}$$

Just in case it appears confusing, the joint probability is crossed out too, as the result of our actions leaves us with a *conditional* probability only (deciding for the course of inference). The last line eventually represents what is called **Bayes' rule.**

An easy way to remember Bayes' rule is to follow the variables clockwise, starting with the posterior, and chant them, going: theta-data // data-theta // theta --- data (with a little break in-between the prior and $P(data)$, which stand for themselves). As it just repeats an inversion of the pair theta and data, it is not too hard to remember (also conceptually).

As mentioned, the posterior can be understood as the result of *the joint being divided by the weight* of $P(data)$, a kind of "counter weight", so to speak. To get a better intuition what this means, let us go back to our Venn diagram.

Our prior $P(theta)$ will represent a weight to the likelihood, resulting in a joint probability. The variable $P(data)$ on the other hand, also called the *model evidence*, works as a kind of counter weight to the joint as a whole (not just the likelihood), resulting in the posterior probability.

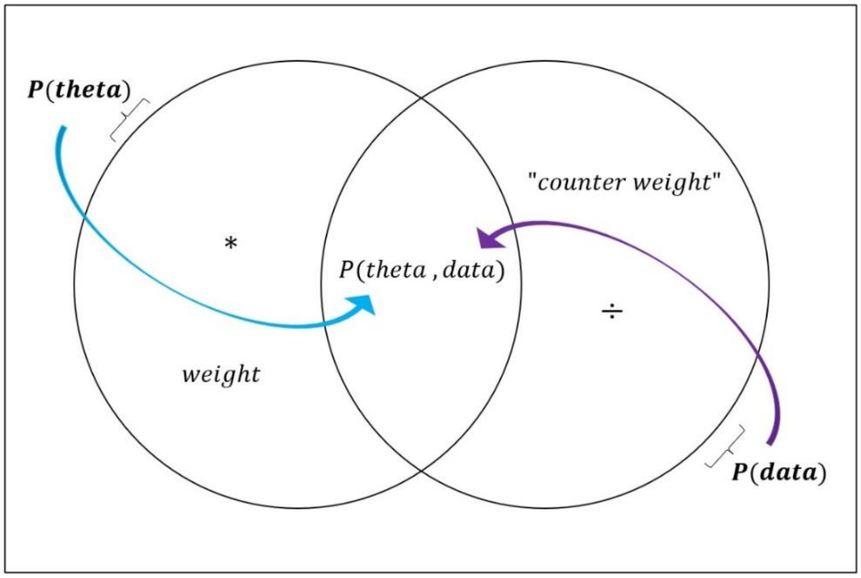


Fig. 10 The optimized Venn diagram suggest that when our prior $P(theta)$ acts as a weight to the *likelihood*, resulting in $P(theta, data)$, then the model evidence, $P(data)$, will work as a kind of counter weight on that joint and will leave us with the conditional probability $P(theta|data)$. The mathematical representation of this process is referred to as normalization. Numerically this process is necessary, as the joint probability has the attribute that it does not sum to 1 for itself. We will get there when we do some simple calculations via R.

Another way of reflecting on the joint probabilities is via a conditional probability table, which also reveals a way how we can obtain $P(B)$ or $P(A)$, given the joint probability:

	A	$\bar{A}$	$\sum_A$
B	$A \cap B$	$\bar{A} \cap B$	$P(B)$
$\bar{B}$	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	$P(\bar{B})$
$\sum_B$	$P(A)$	$P(\bar{A})$	1

Table. 2 Conditional probability table (CPT). Here we will adapt the classic theme for a change, i.e., a relation between *A* and *B*, instead of *theta* and *data*. Note that "∩" is equivalent to ",", i.e., spoken "and". The content of this table is just another way to reflect on our upper *Venn diagram*. The table also indicates that $P(A) = P(A,B) + P(A, not - B)$, i.e., is the sum of the *joint probabilities* between given and not given the B that fits the A – the process is also referred to as 'to sum over B' (or 'to sum out B'). Mathematically this represents the application of the *sum rule of probability* – which also exactly represents our linguistic definition of a prior: theta as the sum of all past / possible data ("all possible" refers to data and not-data). This process also reflects in the sum sign "Σ" (capital sigma) in the CBT. The model evidence $P(B)$ is also called the ***marginal likelihood***, as it is the summed-out relation of the *likelihood* $P(B|A)$ and *prior* $P(A)$ and the *complementary*

joint including not-A. Here we sum over all A. This process is also referred to as **marginalization**. In a CPT $P(B)$ is located in the tables *margin* – therefore the name (I have also encountered the term 'average likelihood' for the mod. e. once; the prior is theoretically a "marginal posterior", or just the marginal probability of A).

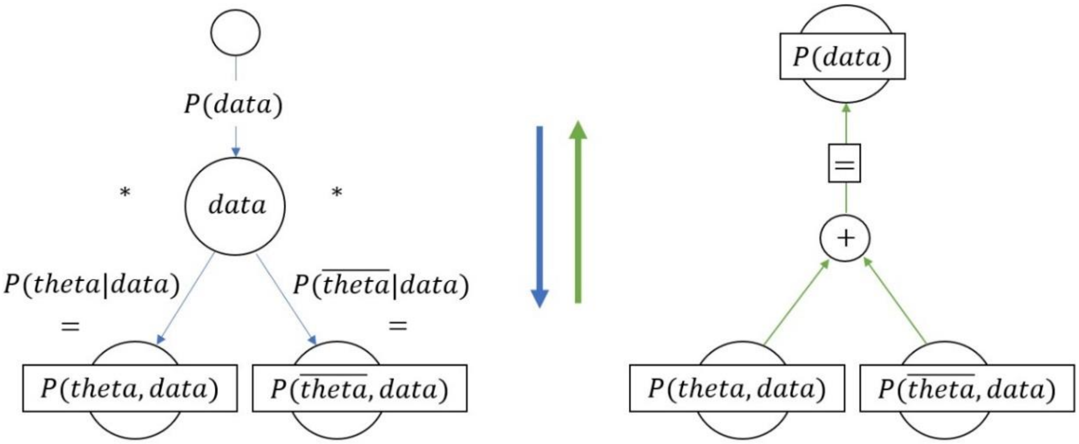

$P(data)$
$data$
*
*
$P(theta|data)$
=
$P(\overline{theta}|data)$
=
$P(theta, data)$
$P(\overline{theta}, data)$
$P(data)$
=
+
$P(theta, data)$
$P(\overline{theta}, data)$

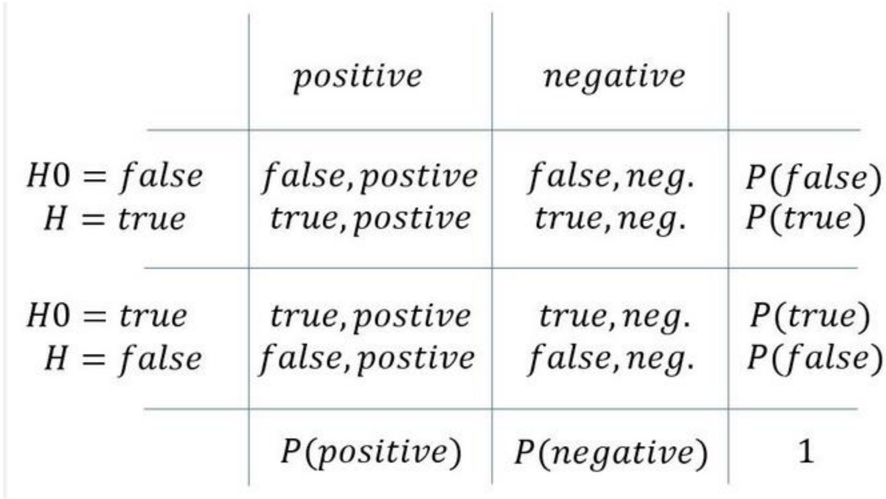
Fig. 11 This figure demonstrates that the sum rule reverses the actions of the chain rule (our decision tree). Note that the figure above this time refers to a different path of inference, a different chain, starting with $P(data)$ this time. The sum rule leads to $P(data)$ only, as data is set "constant" so to speak. If the variabel data was treated "dynamically" and theta constant, it would result in $p(theta)$.

Note that $P(B)$ is a value that can also be provided, in order to correct or compare it with the assumed model evidence of just one event (or the events a model works with). You may have already found respective examples during your own research, where the model evidence is not calculated via summing out, but provided. A common example is weather forecast: What is the probability of tomorrow being sunny, say given that it is rainy today. In mathematical terms this would be denoted as: $P(t + 1 = sunny\ |t = rainy)$ following the scheme of $P(theta|data)$. In such a case of meteorology the general probability of a day being rainy, i.e., $P(rainy)$, may have been externally obtained from an average on the data from a long-range observation and is not just calculated from the information given by the prior and the likelihood of just one or a few events (note that statistical thermodynamics / mechanics relies on the same idea of conditional probability, so it totally makes sense to predict weather with such a probabilistic model; we will provide tutorials on thermodynamics and information theory in the near future as well!).

Apart from that, Bayes' rule is in general used to test a test, i.e., the evidence of a test – how well it predicts. It does so by considering a longer range of observations over time, or a wider population in the form of a general frequency of $P(Event)$, and again not just the one or few events a model works with. The posterior can in such cases act as the positive predictive value (PPV), updating our confidence in a test. In essence such calculations rely on *two conditional dimensions*: the *actual data (set as hypothesis)*, i.e., true and false, and the *predicted data*: positive and negative. The model evidence is here more clearly used to reflect on the evidence of a statistical model, i.e., our joint probability. The far-right refers to probabilities of the joint and the model evidence, to clearly relate it to Bayes rule', even though the middle part of the equation does not yet indicate probabilities (is not *normalized* in this case), but the number of events of a certain kind.

$$\text{PPV} = \text{P(True|Positive)} = \frac{\text{Number of true positive}}{\text{Number of true positive} + \text{Number of false positive}} = \frac{\text{P(True , Positive)}}{\text{P(Positive)}}$$

A CBT for the binary dimensions "true/false" and "positive/negative" is also referred to as a *confusion matrix*. Such a confusion matrix can in a lot of cases be literally confusing. Recall that the classic p-value is mostly referred to as $P(data|\,\overline{theta})$, where $\overline{theta}$ is referred to as the null hypothesis. The aim is to discard the null hypothesis, in order to keep the alternative hypothesis (which is the hypothesis we actually set). However, operating with the distinction "alternative and null hypothesis" is often adding an unnecessary redundancy of binary dimensions, e.g., when saying that the null hypothesis is considered to be true, which is the same as the alternative hypothesis being false (it tilts around, due to redundant binary distinctions). Keep that in mind, when looking for the PPV or the false discovery rate ($1 - PPV$ for false positive), where "false" could be tilted in the sense that it actually relates to the alternative hypothesis being false and positive (significant), *not* the null hypothesis being false and significant. We believe that the redundant *ex negativo* approach is one reason why people fail in understanding p-values and conditional probability in inferential statistics in general, even though it just represents the steps of an argumentation we make.



	$positive$	$negative$	
$H0 = false$ $H = true$	$false, postive$ $true, postive$	$false, neg.$ $true, neg.$	$P(false)$ $P(true)$
$H0 = true$ $H = false$	$true, postive$ $false, postive$	$true, neg.$ $false, neg.$	$P(true)$ $P(false)$
	$P(positive)$	$P(negative)$	1

Table 3. A confusion matrix for H (alternative hypothesis) *and* H0 (null hypothesis). Note that this table has the same general relational structure as a regular CBT.

That was it, **you have now mastered the essentials of hypothesis testing reflected as conditional probability and Bayes' rule respectively,** both on the conceptual and on the mathematical level. In the next chapter we will do some calculations to gain some experience with applications of the math that we have learned so far using R. After that we will introduce the difference between the Bayesian and the frequentist interpretation of conditional probability aka Bayes' rule and finally reveal what a *p-value* actually is (we have indirectly encountered it before).

Below you see the posterior on the far-left side of the equation, the classic formula to obtain such a conditional probability as $P(theta, data)$ in the middle and Bayes' rule on the far-right side of the equation (including a method to obtain the joint, given prior (our hypothesis for itself) and the likelihood in terms of conditioned observations).

$$P(theta|data) \;=\; \frac{P(theta \,\cap\, data)}{P(data)} \;=\; \frac{P(data \mid theta)\, P(theta)}{P(data)}$$

We have learned…

… that the three steps of **hypothesis testing** can be directly translated into conditional probability, i.e., Bayes´ theorem and carry a temporal hierarchy in their names, referring to a course of inference (starting with either $P(theta)$ or $P(data)$, ending with the respective conditional probability).

I. prior => $P(theta)$
II. joint probability => $P(data|theta)P(theta) \quad \coloneqq \quad chain\ rule$
III. posterior => $P(theta|data)$

… that there are several ways to represent the logical relations within conditional probability, either with a CBT, via decision trees or a Venn diagram.

… an easy way to remember Bayes' rule: follow the variables clockwise, starting with the posterior, and repeatedly chant the variables, going:

theta-data // data-theta // theta --- data

$$P(theta|data) \;=\; \frac{P(data \mid theta)\, P(theta)}{P(data)}$$

… about the minor difference between conditional probability and Bayes´ rule, the latter referring to a method to obtain joint probabilities from weighted conditional probabilities.

3 Computing conditional probability / Bayes´ rule in R

After elaborations on the concept and the mathematics behind hypothesis testing (its mathematical logic of inference), we will now do some calculations in R for a change (the application of the mathematical logic). This will help us to get a clear orientation on what the results of the upper formula may look like and we will also numerically prove some assumption we made above. **Note that using R will be much *easier* than it might sound, as we will be using it more as a simple calculator for now (note, all the following can also be done 'by hand' – also holds for the linear regression, as we will see).** We will also introduce R in the second part of this tutorial series again, which also involves things as plotting – so no worries if this either appears too much, or too basic to you for now.

Below you will find code for the programming language R that can be used to calculate the posterior probability. Note again, **if you are new to R** you can either just read this tutorial and with it read the code – as we will provide the output for every line of commands – or you **just download R and**

RStudio, open a new script and copy and paste the code provided below into it, or download and open the R script we provided below (or type it in yourself, if you wish to do so).

R script, corresponding to the tutorial:

Inferential_statistics_I.R

The first lines of our script will be just a test and looks like this:

```
# This is a test, which will also be the name of the 'object'
test = 2 + 5 # Execute this line!
```

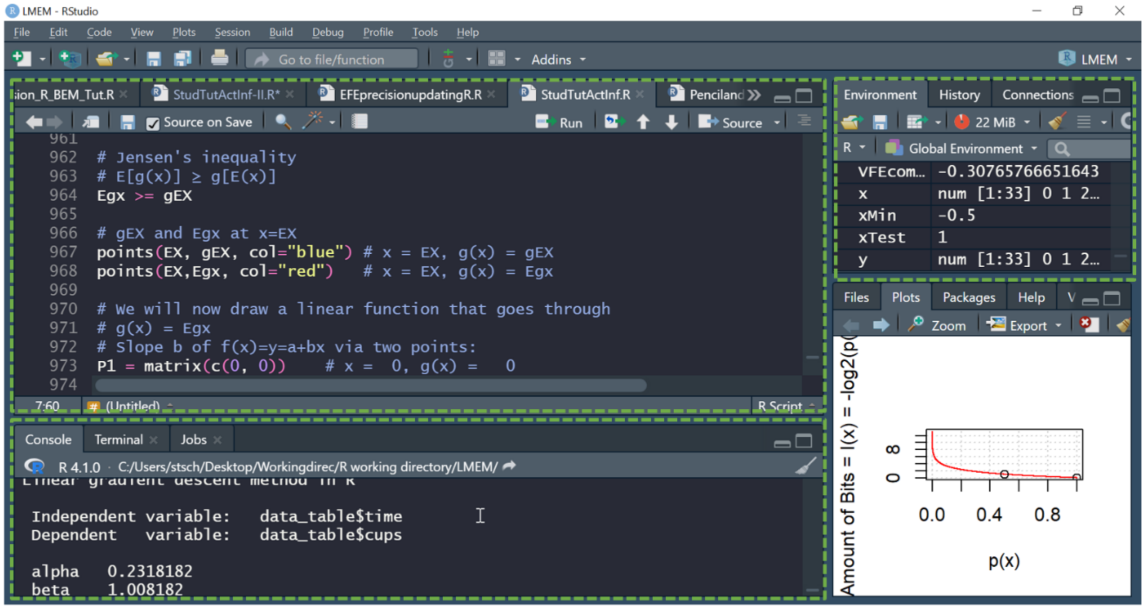
Note that those lines that start with #, as well as any code *within a line after* a # was placed, is considered a "comment" and will not be 'understood as code' by R (so you can also mark and execute it and it will not mess up anything). Otherwise text will be interpreted as code, such as a calculation – which will lead to (mostly enigmatic) errors presented in the console below the script (see figure below for details on interface of RStudio).

Mark the lines you want to execute and **press ALT+ENTER**. **In case you are using Apple hardware, use the command key instead of ALT.** We also recommend using the keyboard when operating within the script: use **SHIFT+ARROW** to mark and demark letters (left/right) or whole lines (up/down).

You can also execute comments, so if a script is set as a whole, one can also mark all and then execute the script as whole. **The result** can be seen in the *environment tab* on the upper right side of RStudio (see figure below). If you ever feel that your script is presenting a funny output (especially after a series of errors), **clear the environment via the brush tool** – there is another brush-icon placed in the console tab to clear the console. **Now mark the name of an object only and press ALT+ENTER again to obtain the results in the console** (below the script) – you won't need to know much more for this tutorial for now, believe us!

The consequence of your actions should result in the following console output (ignore the [1] for a moment).

```
# Console output:
# [1] 7
```



```
# Jensen's inequality
# E[g(x)] ≥ g[E(x)]
Egx >= gEX

# gEX and Egx at x=EX
points(EX, gEX, col="blue") # x = EX, g(x) = gEX
points(EX,Egx, col="red")   # x = EX, g(x) = Egx

# We will now draw a linear function that goes through
# g(x) = Egx
# Slope b of f(x)=y=a+bx via two points:
P1 = matrix(c(0, 0))     # x =  0, g(x) =   0
```

```
Linear gradient descent method in R

 Independent variable:   data_table$time
 Dependent   variable:   data_table$cups

 alpha   0.2318182
 beta    1.008182
```

```
VFEcom…  -0.30765766651643
x        num [1:33] 0 1 2…
xMin     -0.5
xTest    1
y        num [1:33] 0 1 2…
```

Fig. 12 RStudio is an IDE, i.e., an integrated development environment, in general used to ease the use of R – in other words: it is a partial interface for R. For this tutorial you won't need to understand much of R or RStudio in order to follow and execute the script. On the **upper left** you will find the window for the script – *containing unrelated content*. On the **upper right** side, you will find the *environment*. It keeps track of what you executed. You can also find the results in the *console* output, which is usually to be found on the **lower left** side. To also obtain the result of an object in the console, *after* you have executed the whole line, such as test = 2 + 5, mark the name of an object *and the name only* (here "test") and execute (*again, the whole line has to be executed first*). The appearance of RStudio can be changed via *Tools -> Global Options -> Appearance*. The upper theme I am using is called "Dracula".

Note that we can represent our probability variables *either as single value or as a probability vector that sums to 1*, such that we will always consider theta and its complement at the same time. We will start with using single values first.

The following example will refer to anything you consider a *theta* and for which (evident) data is given. For a simple example, we will start with flipping a coin, arguing that the coin is fair, such that a 50%/50% chance is given to encounter either heads or tails (or 0 or 1; true or false…). **As this is an example, we can provide ourselves with data and set the likelihood for ourself.** Recall that we need the likelihood to calculate the joint probability via Bayes' rule and that the likelihood is a conditional probability for itself – reflecting on the transition of step I to step II within hypothesis testing (I. forming a hypothesis. II. making (new) experiences / gathering data). The likelihood can therefore be reflected as the probability of data, *after* we have formed a hypothesis (in contrast to the inverted posterior, reflecting moving from step II to step III: – expressing the probability of the theta, *after* data was obtained).

$$Prior \ := \ P(A) = .5, \ \ Likelihood \ := \ P(B|A) = .5$$

```
# Define you prior, e.g., .5 for heads.
# Note that R is a case sensitive language ("prior" not same as "Prior").
prior = .5
```

```
# Likelihood
likelihood = .5
```

Now we can define the likelihood corresponding to our prior hypothesis. It can either be .5 again, which would represent a truly fair coin (after some rounds of flipping the coin) – and suggests that our hypothesis holds. Or you assign a value of .6 or any other value deviating from .5, which would suggest that the coin is phony and that our prior assumptions about the coin are wrong.

Either way, we can now calculate the joint probability, as well as the model evidence:

$$Joint\ probability := \ P(A,B) = P(B|A)P(A) = \ .5 \ * \ .5 = \ .25$$

```
# Joint probability:
joint = prior*likelihood
# Output:
# [1] 0.25
```

Now we can calculate the model evidence. Recall that the mathematical definition is as follows:

$$Model\ evidence := \ P(B) = \sum_{A} P(A,B) \ = \ P(B|A) * P(A) \ + \ P(B|\bar{A}) * P(\bar{A}) \ = \ .25 + .25 = \ .5$$

```
# Model evidence (note that R does not allow spacing within names!):
model_evidence = .25 + .25
```

The color marking refers to the result of each of the joint probabilities – the sum of both results in $P(B)$. Also recall our CBT in that respect – its first row:

	A	$\bar{A}$	$\sum_{A}$
B	$A \cap B$	$\bar{A} \cap B$	$P(B)$
$\bar{B}$	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	$P(\bar{B})$
$\sum_{B}$	$P(A)$	$P(\bar{A})$	1

Table 3 Our CBT for comparison. P(B) is on upper right margin and represents the *sum* over all possible A (applying the sum rule of probability) – the sum of the first row of joint probabilities.

We can now calculate the posterior:

$$Posterior := \; P(A|B) = \frac{P(A,\; B)}{P(B)} = \frac{.25}{.5} = \; .5$$

```
# Posterior:
posterior = joint/model_evidence
# [1] 0.5
```

In this case the prior hypothesis was *not really* updated, such that *prior and posterior remain the same*! What if the likelihood or prior changes?

To get a better overview of the process above, we are now going to slightly expand the math: Next we are not only using probability values, but probability *vectors* to do Bayes', which lets us calculate both $theta$ and $\overline{theta}$ at the same time. To represent our vectors in R, we will use the combine function **`c()`**, with which objects of any kind can be combined as a list of values, objects or even a list of lists (note that there is also a function called **`list()`** as well, which is not structured in rows and columns, but sequentially numerates elements – also of various kinds of classes (vector, matrix, single values, whole data frames, all in one list)).

However, the upper formula and code using probability vectors *just slightly differs* from what we have gone through so far. Note that for non-binary outcomes the vector would just be expanded holding 3+ values. The prior probability distribution you see below is also referred to as ***uniform prior* (will be important in a bit)**. Let us first go through the formulas and then do some computation with R. **Play around with the input values and you might notice a special characteristic of the posterior for yourself, when changing the values of the likelihood, but keeping the prior equally distributed (uniform).** The formula below formally does not include the complement, but it does so by using the probability vectors. Keep in mind that every vector has to sum to 1, when changing the input values below. **Also keep in mind that if you change a line you have to execute the respective line and every line involved *again* to obtain the new results.**

$$Prior \; := \;\; P(A) = \begin{bmatrix} .5 \\ .5 \end{bmatrix}, \;\; Likelihood \; := \;\; P(B|A) = \begin{bmatrix} .5 \\ .5 \end{bmatrix}$$

$$Joint\,probability := \;\; P(A,B) = P(B|A)P(A) = \begin{bmatrix} .5 \\ .5 \end{bmatrix} * \begin{bmatrix} .5 \\ .5 \end{bmatrix} = \begin{bmatrix} .25 \\ .25 \end{bmatrix}$$

$$Model\,evidence := \;\; P(B) = \textstyle\sum_A P(A,B) = \; .25 + .25 = \; .5$$

$$Posterior := \;\; P(A|B) = \frac{P(A,\; B)}{P(B)} = \frac{[.25\;.25]}{[.5]} = \begin{bmatrix} .5 \\ .5 \end{bmatrix}$$

```
# Define prior
prior = c(.5, .5)

# Likelihood
likelihood = c(.5, .5)

# Joint probability
joint = likelihood*prior
```

```
# Model evidence
model_  evidence = sum(joint)

# Posterior
posterior = joint / model_evidence
```

Above we have calculate the `model_evidence` using the `sum()` function. Basically, this function does what is says and uses an input, such as our joint probability vector with two values – .25 and .25 – and sums up every element of that vector (we will further get into what functions are in general in the next part of this tutorial series). **Keep in mind that an R script is executed from top to bottom.** The R script we provided can theoretically be executed as a whole (mark all and execute). However, it may be that a variable, e.g., with the name **prior** gets *redefined* in lower parts of a script (changes in the environment!). In other words: The content of the previous object with the same name **prior** will be "overwritten" so to speak.

Have you figured out what happened, when using a uniform prior, changing the likelihood only? You are right! *The posterior will always match the likelihood*. How is this possible? The reason is that in such a case the weight and the "counter weight" eliminate each other. Let us take a look at the formula to understand what that means mathematically:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = \frac{[.6\ \ .4] * [.5\ \ .5]}{[.5]} = \frac{[.3\ \ .2]}{[.5]} = [.6\ \ .4]$$

The above can also be simplified to the following, where $P(A)$ and $P(B)$ eliminate each other under certain conditions:

$$P(A|B) = \frac{P(B|A) * \cancel{P(A)}}{\cancel{P(B)}} = P(B|A),\ \ if\ and\ only\ if\ P(A)\ is\ uniform.$$

With this we have already revealed the most essential difference between the frequentist and Bayesian interpretation of Bayes' rule, as the frequentist apporach always assumes a uniform prior, such that the posterior will be equivalent to the maximum likelihood estimate (to which we will come in the *third* part of this tutorial series in detail).

As we know, the likelihood $P(data|theta)$ relates to what is called the alternative and is therefore *almost* what is typically referred to as **the p-values, testing for the null hypothesis** – which would be denoted as the probability of $data$ given $\overline{theta}$, i.e., $P(data|\overline{theta})$. In other words: the probability of having the $data$ or making an experience, given the complement of theta, the null-hypothesis (in other word: the probability that our hypothesis theta is false, given (new) $data$). As mentioned, you may come across phrases such as "the p-value is the probability of the $data$ given that the null hypothesis is *true*" – don´t get confused, it is the same as saying *theta* to be *false*. **There are also some other minor twists concerning the t-test, the confidence interval, the power of a study, density functions etc. that we have to go through when trying to fully understand how we get to a p-value *given a linear regression model* (again this will be done in the *third* part of this series).**

However, we are still well prepared to discuss the general difference between the Bayesian and the frequentist approach to probability in the next chapter.

$$P(theta|data) = \frac{P(data \mid theta)P(theta)}{P(data)}$$

$\textbf{\textit{Posterior probability}}, \textit{which includes both unweighted (uniform) and weighted prior}$

$$P(theta|data) = \frac{P(data \mid theta)\cancel{P(theta)}}{\cancel{P(data)}}$$

$\textbf{\textit{Maximum likelihood}}, \textit{given the special case of a uniform prior probability distribution}$

$$P(\overline{theta}|data) = \frac{P(data \mid \overline{theta})\cancel{P(\overline{theta})}}{\cancel{P(data)}}$$

$\boldsymbol{p-value}$ $\textit{in the common sense: the data given the null hypothesis} - \overline{theta}$

Fig. 13 Different cases of conditional probability, reflected via Bayes´ theorem. The posterior will always fully adapt the likelihood, given a uniform prior, and will at least move towards the likelihood, given an *informed prior*. Again, because it is the world, the realm of experiences we want to relate our model to.

We have learned…

… basic functions in R, such as defining an object, executing lines of code and presenting them as console output. We did some testing and calculated the posterior given some chosen values for the prior and the likelihood. We also learned what probability vectors are.

... that the posterior and the likelihood are the same, given a uniform prior probability distribution, which can also be cast as unweighted prior. We also found out how the basic structure of a p-value looks like: $P(data|\overline{theta})$

4 The Bayesian and the frequentist approach to (conditional) probability

How does the special case of equivalence between prior and likelihood fit our intuition? In general, a uniform prior can be considered as taking a "neutral position a priori" – at least neutral regarding the weight of the *pre-set contingencies* of our categories (above it was binary). Equivalent to our coin example we assume a kind of 'fairness' *a priori*. This makes sense in a lot of ways – in others it absolutely doesn´t and contradicts a neutral position due to 'false balance', due to a uniform prior. In order to get a grip on what that means, we will have a look at another famous xkcd webcomic:

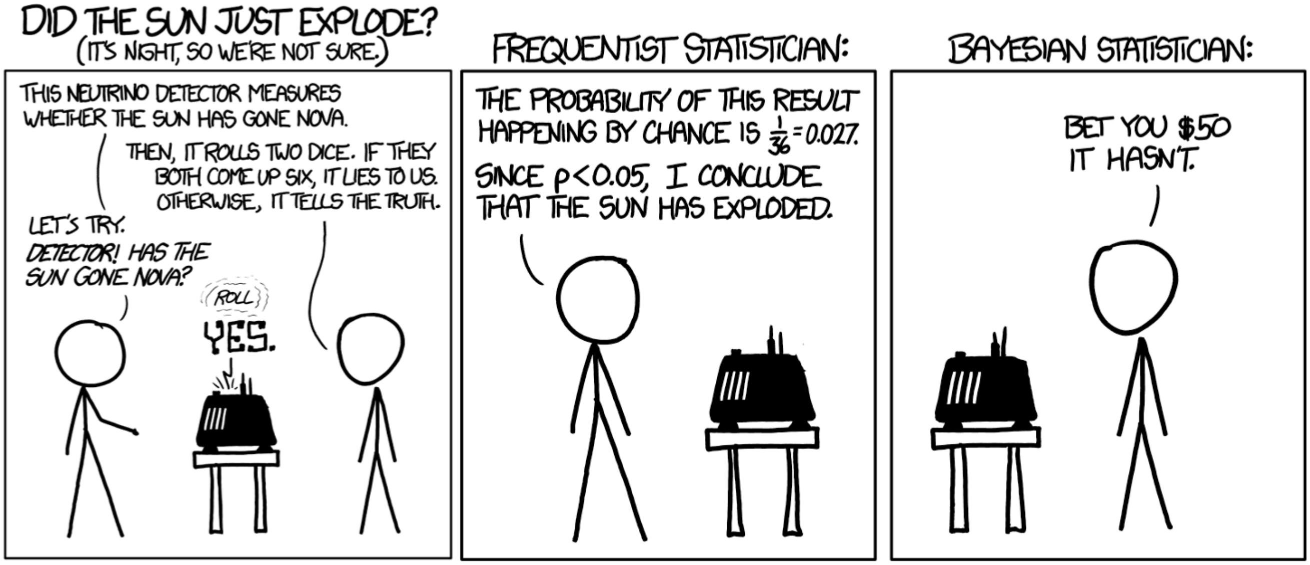

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Fig. 14 Another xkcd comic on mathematics. The comic has a lot of wonderful implications concerning the difference between the two approaches to probability. The **far-left part** argues that a machine does some measurements (gathers data) regarding the state of the sun (Nova / not-Nova) and then reflects on how likely the measurement is to be true amongst a series of measurements, where in rare cases (1/36) the machine will "lie to us" (refers to a threshold probability of .05, in order to discard the null hypothesis, as 0 is not what we usually get (due to normal exceptions, but in essence it's an arbitray threshold)). In the **middle part** the frequentist statistician reflects on how rare such an event would be, in order to check if the null hypothesis can be discarded (again, does not have to be a zero likelihood, but a defined cut-off value, such as 0.05, or 5%; the calculation would also include reflections on the sample size). However, the important part in this webcomic is the reaction of the Bayesian statistician in the **far-right part** of the comic: You may ask, how come so certain? In general, the comic beautifully conveys the problem of overfitting. Recall that the frequentist approach involves a uniform prior, in order to stay neutral, i.e., this approach implicitly argues that the states "Nova" and "Not-Nova" have a probability of .5 assigned to each *a priori*. As we know, this results in a model that will fully adapt the data (our measurement) – the posterior being equal to the likelihood. The Bayesian approach on the other hand allows weighting the likelihood (also called *informed prior*). Knowing that a Nova is in general a very rare event, the frequentist statistician indirectly overweight's the event "Sun has gone Nova". The Bayesian statistician, aware of this circumstance, therefore immediately bets against the frequentist model. On the other hand, this also implies that an overconfident prior can lead to overfitting in both interpretations of Bayes´ rule. All of this also relates well to the difference between evidence and significance, as discussed in chapter 1.2.

There is more to overfitting and we will probably come up with a tutorial on this in the future too. The take home message that we intended to convey is that both interpretations of Bayes' rule / conditional probability lead to problems in similar forms, when trying to gain evidence of any kind from a statistical analysis. The Bayesian approach is trying to get hold of issues such as overfitting, by setting informed priors (e.g., on prior knowledge from previous research). The frequentist approach will have similar issues, just more related to the interpretation of the results and less to the way data is gathered (when likelihood is unweighted). There is a great number of methods trying to overlook and overcome such boundaries in any of the two "fields", both mathematically and intellectually ('What is evidence?'). At the end, the attribute "fields" is somewhat over the top, as both approaches refer to the same equation, just under different prior assumptions. The 'tilting effect' between likelihood and posterior in combination with the *ex negativo* likelihood $P(B|\bar{A})$, when testing for the null hypothesis, eventually blurs away the fact how banal the difference between the two approaches is in the end.

When doing your own research, you may have come across further distinctions between the Bayesian and the frequentist approach. E.g., in Bayesian statistics it is said that the hypothesis is dynamic or changes, the data being something constant. In frequentist statistics it is said that the data or likelihood changes, and the hypotheses stay stable (i.e., binary (0% and 100%). This may seem complicated or even enigmatic, but it essentially just refers to the denotation of the variables:

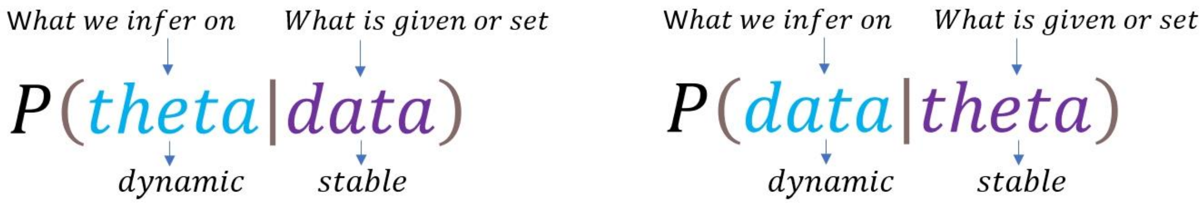


Fig. 15 In the figure we see the reason that theta or data is either considered stable or dynamic, as stable just refers to the temporal course of inference: the posterior being the probability of theta *after* data was gathered (set), and the likelihood is the probability of data, *after* a hypothesis was formed or set. Note again that in frequentist statistics posterior and likelihood are actually the same, given uniform prior, so the difference is rather synthetic and actually concerns the prior only, not the conditional probability. In case you wonder, what a dynamic hypothesis is, here is an interesting example: a prior of [.1 .9] and a likelihood of [.9 .1] will result in a uniform posterior of [.5 .5] (try yourself in R!). Updating the model with the new prior will eventually result in the set likelihood (as data is considered stable) and therefore results in a change, either from betting on theta to betting on its complement (or vice versa) *over time*. Apart from that it is again just tilting around with conditions on variables and does not add much to the story.

As mentioned, the difference between the Bayesian and frequentist approach is often cast as philosophical discussion on probability as such: Bayes' rule assumes that experience changes the way we hypothesize or expect the world to be *a priori in the future* (posterior becoming the new prior), where the frequentist approach assumes the possibility of a stable neutral position *a priori* (always uniform, never updated!) and casts evidence as a frequency of an assumption (being able to "frequently discard" the null hypothesis). Given a uniform prior, we will obtain an equivalence between posterior and likelihood, **so the course of inference does not matter in such a case**. So again, the difference is rather synthetic and it logically makes at least to me much more sense, to always consider the outcome probability of our statistical inference as the probability of a hypothesis after we gather data, under specific prior conditions (either uniform or informed prior). The p-value under frequentist considerations is still always denoted a likelihood, even though given a bi-directionality of inference so to speak; this can also be understood as trying to representing the "present" only (focusing on the data only, though: a uniform prior is still a prior assumption!).

However, arguing this to be a "philosophical" discussion is somewhat misleading, as we learned that the above is again just a reflection on Bayes' rule (in particular likelihood and posterior). So in general, just keep in mind that the frequentist approach to probability just reflects a special case of Bayes' theorem.

However, being aware of different approaches to probability theory does not involve choosing for a specific side in that discourse. We rather believe it to be essential to be aware of the prior considerations of hypothesis testing no matter how the chosen prior weight may look like. Still, in a lot of cases it is not just a matter of style or opinion, which side or method we choose, as we are all Bayes' when it comes to, e.g., the positive predictive values, or when performing differential diagnostics ("investigative reasoning", see below).

One last thing before we close our reflections on Bayes and frequentist stats: As mentioned, the Bayesian complement of a "frequency" is the posterior to become the prior . This can be understood as a structural recursion / iteration of hypothesis testing. Investigative reasoning has therefore often been related to Bayes' rule: Vanessa Holmes' new case involves four suspects, one of them being the murder of an innocent racoon. A *prior probability* could for now look something like [25 .25 .25 .25] for each suspect, when there are no specific prior assumptions on a suspect given so far (no clues) – staying elementary neutral and letting the facts speak first, so to speak (initial uniform prior). Holmes checked on one of the suspects, but the possible suspect has a clear alibi. With this new observation, our prior probabilities will subsequently change to [.33 .33 .33 .0], so our model (joint) was subsequently updated (as it consists of prior and likelihood). Below you will find the code to do more iterations. Another example would be reasoning in the medical field in terms of differential diagnostics, as well as evidence-based medicine in general, as we now know. Here is some code to replicate the Vanessa Holmes example:

```
# First iteration:
joint = c(.25,.25,.25,25)*c(.33,.33,.33,0) # prior*likelihood
model_evidence = sum(c(.25,.25,.25,.25)*c(.33,.33,.33,0))
posterior = joint/model_evidence

# Result:
# [1] 0.3333333  0.3333333  0.3333333  0

# Second iteration (another suspect ruled out):
joint = c(.33,.33,.33,0)*c(.5,.5,0,0)
model_evidence = sum(c(.33,.33,.33,0)*c(.5,.5,0,0))
posterior = joint/model_evidence

# Result:
# [1] 0.5  0.5  0  0

# Third iteration - let's say all of the suspects are ruled out:
joint = c(.5,.5,0,0)*c(0,0,0,0)
# [1] 0 0 0 0
model_evidence = sum(c(.5,.5,.33,0)*c(0,0,0,0))
# [1] 0
posterior = joint/model_evidence # Note that it says: 0 divided by 0!
# [1] NaN NaN NaN NaN

# NaN = Not a number.
# This is due to the fact that 0/0 is not defined. At this point
# Vanessa Holmes would need to find another suspect to be able
# to do further inference...
```

The first part of our series on statistical inference has come to an end. We hope that this has given you a stable overview over what hypothesis testing is in general all about and how it is related to conditional probability. You will definitely come across conditional probability in various forms, when further digging into statistics. A lot of our section members are also interested in computational neuroscience, bioinformatics and data science in general, so we will therefore soon also provide tutorials on topics such as information theory (Shannon, Akaike; both relies on reflections of thermodynamic analogies, which also involves conditional probability (Boltzmann and Gibbs´ entropy, free energy etc.)), predictive processing / active inference (disconnection hypothesis arguing schizophrenia to be a weighting problem) and many more, which all rely on the basic concept of hypothesis testing in the sense of Bayes' rule (tutorials on these topics are also just about to be finished, so stay tuned!).

The information theoretic use of conditional probability is one of the most fascinating and mind blowing and refers to the "bit" as a statistical quantity, showing that communication is a stochastic process and comes without conveying meaning in the actual sense (Shannon). Something that especially research in the field of humanities struggled a lot with conceptually when trying to understand information technology (unfortunately there a lot of heavily faulty interpretations of information theory and what a computer actually does).

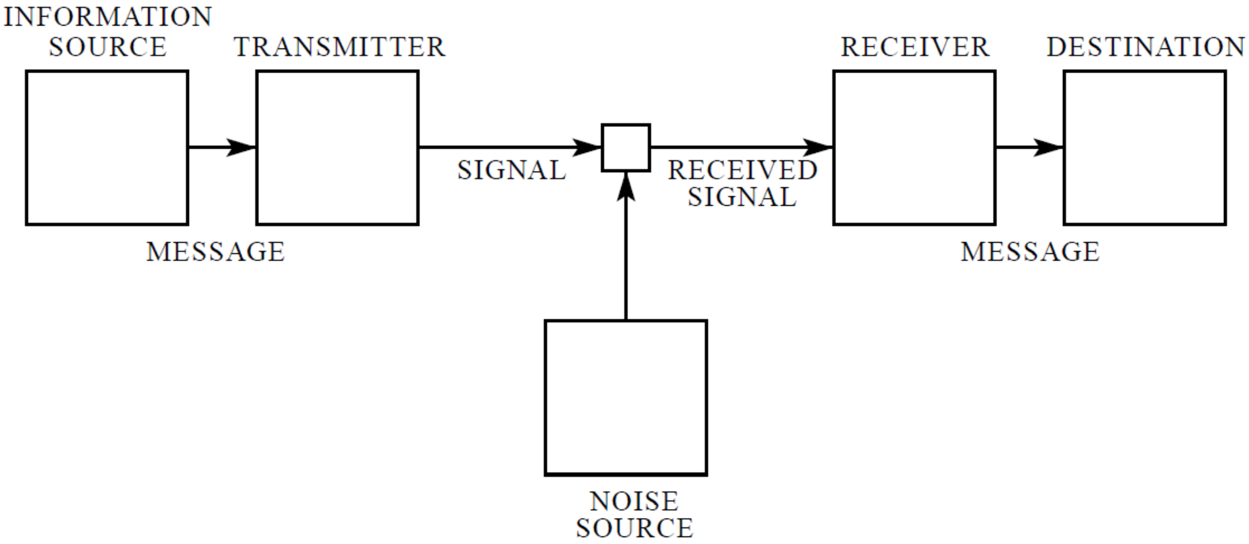


Fig. 1 — Schematic diagram of a general communication system.

Fig. 16 The original schematic diagram of a general communicational system from the ground breaking paper "A mathematical theory of communication" from 1948, by Claude E. Shannon. Shannon showed that sending a signal and receiving a signal can be understood as a stochastic process, as a course of inference as well, so to speak. The channel can be understood as a joint probability, and the goal of the receiver is to correctly infer the sent signal, given the received signal $P(sent = 1 \mid received = 1) = 1$, given a noiseless channel (inferring the probability that a binary 1 was sent given that a 1 was received as being 100%). A 1 or 0 can also be assigned a prior probability of failure, considering noise (again just weighted probability).

However, note that hypothesis testing is often also referred to as **abductive inference in contrast to deductive and inductive reasoning** and was defined by the mathematician and philosopher C.S. Peirce that had a great influence on the development of information theory (logic gates, further development of Boolean algebra, abductive inference and communication; the temporal triad we referred to is essentially referring to the triadic structure of abductive inference (related to semiotics)). The difference to deductive and inductive reasoning is essentially that abductive inference / hypothesis testing involves developing and evaluating something *new*, a hypothesis, and does not just evaluate a pre-set part-to-whole relation between a rule and a single event as in the other two forms of inference. Arguing that this can be looked at as a hierarchical and recursive process, such that a deduction is just an abduction of an abduction, then one can say that statistical reasoning in the sense of a mathematical method can be referred to as deductive reasoning (also under "neutral assumptions" in particular), as a process of forming a hypothesis has already been performed cognitively before and therefore on a lower-level of the hierarchy so to speak. However, it does not change our general intuition on hypothesis testing, it rather expands its influence and the possible complexity of representing argumentations in the form of statistical inference.

We have learned…

… some of the basic difference between the Bayesian and the frequentist approach, which basically most of time relies on how priors are involved in the process of statistical inference.

... that conditional probability is related to wide range of methods in (comp.) science (even thermodynamics, information theory…).

We hope that you have learned…

... that statistics is of course not a trivial task, but is far more related to our intuition and general approach to inferring on the world and ourselves as the general (public) tenor suggest.

Fig. 16 Ms. Miranda, longing for feedback data. Did any of this makes sense? We would like to know from you! Similar to our open review process, **every of our articles can be commented by you**. Our journal still relies on the expertise of other student and professionals. However, the goal of our tutorial collection is to especially come in contact with you, helping us to create an open and transparent peer-teaching space within BEM.

Appendix: Our Submission for the >>Summer of Math Exposition II<<

This tutorial was submitted for the Summer of Math Exposition II, organized by Grant Sanderson and James Schloss. We are proud that we made it into the top 10% of all entries (the top 100 entries of which 25 were non video entries).

To give some other submission of the competition some attention, here is one that fits very well to the topic of our tutorial: The article by Jeffrey Wang discusses misinterpretations of conditional probability by a medical doctor that played the role of an "expert" in court trial. The misinterpretation led to the false accusation of a mother to be the cause of the death of her infant that actually "just" died from sudden infant death syndrome (SIDS). The false accusations and imprisonment eventually led to lethal self-destruction of the mother, Sally Clark (which eventually died of alcohol intoxication). It is a drastic example of the misconception of statistics by medical personal in this case (which are unfortunately not experts on that matter). The example also involves conditional *independency*. **Above we have only gone through examples of a conditional dependency, but will eventually get back to the topic some other time as well.**