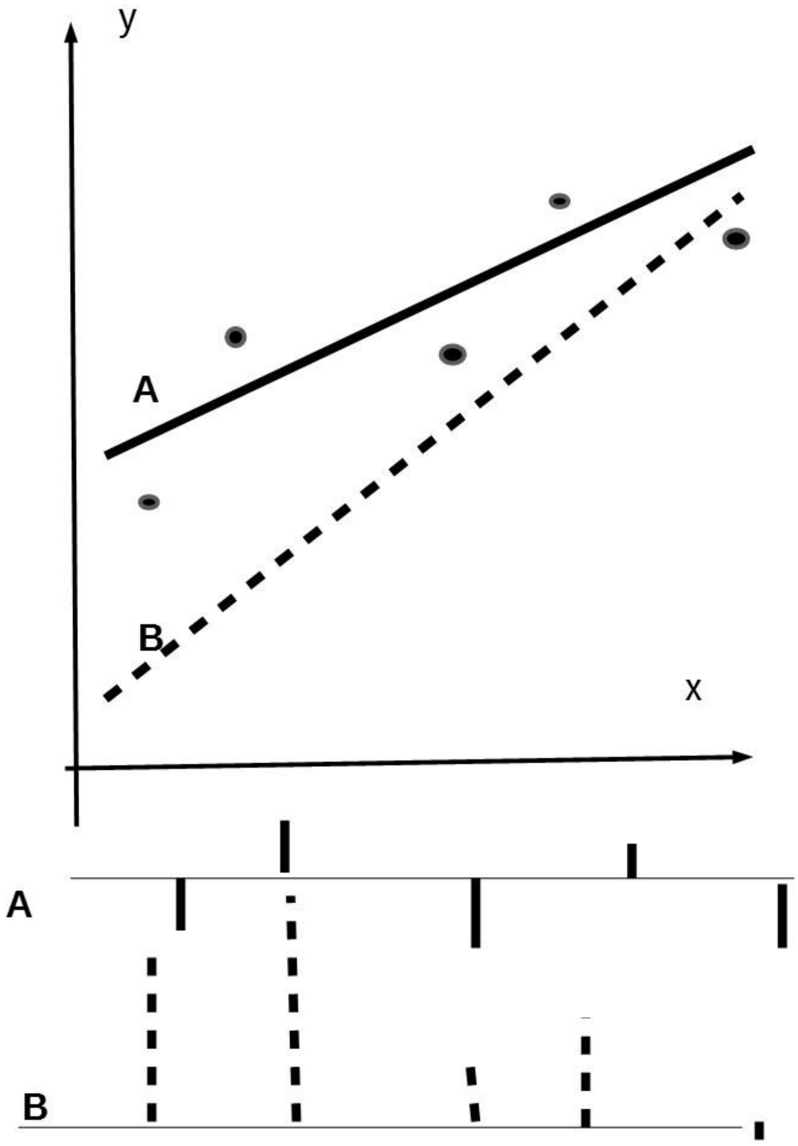
Information and statistical inference

One can view Shannon's model of communication as a particular case of a more general problem. Interestingly, this more general problem predates Shannon by nearly two hundred years, and was the concern of philosophers even before mathematicians took it up – it is the problem of induction. The logic formulated by Aristotle is deductive – going from cause to consequence. Most scientific investigation tries to do the opposite – to infer the cause from the observed effects. It is well known that this is not unique – the same effect can emerge from different causes. It is also known that even in the forward direction – from cause to effect, the result may not be unique. The process might be inherently random, or there might be noise in the measurement process. The problem of communication is a special situation in which one can infer the transmitted message (the cause) from the received message with negligible error, even in the presence of noise, by proper design of the experiment. We may not always be that lucky. Even after the measurement, uncertainty may still remain. The job of the experimenter, then is both to reduce and to quantify this uncertainty.

Going from the sublime to the mundane, let us revisit something everyone does – least squares. Let us say we have a model for how the price of potatoes increases with time - $y=mx+c$. The truth that we wish to learn is what the values of *m* and *c* are. If there were no noise, just two points would be enough to draw the straight line, and get the values of *m* and *c* – two equations for two unknowns. But in the real world, there is noise. Our model for the experiment is that $y_i=m x_i+c+r_i, \quad 1\leq i\leq N$ We have *N* points and the ith point has a random residual r_i . We usually call this an 'overdetermined' problem because we have N points and only two unknowns. And the recipe, which makes a lot of intuitive sense, is to minimise the sum of the squares of the residuals, and get the value of the two things we are interested in. What is good enough for Gauss is good enough for us! The legend goes that he invented this method to find out the orbit of the asteroid Ceres from a small number of observations, so that the observers would know where to look when it reappeared from behind the sun.

But notice that this should really be called in 'underdetermined' problem if we include the random residuals among the unknowns – there are always two more equations than unknowns. In the framework of induction, our final answer should not be unique values for the two parameters *m,c* but probability distributions for them based on probability distributions for the random residuals r_i . The evidence is that Gauss was aware of this, and in fact used his own distribution, for the random errors of measurement, to justify his method. The probability of the residuals taking some values is proportional to $\exp(-r_1^2/2\sigma^2)\times\exp(-r_2^2/2\sigma^2)....\times.\exp(-r_N^2/2\sigma^2)$. Here the parameter σ measures the width of the gaussian distribution. Thanks to the convenient property of exponentials, this probability is nothing but $\exp(-(r_1^2+r_2^2......+r_N^2)/2\sigma^2)$. By minismising the sum of squares, we are maximising this probability. We are choosing our parameters *m, c* so as that the residuals implied by this choice are most likely to occur. Hence the name, maximum likelihood for this approach. This is a well known statistical methodology. C.R.Rao, the doyen of statisticians who have emerged from Indian soil, has done great work on this topic, amongst many other things.

The situation is illustrated in the figure below. A set of five points is shown along with two lines, A and B, (B is dashed). For each line, the correspoding residuals are shown below the graph. Our preference for line A rather than line B is based on our preference for the A residuals in the solid lines rather than the B residuals in the dashed lines. Interestingly we are not using any knowledge of the size of the residuals.



This brings us to a fundamental divide which has persisted in the statistical community to the present day. The divide is hidden in the word – likelihood. The quantity we wrote down is the probability of getting some values of y_i, given *m* and *c* . It is a probability distribution for y_i . It is not a probabilty distribution for *m* and *c*! In fact, we could argue that it doesn even have the right dimensions! It is meant to give the probability per unit interval of y (price of potatoes), not unit interval of m and c. The probability density for y would have dimensions of kg/rupees, while that for m would be time x kg / rupees) If for no other reason, we need an additional factor to convert it into a proper probabilty distribution for the quantities we are interested in, slope and intercept. But there is a better reason, and thereby hangs the tale of Bayes theorem, into which we now digress – see box below if you are new to Bayes theorem, or skip to move on.

…………………

Box: Bayesianity – edited version of a blogpost by RN

The Reverend Thomas Bayes (1701-1761) was an English Presbyterian minister at a time when such men learnt and wrote on calculus and probability. But his real legacy was published posthumously, a theorem in probability close to trivial (see below). The later application was to statistical inference, and completely nontrivial.

For "statistical inference"read "how we acquire knowledge through experience", epistemology, no less. It deeply divided the world of people who worry about such things, philosphers and statisticians,   ever since. One school holds it as a 'basic" tenet and the other "bays" for their blood!

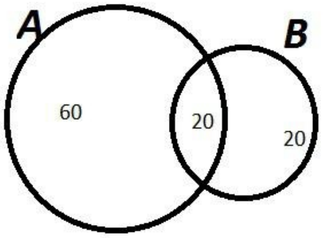
Here is the story , told informally and with a personal angle thrown in Astronomers try to tell you what is out there from what enters their telescopes, a problem of inference which my then colleague Ramesh Narayan and I worked on in the late seventies and early eighties I first heard of the proof of Bayes theorem using two beer mats from the astronomer Steve Gull of Cambridge, apparently conveyed to his colleague John Skilling, presumably some beer was also involved, They later set up a company based on Bayesian infrerence which did rather well.

*. Let us begin with the theorem. Its really all about definitions, best illustrated by an example. .A class of a hundred students has forty who like bananas, eighty who like apples, and twenty (the minimum possible number, incidentally) who like both. In the language of probabilities, thinking of them as fractions, and ignoring the issue of whether hundred is a large enough number, one can say the the probability of A (the event of liking apples) written as **P(A) = 80/100=0,.8** and similarly **P(B)=40/100 =0,4**. The probability of someone liking both a is called the joint probability of A and B, written **P(A,B)** and it equals **20/100= 0,.2**.*

*Now we move to conditional probabilities. Given that someone likes apples, i.e given A, what is the probability that that person likes bananas? This is called the probability of B, given, or conditioned on, the truth of A, and is written **P(B|A)** and is clearly **0.2/0.8= 0.25**. Notice that this is not the same as **P(A|B).** This is the the answer to a different question, the probability of a given banana lover likes apples, which is **0.2/0.4=0.5**.*

*There are now two routes to the joint probability **P(A,B).** First start with the probability that a student picked at random likes apples, viz. 0.8. Then note that the fraction of these who like bananas is 0.25, and hence the fraction of the total would be 0,.8 times 0,25, viz. 0.2. Alternatively, starting with the fraction of banana lovers, 0.4, multiply by 0,.5, the fraction of banana lovers who like apples, to get the same answer for the joint probability, 0,.2. In symbols,*

P(A,B)=P(A)P(B|A) =P(B)P(A|B). *Bayes theorem states that these two routes must give the same answer, and the supporting beer mats are sketched below.*



The theorem is usually written in a different form.

$$P(H|D)=P(D|H)\ P(H)/P(D).$$

Note the change in symbols. Have the apples and bananas have been replaced by "Heerekai" and "Donnemensinkai" (Kannada for ridge gourd and capsicum) ? No dear reader, we have plunged into the hornet's nest of epistemology**, H** stands for "hypothesis" and **D** stands for "data". The use of probabilities on the left hand side reminds us that the outcome of our process is not certainty. But we can still judge the relative probabilities of different hypotheses, given some data – surely a posture of becoming humility. Now if only we had been asked to solve the problem of deduction, i.e given a hypothesis **H,** what is the probability that the data would turn out to be **D**, life should be much simpler. In fact, if we do not know **P(D|H)** we should not be working in this field at all! Bayes theorem allows us to make the giant leap from deduction – **P(D|H)** –to induction **P(H|D).** It was Bayes illustrious successor across the Channel, Pierre Simon Laplace, who squarely (no pun intended, but his name translates as "The Square") took this step, and many have followed him.

The price of this breakthrough is the factor **P(H)** the unconditional -, i.e given no data at all!-probability of the hypothesis being true.. This is called the 'prior' – and the probability distribution after getting the data is called the "posterior" . Toba Toba, would be ones first reaction to this notorious "prior distribution". Ramesh and I expressed our scepticism and moved on to explore a range of alternatives, with fair empirical success. But in astronomy, and especially in cosmology where the link from data to hypothesis used to be tenuous, the Bayes methodology refused to die, as indeed in many other situations calling for inference from noisy and incomplete data. This is not the place to go into technical details or specific problems, but just to give an example, let us take the rate at which the universe expands, which is an important quantity for cosmologists, and happily, also denoted by **H.** Thirty years ago, one could only say it lay somewhere between 50 and 100 (in units which do not concern us here). A Bayesian would in fact give a probability distribution **P(H|D)**, known as the "posterior" because we came in with a prior and it got modified by the data. As each new piece of data came in, the distribution would get narrower –today it has a peak around 70, and drops off rapidly as we move away by more than a few.

Standiing back, one can see many virtues in Bayesian inference. To the opponents of this approach, the greatest weakness would seem to be the subjectivity involved in the choice of the prior, but Bayesians take the position that this is the greatest strength. One may approach a problem with an open mind, but it is never a blank mind. The demand for a prior is really a demand to state otherwise hidden assumptions. The literature characterises a prior as broad where the final result – the winning hypothesis **H** with the highest **P(H|D),** for example – is not very sensitive to the choice of **P(H)** . In the opposite case, the dependence on the choice of prior, even when different people make different reasonable choices, is actually a warning to us that we need better data, or that one should not be drawing very strong conclusions with the current data. The fact that the final answer is a probability distribution for **H** is very helpful if we are going to use **H** for some other purpose like making an investment for example. We can then get a probabililty distribution for anything which can be calculated from **H.** The systematic propagation of uncertainty along a deductive chain is definitely desirable. The ability to keep *incorporating more data by using an existing posterior as a prior for the next data set again seems a natural reflection of the 'scientific method'. And finally, methods which claim to be non-Bayesian may simply be hiding the prior in their formulae. As one example, "maximum likelihood" focuses on the noncontroversial* ***P(D|H)*** *(called as the likelihood when viewed as a function of* ***H*** *) and maximises with respect to* ***H*** *even though it is not a probability for* ***H*** *. The Bayesian simply says the prior has been chosen as a 1 (unity) multiplying* ***P(D|H)*** *so this is just a 'flat prior' in disguise.*

The most interesting conceptual issue thrown up by the Bayesian strategy is the meaning to be assigned to the prior and posterior. We had a largish collection of students in our starting example, so fractions can be checked, but a largish collection of universes with respect to which our probabilities are defined boggles the mind (not that there aren't votaries of the multiverse among cosmologists). Incidentally, the probability distribution ***P(D|H)*** *does not suffer from this problem – given one universe, expanding at a given rate, its perfectly easy to imagine many astronomers making repeated measurements and arriving at different results* ***D*** *. The best that I know is to view* ***P(H)*** *and* ***P(H|D)*** *as "degrees of belief" regarding the validity of* ***H*** *before and after the data were taken. . This is not as weak as it seems – assuming ones beliefs are consistent forces these numbers to obey the same laws as the probabilities which we are used to – fractions of a large sample. In fact, to pure mathematicians, probabilities are any set of numbers obeying these properties, (the usual rules to add and multiply probabilities) formalised as the Kolmogoroff axioms. Someone else is free to have different degrees of belief in the same propositions, but they too will obey consistency conditions allowing us to work with these probabilities in the way we are used to. An interesting twist was given to the degree of belief by de Finetti, in the 1930's, converting it into a gambling game. Your subjective "probability of life on Mars" is just the fraction of a dollar at which you would be prepared to buy, or sell, a document with a true promise to pay one dollar if life were found on Mars, and zero if it was not. The wheel has turned full circle. Chevalier de Mere approached Pascal and Fermat with his gambling problems and they created probability theory, and after a few centuries, the meaning of probability in real life situations covered by statistical inference, rests on betting odds!*

Viewed as epistemology, the Bayesian viewpoint seems to have all bases covered. Different belief systems are different priors, the subjectivity is upfront, both in the choice of the priors and in their interpretation. . The Gettier paradox – where someone reads the right time by chance from a stopped clock – is simply a case of the prior not including the possibility that the clock was stopped, and/or the data not including tests of its actually moving. We live and learn, rather than agonise. Justification comes from the likelihood function. And there is a sanity check, in something we have not talked about so far, the denominator of the right hand side, ***P(D****) . Since we are mainly interested in a situation where* ***D*** *is given and we are exploring different* ***H****'s, this is fixed, but does play the role of normalising* ***P(H|D)*** *properly. More to the point, if* ***P(D****) is very small, it either means that our data has a very small probability of being realised regardless of the hypothesis in the chosen set. So either you were unlucky in the way the dice rolled when you did your experiment, or, more likely, you should go back and look at your* ***P(H)*** *. Maybe you missed out on some alternative hypotheses that would have given a more reasonable probability of getting the data that you did get. So even the denominator is useful and goes by the name of "evidence".*

I remember correspondence with a friend, Partha Pratim Mitra, who was using some ingenious method to do his data analysis and I asked him why not a Bayesian approach. His answer well expressed the opposing point of view – he said that he was not religious! And indeed, Bayesianity requires that you need faith in the idea of a prior, and once you put your faith in it, all else is taken care of. You can change your prior based on evidence, though, so its really a meta religion, Happy Bayesing!

……………………………………………………………………………………

Our problem is to convert a probability distribution for data, given a hypothesis, to the opposite – the probability distribution of hypotheses, given data. Bayes theorem tell us how to convert one of these to the other. The box goes into more detail on conceptual issues. With continous distributions, it is now dimensionally correct. We are forced to introduce a prior distribution, ***P(H),*** which is our knowledge about *H* before we were given the data.

$$P(H|D)=P(D|H)\ P(H)/P(D).$$

We are now ready to go back to our least squares problem. We can think of it as Bayesian inference, but with a flat (that is, constant in the region of interest) prior for the parameters *m, c* . But dont be misled by someone who claims that by using a flat distribution, we are being maximally non-commital or open minded. A disitribution which is flat in terms of *m*, and *c*, is not flat in terms of m^3, c^3 . When you deal with continuous variables, no choice (that is, writing unity) is also a choice!

In Shannon's theory, this prior is the probabilty distribution which describes the ensemble of messages, the set of all ***H's*** with all the correlations and redundancies characteristic of the source. The received message is ***D,*** and the noisy channel is characterised by ***P(D|H).*** The probability distribution at the receiving end, over the set of all transmitted messages, is ***P(D) .*** What happens in the case of working at or below the channel capacity is that for a given ***D***, there is a unique ***H –*** a unique transmitted messagefor each received message, even with noise. This means that ***P(H|D)*** is concentrated at one ***H*** for each ***D***. In the more general case, we are left with a probability distribution for ***H.*** If Bayes had come after Shannon, we would have accused him of being theological, of thinking of science as interpreting messages sent by the creator! (He was Rev, after all!).

Bayesian inference has some nice properties. Let us imagine repeated experiments giving data $D_{1,}$ $D_{2,}$ ...D_k . We should then be judging the hpothesis according to all the data, that is, look at $P(H|\ D_1, D_2, ...D_k\ )$. In practice, the data may come in sequentially. So one may be tempeted to use the posterior after the first piece of data has arrived, as the prior when analysing the second piece, and use the second posterior as the prior for the third stage, etc. The nice thing is that the final result is that same as if we had all the data at one go. Another property connects with Shannon information. It seems very reasonable that the entropy associated with the posterior should be less than that related to the prior. Otherwise, the data is reducing the information which we have about H, which would be unfortunate. Fortunately, there is a theorem assuring us that the entropy associated with ***P(H***) is always greater than that associated with ***P(H|D).*** We can also view coding in the Bayesian framework. It is just changing the variables which describe the message – that is ***H,*** with a new set which have a flat distribution. Once one has compressed the message to N independent bits, they are independent and also have equal probability for being one or zero. (If not, the message could be further compressed).

A final remark, Even if one does not want to follow Bayesian inference because of allergy to the prior, everyone should know the theorem, and in particular, that ***P(A|B)*** is not the same as P***(B|A)***. The fact that the majority of phishing e-mails emanate from Nigeria doesnt mean that the majority of Nigerians are phishers! While this is obvious, a lot of opinions that one hears, are based on this confusion- and some can have seriuos and tragic consequences.