UNIVERSITÀ DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche, aziendali, matematiche e statistiche "Bruno de Finetti"

Bayesian Statistics

Introduction

Francesco Pauli

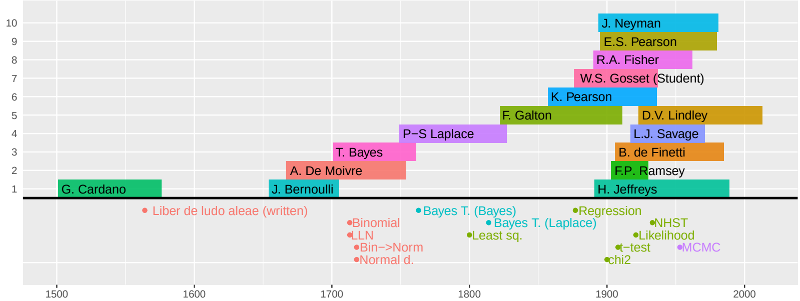
A.A. 2018/19

Indice

Why starting from history

We will start with an overview of the history of Bayesian statistics: its development from probability calculus and its relationship with the so called classical statistics.

This is useful to better understand the differences between Bayesian and classical statistics.



Games of chance are the cradle of probability

Probability calculus is initially developed to study games of chance: developing strategies to win in games was of interest to nobles, who were willing to pay scholars for them.

For example Galileo in 1620 wrote a note offering the solution of this issue:

> *suppose three dice are thrown and the three numbers obtained added. What is the probability that the total equals 9?*

This circumstances help not only because of the money, but also because of the simple structure of the problems involved.

Elementary probability

The first examples of probability problems are concerned with simple random mechanisms whose symmetry offered the solution.

> *Q: A marble is randomly drawn from an urn containing R red marbles and W white marbles, what is the probability that the marble is red?*
> *A: By symmetry*
>
> $$P(red) = \frac{R}{R+W}$$

Games of chances are easily tackled using the first definition of probability, based on symmetry

$$\text{prob. of event} = \frac{\#\ \text{favourable outcomes}}{\#\ \text{possible outcomes}}$$

Elementary probability and combinatorics

For more "complicated" questions tools were developed to count favourable and unfavourable outcomes.

Q: We draw a marble from an urn containing R red marbles and W white marbles m times (putting it back in the urn after each draw), what is the probability that r out of m are red?

A: Still by symmetry

$$P(r \text{ red out of } m) = \binom{m}{r} \left(\frac{R}{R+W}\right)^r \left(1 - \frac{R}{R+W}\right)^{m-r}$$

Girolamo Cardano (1501-1576), wrote the first systematic treatment of probability in 1576: *Liber de ludo aleae*; this, however was not published until 1663. He was a polymath with interests ranging from mathematics to biology. He was also a gambler (and a rumor exists that he did not publish his book on probability because his knowledge gave him an advantage in betting).

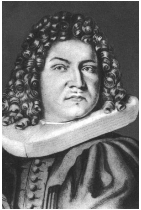
Limiting frequency

Moreover, in the context of game of chances it is easy to think of "repeating" events, so the probability of an event materializes as the

> *limiting relative frequency of occurrence of the event in a number of repetitions.*

This idea was developed and made more precise by Jakob Bernoulli (Ars Conjectandi, 1713) and Abraham De Moivre (1733) in the **law of large numbers** which links theoretically the probability of an event to the relative frequency in infinite repetitions.

Jakob Bernoulli (Basel 1654-1705), (Jacob, Jacques or James) in *Ars Conjectandi (1713)* discusses the application of probability to gambling. He develops techniques based on combinatorics calculus (and the binomial distribution) and a first version of the law of large numbers.



The law of large numbers

Theorem ((Strong) Law of large numbers)

Let $E_1, \ldots, E_n, \ldots$ be a sequence of independent events such that $P(E_i) = p$ for all i. Let $S_n = \sum_{i=1}^{n} |E_i|$ be the number of events occurring among the first n. Then

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = p\right) = 1.$$

Note that the theorem was already stated, without proof, by Cardano.

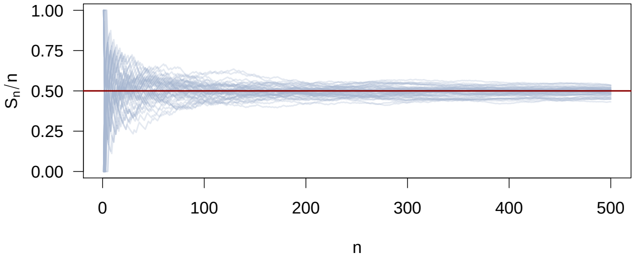
Abraham De Moivre (1667-1754) in *Laws of Chances (1718)* builds on Bernoulli's (and others) works. One of his main achievements is the formula for the normal distribution and the link between the binomial and the normal distribution.

The law of large numbers

Theorem ((Strong) Law of large numbers)

Let $E_1, \ldots, E_n, \ldots$ be a sequence of independent events such that $P(E_i) = p$ for all i. Let $S_n = \sum_{i=1}^{n} |E_i|$ be the number of events occurring among the first n. Then

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = p\right) = 1.$$



Direct problems, known mechanism

Up to this time, all developments were limited to **direct problems**: I know the random mechanism which generates the observations and I can compute the probability of the various outcomes.

An urn contains 10 marbles, R of which are red, $R \in \{1, \ldots, 10\}$, we draw a marble from the urn 5 times (putting it back after each draw) and record its colour, let X be the number of times a red marble is observed.

Then $X = 0, 1, 2, 3, 4, 5$ and, if we let $\theta = R/10$,

$$P(X = x) = \binom{5}{x} \theta^x (1 - \theta)^{5-x}$$

Direct problems, known mechanism

Up to this time, all developments were limited to **direct problems**: I know the random mechanism which generates the observations and I can compute the probability of the various outcomes.

An urn contains 10 marbles, R of which are red, $R \in \{1, \ldots, 10\}$, we draw a marble from the urn 5 times (putting it back after each draw) and record its colour, let X be the number of times a red marble is observed.

Then $X = 0, 1, 2, 3, 4, 5$ and, if we let $\theta = R/10$,

		Urn composition (θ, proportion of red marbles)									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Sample (x)	0	.5905	.3277	.1681	.0778	.0312	.0102	.0024	.0003	.0000	0
	1	.3280	.4096	.3601	.2592	.1562	.0768	.0284	.0064	.0004	0
	2	.0729	.2048	.3087	.3456	.3125	.2304	.1323	.0512	.0081	0
	3	.0081	.0512	.1323	.2304	.3125	.3456	.3087	.2048	.0729	0
	4	.0005	.0064	.0284	.0768	.1562	.2592	.3601	.4096	.3280	0
	5	.0000	.0003	.0024	.0102	.0312	.0778	.1681	.3277	.5905	1

A more complicated direct problem

Consider the following experiment

- An urn contains 10 marbles, R are red, R was decided by throwing a 10-sides die, the result is unknown to us.
- We draw a marble from the urn 5 times ... we observe X red marbles.

What is $P(X = x)$?

This is still a direct problem, the solution is obtained through the

Theorem (Law of total probability)

Let $\{H_i | i = 1, \ldots, n\}$ be a partition of Ω,

1. $\bigcup_{i=1}^{n} H_i = \Omega$ *(exhaustive),*
2. $H_i \cap H_j = \phi$ *if* $i \neq j$ *(pairwise incompatible),*

then

$$P(E) = P(E \cap \Omega) = \sum_{i=1}^{n} P(H_i \cap E) = \sum_{i=1}^{n} P(H_i)P(E|H_i)$$

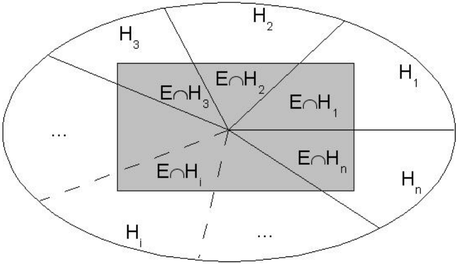
A more complicated direct problem

Consider the following experiment

- An urn contains 10 marbles, R are red, R was decided by throwing a 10-sides die, the result is unknown to us.
- We draw a marble from the urn 5 times ... we observe X red marbles.

What is $P(X = x)$?

This is still a direct problem, the solution is obtained through the law of total probability



A more complicated direct problem

Consider the following experiment

- An urn contains 10 marbles, R are red, R was decided by throwing a 10-sides die, the result is unknown to us.
- We draw a marble from the urn 5 times ... we observe X red marbles.

What is $P(X = x)$?

This is still a direct problem, the solution is obtained through the law of total probability

$$\begin{aligned} P(X = x) &= \sum_{i=1}^{10} P(X = x \cap R = i) \\ &= \sum_{i=1}^{10} P(R = i) P(X = x | R = i) \end{aligned}$$

LAw of total probability in tabular form

$$P(X = x|R = 10\theta)$$

x	Urn composition (θ, proportion of red marbles) 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	.5905	.3277	.1681	.0778	.0312	.0102	.0024	.0003	.0000	0
1	.3280	.4096	.3601	.2592	.1562	.0768	.0284	.0064	.0004	0
2	.0729	.2048	.3087	.3456	.3125	.2304	.1323	.0512	.0081	0
3	.0081	.0512	.1323	.2304	.3125	.3456	.3087	.2048	.0729	0
4	.0005	.0064	.0284	.0768	.1562	.2592	.3601	.4096	.3280	0
5	.0000	.0003	.0024	.0102	.0312	.0778	.1681	.3277	.5905	1

$$P(X = x \cap R = 10\theta) = P(R = 10\theta)P(X = x|R = 10\theta)$$

x	Urn composition (θ, proportion of red marbles) 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	$P(X = x)$
1	.05905	.03277	.01681	.00778	.00313	.00102	.00024	.00003	.00000	0	.12083
2	.03280	.04096	.03601	.02592	.01562	.00768	.00283	.00064	.00004	0	.16252
3	.00729	.02048	.03087	.03456	.03125	.02304	.01323	.00512	.00081	0	.16665
4	.00081	.00512	.01323	.02304	.03125	.03456	.03087	.02048	.00729	0	.16665
5	.00005	.00064	.00284	.00768	.01562	.02592	.03602	.04096	.03281	0	.16253
6	.00000	.00003	.00024	.00102	.00313	.00778	.01681	.03277	.05905	.1	.22083

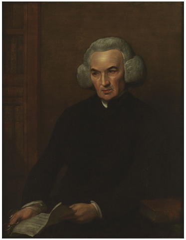
$$P(X = x) = \sum_{r=1}^{10} P(X = x \cap R = r)$$

Indirect problems: the probability of causes

Within the above experiment, we can also ask the following question

> *Having observed $X = x$, what is the probability that the urn contains R red marbles?*

This is solved by **Bayes theorem**.



Thomas Bayes (c. 1702-1761) was a Presbyterian minister. In *Essay Towards Solving a Problem in the Doctrine of Chances (1763)* he considers the inverse probability problem for which he formalizes a solution.
His work was published posthumously by his friend **Richard Price (1723-1791)**.

Bayes theorem: original formulation

In *Essay Towards Solving a Problem in the Doctrine of Chances (1763)* we find

Theorem (PROP. 3)

The probability that two subsequent events will both happen is a ratio compounded of the probability of the 1st, and the probability of the 2d on supposition the 1st happens.

Corollary (PROP. 3)

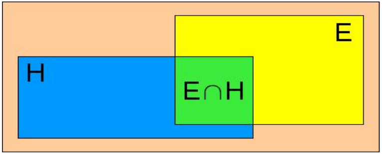
Hence if of two subsequent events the probability of the 1st be a/N, and the probability of both together be P/N, then the probability of the 2d on supposition the 1st happens is P/a.

Bayes theorem

Theorem (Bayes theorem)

Let E and H be two events, assume $P(E) \neq 0$, then

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{P(H)P(E|H)}{P(E)}$$

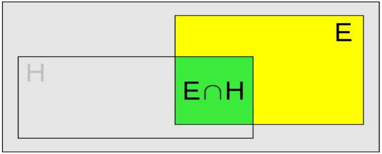


Bayes theorem

Theorem (Bayes theorem)

Let E and H be two events, assume $P(E) \neq 0$, then

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{P(H)P(E|H)}{P(E)}$$



Bayes theorem for the urn example

> *Having observed $X = x$, what is the probability that the urn contains R red marbles?*

The answer from Bayes theorem is

$$P(R = 10\theta | X = x) = \frac{P(R = 10\theta)P(X = x | R = 10\theta)}{P(X = x)}$$

Bayes theorem for the urn example

Having observed $X = x$, what is the probability that the urn contains R red marbles?

Assume $X = 3$

Consider the joint probabilities $P(X = x \cap R = 10\theta)$

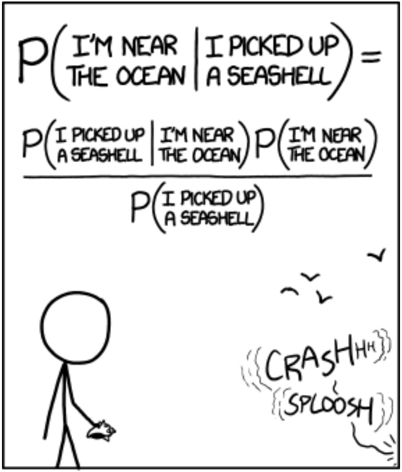
x	Urn composition (θ, share of red marbles) 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	$P(X = x)$
1	.05905	.03277	.01681	.00778	.00313	.00102	.00024	.00003	.00000	0	.12083
2	.03280	.04096	.03601	.02592	.01562	.00768	.00283	.00064	.00004	0	.16252
3	.00729	.02048	.03087	.03456	.03125	.02304	.01323	.00512	.00081	0	.16665
4	.00081	.00512	.01323	.02304	.03125	.03456	.03087	.02048	.00729	0	.16665
5	.00005	.00064	.00284	.00768	.01562	.02592	.03602	.04096	.03281	0	.16253
6	.00000	.00003	.00024	.00102	.00313	.00778	.01681	.03277	.05905	.1	.22083

then

$$P(R = 10\theta | X = 3) = \frac{P(X = 3 \cap R = 10\theta)}{\sum_{r=1}^{10} P(X = 3 \cap R = r)} = \frac{P(X = 3 \cap R = 10\theta)}{P(X = 3)}$$

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$P(R = 10\theta \mid X = 3)$	.0437	.1229	.1852	.2074	.1875	.1383	.0794	.0307	.0049	0

Assessment: do you understand the comic?



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

What's so strange?

What we have obtained, the probability of each urn composition, is uncontroversial and straightforward.

Let us make this problem more interesting

> *An urn contains 10 marbles, R of which are red ($R \geq 1$), we draw a marble from the urn 5 times . . . we observe X red marbles. What can we say about R?*

R may or may not have been decided with a random mechanism, this is unimportant to us now (but it is what made the problem a standard problem before).

This is what we call a statistical problem (in today's language): we have observations which have been produced by a random mechanism which is not fully known and we want to induce its characteristics.

Bayes: inference for a probability

This is stated and more or less solved in Bayes essay as follows.

> *Given the number of times on which an unknown event has happened and failed:*
>
> *Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.*

If nothing is known of an event but that it has happened p times and failed q in $p+q$ or n trials, and from hence I judge that the probability of it's happening in a single trial lies between $\frac{p}{n}+z$ and $\frac{p}{n}-z$ my chance to be right is *greater* than $\frac{\sqrt{\mathrm{K}pq}\times h}{2\sqrt{\mathrm{K}pq}+hn^{\frac{1}{2}}+hn^{-\frac{1}{2}}}\times \overline{2\mathrm{H}-\frac{\sqrt{2}}{\sqrt{\mathrm{K}}}\times\frac{n+1}{n+2}\times\frac{1}{mz}\times 1-\frac{2m^2z^2}{n}}|^{\frac{n}{2}+1}$ and less than $\frac{\sqrt{\mathrm{K}pq}\times h}{2\sqrt{\mathrm{K}pq}-hn^{\frac{1}{2}}-hn^{-\frac{1}{2}}}$ multiplied by the 3 terms $2\mathrm{H}-\frac{\sqrt{2}}{\sqrt{\mathrm{K}}}\times\frac{n+1}{n+2}\times\frac{1}{mz}\times \overline{1-\frac{2m^2z^2}{n}}|^{\frac{n}{2}+1}+\frac{\sqrt{2}}{\sqrt{\mathrm{K}}}\times\frac{n}{n+2}\times\frac{n+1}{n+4}\times\frac{1}{m^3z^3}\times\overline{1-\frac{2m^2z^2}{n}}|^{\frac{n}{2}+2}$ where m^2, K, h and H stand for the quantities already explained.

Bayes solution was not actually very clear, the one from Laplace was better.

Probability of a female birth, Laplace

Laplace was the first to formulate a statistical problem and solve it with Bayesian statistics.

The question he poses was whether the probability of a female birth (θ) is or is not lower than 0.5.

The problem is analogous to that of the urn above, but for the fact that "there is a continuum of possible urn compositions".

He observed that in Paris, from 1745 to 1770 there were $493,472$ births, of which $241,945$ were girls and derived that

$$P(\theta \geq 0.5|\text{data}) \approx 1.15 \times 10^{-42}$$

giving "moral certainty" that $\theta < 0.5$.

Pierre-Simon Laplace (1749-1827) in *Essai philosophique sur les probabilités (1814)* gives a systematic treatment to the approach which we call Bayesian today.

Laplace statement of Bayes theorem

In *Essai philosophique sur les probabilités (1814)*, by Laplace, Bayes' theorem is formulated as

> Quand deux événemens dépendent l'un de l'autre ; la probabilité IVe Principe.
> de l'événement composé est le produit de la probabilité du premier événement, par la probabilité que cet événement étant arrivé, l'autre aura lieu. Ainsi, dans le cas précédent de trois urnes A, B, C, dont deux ne contiennent que des boules blanches, et dont une ne renferme que des boules noires ; la probabilité de tirer une boule blanche de l'urne C est $\frac{2}{3}$, puisque sur trois urnes, deux ne contiennent que des boules de cette couleur. Mais lorsqu'on a extrait une boule blanche, de l'urne C ; l'indécision relative à celle des urnes qui ne renferme que des boules noires, ne portant plus que sur les urnes A et B ; la probabilité d'extraire une boule blanche, de l'urne B est $\frac{1}{2}$; le produit de $\frac{2}{3}$ par $\frac{1}{2}$, ou $\frac{1}{3}$ est donc la probabilité d'extraire à-la-fois des urnes B et C, deux boules blanches.

Laplace and the probability of causes

but we have more

VIe Principe. Chacune des causes auxquelles un événement observé, peut être attribué, est indiquée avec d'autant plus de vraisemblance, qu'il est plus probable que cette cause étant supposée exister, l'événement aura lieu; la probabilité de l'existence d'une quelconque de ces causes, est donc une fraction dont le numérateur est la probabilité de l'événement, résultante de cette cause, et dont le dénominateur est la somme des probabilités semblables relatives à toutes les causes : si ces diverses causes considérées *à priori*, sont inégalement probables, il faut au lieu de la probabilité de l'événement, résultante de chaque cause, employer le produit de cette probabilité, par celle de la cause elle-même. C'est le principe fondamental de cette branche de l'analyse des hasards, qui consiste à remonter des événemens aux causes.

Laplace extended the scope of Bayes theorem to n possible causes of an event E.

back

The statistical problem

Back to

> *An urn contains 10 marbles, R of which are red ($R \geq 1$), we draw a marble from the urn 5 times . . . we observe X red marbles. What can we say about R?*

the point here is that R (θ) is not random in the sense of being generated through a random experiment (such as the die).

Rather, R (θ) is *unknown* to us.

How are we then to interpret the probability we attach to θ: $P(\theta|X = x)$?

Can it represent our beliefs on the value of θ?

According to some it could, according to other, this was nonsense.

Bayesian approach put aside

Since Laplace, and for a relatively long time, the Bayesian approach was put aside because it was deemed unscientific.

- The idea that the probability could be used to model ignorance/beliefs was ridiculed.
- Also, in order to get $P(\theta|X = x)$ we need to start from $P(\theta)$, a *prior* belief about θ (which comes before observations), this amounted to introducing an element of subjectivity in the analysis, which was, again, deemed unscientific.
- Moreover, there were practical problems: even for relatively simple problems, the Bayesian approach easily leads to intractable computations (Laplace used clever approximations to get his inference about θ)

Indice

New questions, new answers

Between XIX and XX-th centuries new field of application of statistical techniques rise

- quality control
- heredity and genetics

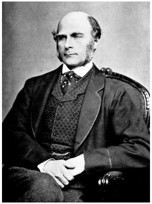
New approaches are developed in which

the parameter θ is a fixed number

inference is based on

- the likelihood: we compare $P(Data|Model)$ for the different models (In Bayesian statistics we compare $P(Model|Data)$),
- the performance in repeated sampling: procedures are evaluated based on fictitious repetitions of the experiment.

William Gosset (1876-1937)
Working for Guiness, he developed the Student-t distribution to evaluate quality of barley.



sir Francis Galton (1822-1911)
Founded the Eugenics Record Office in London, later the Galton labooratory. Develops linear regression.

Karl Pearson (1857-1936)
Introduces the concept of correlation and of goodness of fit.

Likelihood

The likelihood summarizes information on θ coming from $X = x$

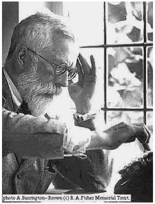
$$L(\theta) \propto P(X = x | R = 10\theta)$$

x	Urn composition (θ, share of red marbles) 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	.5905	.3277	.1681	.0778	.0312	.0102	.0024	.0003	.0000	0
1	.3280	.4096	.3601	.2592	.1562	.0768	.0284	.0064	.0004	0
2	.0729	.2048	.3087	.3456	.3125	.2304	.1323	.0512	.0081	0
3	.0081	.0512	.1323	.2304	.3125	.3456	.3087	.2048	.0729	0
4	.0005	.0064	.0284	.0768	.1562	.2592	.3601	.4096	.3280	0
5	.0000	.0003	.0024	.0102	.0312	.0778	.1681	.3277	.5905	1

From the likelihood alone we get answers in the form of

- maximum likelihood estimator: $\hat{\theta} = X/5 = 0.6$
- p-values: the p-value for the hypotheses $\theta \leq 0.2$ is 0.0579

sir Ronald Fisher (1890-1932)
introduces, among other things, the concepts of likelihood, analysis of variance, experimental design. Also, he originates the ideas of sufficiency, ancillarity, and information. His main works: *Statistical Methods for Research Workers (1925)*, *The design of experiments (1935)*, *Contributions to mathematical statistics (1950)*, *Statistical methods and statistical inference (1956)*



Repeated sampling principle

According to the repated sampling principle, we evaluate our procedures based on how they would behave in the long run with new sets of data.

Using the repeated sampling principle we can evaluate the performance of

- estimators → Mean Square Error
- confidence intervals → coverage probability

Neyman-Pearson hypotheses testing has the most evident link with repeated sampling:

- significance level is the relative frequencies with which we expect to reject a null hypotheses if we were to perform the test on a number of samples coming from a population for which the null is true;
- power is ...

Egon Pearson (1895-1980)
With Jerzy Neyman develops the theory of hypotheses testing.

Classical inference for female births

As far as the probability θ of a female birth is concerned, Laplace observations that in Paris, from 1745 to 1770 there were $493,472$ births, of which $241,945$ were girls would lead to

ML estimate : $\hat{\theta} = \frac{241{,}945}{493{,}472} = 0.4903$,

a 95 percent confidence interval:
$[0.4889, 0.4917]$,

p-value for the hypotheses $H_0 : \theta \geq 0.5$:
≈ 0.

Classical inference for female births

As far as the probability θ of a female birth is concerned, Laplace observations that in Paris, from 1745 to 1770 there were $493,472$ births, of which $241,945$ were girls would lead to

ML estimate : $\hat{\theta} = \frac{241{,}945}{493{,}472} = 0.4903$,

The best guess for θ is 0.4903,

a 95 percent confidence interval: $[0.4889, 0.4917]$,

we obtained an interval $[0.4889, 0.4917]$ as a realization of a random interval which has probability 95% of covering the true value of θ,

p-value for the hypotheses $H_0 : \theta \geq 0.5$: ≈ 0.

if $H_0 : \theta \geq 0.5$ were true, the probability of observing a sample as extreme as the one we saw would be ≈ 0.

What these tell us about θ is not obvious, where by this I mean that we need to make a further step to translate it in information on θ.

Classical approach or approaches?

Note that within the classical approach different views can be distinguished, this is particularly evident in hypotheses testing.

A Fisherian approach is to view the likelihood as central as a measure of evidence brought by the data. As such, a *p*-value is a measure of evidence against a given hypotheses.

The Neyman-Pearson view is behavioural, they devise a decision rule which controls the probability of error (not the overall one, but at least the conditional ones).

The above is a very simplistic summary, however it is true that the two approaches are incompatible and there have been harsh debates between the proponents.

Interpretation of results in BS and CS

> *A primary motivation for Bayesian thinking is that it facilitates a common sense interpretation of statistical conclusions (Gelman).*

Contrast interval estimation or hypotheses testing, BS tells us what we want to know, classical statistics does not, and it is likely that many "users" would incorrectly interpret classical statistics results the Bayesian way (luckily in many cases this is ok).

> *Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations (Gelman).*

Classical and Bayesian statistical inference, differences

In **CLASSICAL INFERENCE**

- the parameter is a constant.
- the conclusion is not derived within probability calculus rules (these are used in fact, but the conclusion is not a direct consequence)
- the likelihood and the probability distribution of the sample are used;

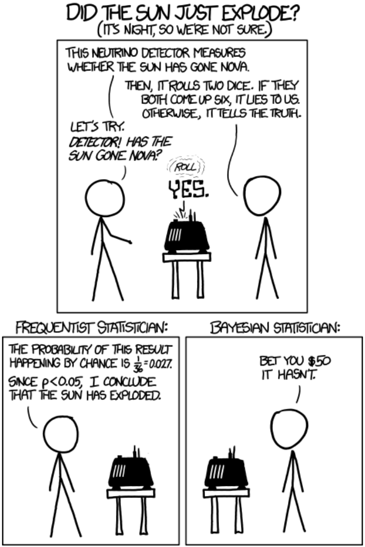
Framework to extract evidence from data.

In **BAYESIAN INFERENCE**

- the parameter is a r. v.
- the reasoning and the conclusion is an immediate consequence of probability calculus rules (of Bayes' theorem in particular);
- the likelihood and the prior distribution are used;

Framework to update information.

Assessment: do you understand the comic?



Today

Bayesian statistics was rediscovered in the XXth century.

Interesting uses included

- breaking the enigma codes during WW2
- combining historical and current information in setting insurance rates (the actuarial technique known as credibility theory turns out to be based on Bayesian reasoning)
- estimating the probability of events such as
  - probability of an aviation accident involving two planes (in the 50s)
  - probability of an accidental explosion of an H-bomb

The availability of computers helped a lot: Bayesian analytical results are available only for simple problems and the computational approaches are rather intensive, Monte Carlo methods are fundamental.

Different questions, different methods

Behind the choice of the preferred statistical approach, frequentist or Bayesian, there might be the question which is asked and the information available.

Fisher, working in genetics, was actually performing experiments

- no need for a prior information
- easy to frame the interpretation in the repeated sampling paradigm

Many applications of Bayesian inference in this period involved the need/desire to

- assess probability of events which were never observed
- combine different sources of information

Indice

Problems with probability

Although employed in special contexts did not achieve general acceptance (far from it).

One of the reasons why Bayesian statistics was difficult to accept is related to the frequentist definition of probability.

It is conceptually difficult to frame a prior distribution as a frequentist probability.

We need another probability!

Let us take a step back and discuss about this.

Limits of the frequentist interpretation of probability

Frequentist definition of probability applies to a narrow set of events, those which can be embedded, at least ideally, in a sequence of repetitions.

This is easily done for situations such as the toss of a coin with a head and a tail.

However we can easily think of events for which it does not work:

- Italy wins the next world cup;
- III WW between USA and Russia happen in the next two years;
- coin shows head when we know that the coin is double-headed or double tailed, but we do not know which.

It makes intuitive sense to consider these events, however a sequence of repetitions is not even thinkable.

The model is a random thing

In statistical applications, the situation is analogous to the two-headed/two-tailed coin example.

Suppose you want to estimate 'the number N of non UE citizens in Italy on 1/1/2019', N is then a random number in the Bayesian setting, one may object that

- N is not a random quantity (it is intrinsically a fixed number, albeit unknown);
- how can I specify a probability distribution on a non random quantity?
- The frequentist definition does not help in interpreting a probability distribution on N.

What is probability?

We can free us from the frequentist interpretation by taking the axiomatic definition of probability.

Definition (Probability)

Probability is a measure on a set of events (outcomes) such that

- it is non negative
- it is additive over mutually exclusive events
- sums to 1 over all possible mutually exclusive outcomes

This does tell us nothing about what can be used for.

We will all agree that we can use it to describe limiting relative frequencies of occurrence of events in repeated sequences, we may not agree on whether we can use it for something else?

Do we need another probability?

First, should we use it for something else?

We may take the stance that only events for which a sequence of ideal repetition is thinkable are permitted.

This is unsatisfying intuitively and practically since we have to deal with more general kinds of uncertainty (and they are relevant, think the H-bomb accidents) and we do routinely deal with them, that is we do take decisions based on some evaluation of such uncertain (non repeatable) events (think betting or weather forecasts).

Still, we might say that this kind of events is dealt with by common sense and is out of scope for a formal treatment by probability, but it might also be the case that probability could describe how common sense works.

Common sense: deductive logic → plausible logic

An example of common sense is an inference like

$$\left.\begin{array}{c}\text{if } A \text{ then } B\\ A \text{ is true}\end{array}\right\} \Rightarrow B \text{ is true}$$

which is described by deductive logic.

We also do inferences like the following

$$\left.\begin{array}{c}\text{if } A \text{ then } B\\ B \text{ is true}\end{array}\right\} \Rightarrow A \text{ more plausible}, \qquad \left.\begin{array}{c}\text{if } A \text{ then } B\\ A \text{ is false}\end{array}\right\} \Rightarrow B \text{ less plausible}$$

or even

$$\left.\begin{array}{c}\text{if } A \text{ then } B \text{ is more plausible}\\ B \text{ is true}\end{array}\right\} \Rightarrow A \text{ is more plausible}$$

This is a common type of reasoning (even in everyday life), it is sensible to try to describe it, that is, to quantify less/more plausible.

Probability as extension of true-false logic

If the aim is to represent the state of uncertainty on a "fact", then conditional probability is the only system which satisfies the axioms

I. States of uncertainty are represented by real numbers.
II. Qualitative correspondence with common sense.
    1. If the truth value of a proposition increases, its probability must also increase.
    2. In the limit, small changes in propositions must yield small changes in probabilities.
III. Consistency with true-false logic.
    1. Probabilities that depend on multiple propositions cannot depend on the order in which they are presented.
    2. All known propositions must be used in reasoning – nothing can be arbitrarily ignored.
    3. If, in two settings, the propositions known to be true are identical, the probabilities must be as well.

Coherence of bets

Another “proof” that probability as defined by the axioms is the only reasonable way to describe uncertainty is the Dutch book argument.

Let us define the probability of an event $P(E)$ as

- the price you would pay in exchange for a return of 1 if the event occurs and 0 otherwise,
- the price you would accept in exchange for having to pay 1 if the event occurs and 0 otherwise.

In other words, once you state $P(E)$ you would buy or sell the random amount $|E|$ in exchange for $P(E)$.

Suppose that you assess probabilities for a set of events, then if your probabilities do not satisfy the axioms it is possible to devise a combination of bets leading to a sure loss (gain). (That is, there is a combination of bets such that you would loose money no matter what happens.)

Probability to describe uncertainty

To some, these considerations make using probability to represent uncertainty a compelling choice and so Bayesian reasoning (which is a consequence of probability) the only reasonable way to update information (probabilities).

> *Bayesian Statistics offers a rationalist theory of personalistic beliefs in contexts of uncertainty, with the central aim of characterising how an individual should act in order to avoid certain kinds of undesirable behavioural inconsistencies (Beranardo and Smith).*

This leads quite naturally to the subjective definition of probability.

Subjective probability

For 'frequentist-friendly' events ('tail is observed when a coin is thrown')

- everyone (presumably) would agree on the value of the probability;
- the frequentist definition is intuitively applied;
- $\rightarrow$ this is an 'objective' probability.

For more general events events such as 'Italy wins the next world cup',

- it is still possible to state a probability;
- everyone would assign a different probability;
- the probability given by someone will change in time.

Bruno de Finetti (c. 1906-1985), Italian probabilist and actuary (for Generali) proposes the subjective definition of probability and the coherence framework, based on the bet interpretation (see *Theory of probability (1970)*).
In *Theory of probability* he wrote

> *Probability does not exist*

Subjective probability

One then accepts that the probability is not an objective property of a phenomenon but rather the opinion of a person and one defines

Definition (Subjective probability)

The probability of an event is, for an individual, his degree of belief on the event.

Bruno de Finetti (c. 1906-1985), Italian probabilist and actuary (for Generali) proposes the subjective definition of probability and the coherence framework, based on the bet interpretation (see *Theory of probability (1970)*).
In *Theory of probability* he wrote

> *Probability does not exist*

Nature of randomness

If the probability is a subjective degree of belief, it depends on the information which is subjectively available, and it is also clear that by random we mean not known for lack of information.

Given this, the following are random [=uncertain] quantities/events on which a probability distribution may be given

- date of birth of Manzoni
- number of non UE citizen in Italy today
- value of FIAT share will rise over the next month
- Italy's PIL growth in 2018
- exposure to mobile phones increase chances of getting cancer
- "9/11 was an inside job"

Then, there is no problem in saying that a parameter is random because is unknown.

Bayesian statistics and subjective probability

The subjective definition of probability is most compatible with the Bayesian paradigm, stated as follows

- the parameter to be estimated is a well specified quantity but is not known for lack of information
- a probability distribution is (subjectively) specified for the parameter to be estimated, this is called the *prior distribution*
- after seeing experimental results the probability distribution on the parameter is updated using Bayes' theorem to combine experimental results (likelihood) and prior distribution to obtain the *posterior distribution*.

Subjective probability and Bayesian update rule (which is actually a consequence of probability rules) establish a system to describe inference whose input are the prior beliefs and the data and the output is updated (posterior) beliefs.

Subjective Bayes

This approach is sometimes called subjective Bayes, it had a lot of followers since the 60s (see Lindley (1970, 2013), Savage (1972)).

In fact, for many it became the only coherent foundation of statistics, whereas the alternatives (Fisher, Neymann-Pearson and alike) looked like a collection of *ad hoc* tools lacking a proper justification.

This lead to the formation of two factions each rejecting the methods of the other, on part of the anti Bayesians the criticism were focused on the fact that admitting a subjective nature of the conclusions made them useless from a scientific point of view.

Even if we accept that subjective Bayes is a good description of reasoning under uncertainty in broad sense, it is still relevant to discuss whether this is acceptable in a scientific context: simplifying a bit, the role of prior distributions is central to this.

Indice

Need for prior

A critical issue in Bayesian inference (and one of the reason why it did not got acceptance at the beginning) is the need for prior information.

While classical statistics is only concerned with the information coming from the data, Bayesian statistics is a rule to update information based on the data: we must start somewhere.

This was seen as a major issue since it introduces an element of subjectivity in the analysis.

This will be discussed later, we make now two preliminary notes concerning

- where the prior comes from;
- the subjectivity (in the sense of arbitrariness) of results.

Source of prior information

Think of the female birth example again, but with the following sample:

> *In 2010 in Muggia (small city near Triest) 38 males and 47 females were born.*

According to likelihood inference (for θ, pr. of a female birth)

- the ML estimate is $\hat{\theta} = 0.553$
- the 95% c.i. is $[0.441, 0.659]$
- the p-value for $H_0 : \theta \geq 0.5$ is 0.8

What do you think of this information?

You probably think something along the lines of 'This sample tells me nothing'.

Why is that? Well, because you have, in fact, prior information.

We usually have prior information

In fact, it would be rare that we model a situation were we have no prior information at all.

Prior information may come from

- substantive knowledge about the process generating the data (we may be unsure about the exact mechanism but we usually know something),
- observations made in the past.

With this in mind, the prior distribution should not look so strange.

Subjectivity of results: vanishing priors

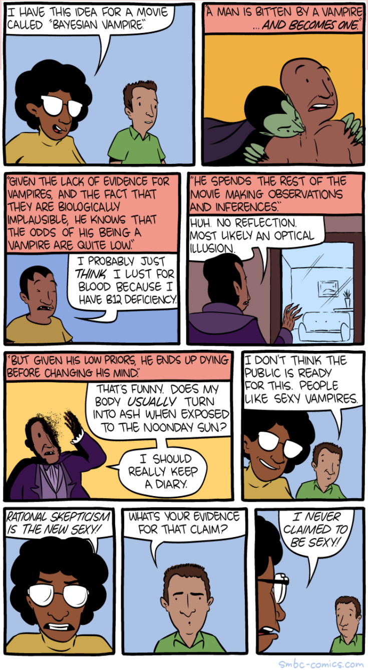
People do not like prior information because they do not like that two persons with the same data may reach different conclusions because they start from different prior informations.

While this is true, it is also true that, if the prior information is not unreasonable, the conclusions tend (asymptotically) to become equal as more data are gathered.

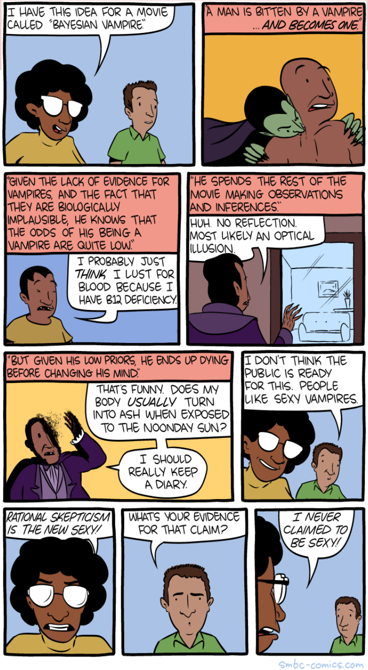
We will discuss what does “unreasonable” means, but the basic requirement is that we do not exclude any possibility (by assigning it a null prior probability).

Moreover, we will discuss how to distinguish prior distribution with respect to how much they weigh on the conclusion (how informative they are): there are methods to ensure that the conclusions are less influenced by the prior.

Assessment: on reasonable priors



Assessment: on reasonable priors



For another example see episode 17 of season 2 of *Star Trek: Voyager*, "Dreadnought" (in particular at approximately minute 20).

Subjectivity of results: standard priors

What we have said above assumes that the prior can (and should) represent the beliefs prior to the observations.

It is also possible to take a different approach, within the Bayesian paradigm.

In the example of female birth Laplace assumed a uniform prior on θ: he viewed this as a way to express indifference with respect to the possibilities.

This is kind of reasonable, although problematic for some aspects, the idea can be made more precise.

Subjectivity of results: make the prior irrelevant

The idea is that the prior does not need to convey information, rather it is regarded as a technical component of the model.

This idea lies behind the so called

- non informative priors
- reference priors

whose name tells it all, although maybe too optimistically:

- 'informativeness' is not a well defined concept, beware of attaching a precise meaning to the intuitive idea
- the posterior still depends on the prior

With these caveats let's say that particular distributions can be defined to avoid the subjective interpretation of the prior distribution.
This approach is sometimes called objective Bayes (or automatic Bayes).

Indice

What now?

We still lack a clear foundation of statistical inference which is agreed upon.

This is not only a abstract issue, it has been argued that it is at the root of practical problems in applications of statistics: the issue of hypotheses testing in applied science (see Nuzzo (2014); Goodman (2016), see also Pauli (2018) for an overview of the issue).

In what follows two modern overviews of the scenario on the foundations of statistics are discussed.

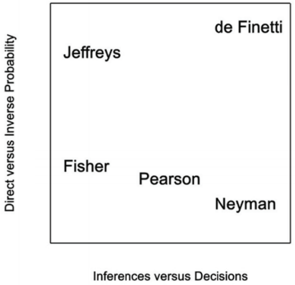
Map of approaches by Senn

Senn (2011) maps the various approaches we have briefly considered according to whether they focus on

- direct or inverse probabilities on one hand;
- on inference or decision on the other hand (here the latter means that we are interested in the consequences of using a certain criterion).

(Keep in mind that any scheme like this is bound to oversimplify.)

- Classical statistics
  - Likelihood: Fisher
  - Hyp. test: Neyman-Pearson
- Bayesian statistics
  - Objective: Jeffrey
  - Subjective: de Finetti



Different approaches for different questions (Royall)

Another way of looking at the different approaches is based on the questions they can answer to, Royall (2004) distinguishes methods based on the question they seek to answer.

Three questions can be asked to the data

(1) What should I believe?

(2) How should I behave?

(3) What is the evidence?

Royall stance is that

- (3) is answered by the likelihood alone,
- (1) is answered by the posterior (needs the likelihood and the prior),
- (2) needs the posterior and the costs of errors.

(Note that (2) is different from the 'decision' Senn has in mind.)

Mixing

The good note is that the factions are no more: to some extent at least, statistician are keen on taking what is relevant from each approach.

In practice this has meant that

- it is now deemed reasonable by many Bayesians to assess model adequacy (this is incoherent with looking them as beliefs, which can not be wrong),
- frequentist properties of Bayesian procedures are studied.

Today

On pragmatic grounds, it is reasonable to use whatever approach is best suited for the situation at hand, this is the most common attitude among applied statisticians.

It is also reasonable to interpret Bayesian techniques as modelling techniques rather than a philosophical stance (thus disconnecting it from the subjective interpretation), in this sense the role of the prior can be downplayed, from a source of information to a regularization device (part of a model).

We will take this attitude in what follows (keeping in mind, however, that the Bayesian approach is the only correct one and all other procedures are justified only as approximations of the B. ones).

Indice

Compelling nature of Bayesian reasoning

Recall that

> *Bayesian Statistics offers a rationalist theory of personalistic beliefs in contexts of uncertainty, with the central aim of characterising how an individual should act in order to avoid certain kinds of undesirable behavioural inconsistencies (Bernardo and Smith).*

we noted that this has lead some to argue for taking Bayesian reasoning as the foundation of statistical inference.

In fact, we have said that Bayesian reasoning could be the paradigm to extend deductive logic to plausible logic.

Role of Bayesian reasoning

These circumstances lead some to think that Bayesian reasoning could (should) be used as the paradigm of inductive logic, that is, beyond its statistical scope: a recipe for human reasoning in general.

Let us then consider contexts were beliefs are important (central) and discuss to what extent Bayesian reasoning fits practice:

- science (epistemology): where interest lies in the truth of a theory,
- law: where interest lies in the belief on guilt or innocence of a defendant.
- diagnostic: where interest lies on whether a tested person is ill

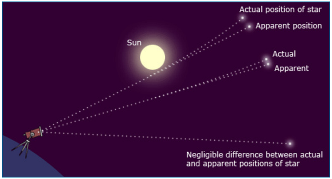
The question is whether (to what extent, under which conditions) Bayesian reasoning can describe (model) the reasoning process of a scientist (judge/juror, clinician) who accept/rejects theories (decides over guilt/innocence, diagnose patients).

Science: Eddington experiment

In 1919 the astronomer Eddington made, during a solar eclipse, a series of measurements of light deflection.

Under the circumstances of the experiment he knew that

- N: Newton law predicted a deflection of 0.875
- $\bar{N}$: Einstein relativity predicted a deflection of 1.75

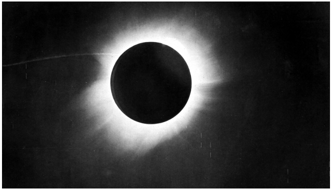


Let us assume, for the sake of the examples, that these two theories are a priori equally likely, that is, $P(N) = P(\bar{N}) = 0.5$.

Science: Eddington experiment

In 1919 the astronomer Eddington made, during a solar eclipse, a series of measurements of light deflection.

Eddington obtained 5 measurements of the deflection with mean 1.98 and standard error 0.16, with these, assuming a Gaussian error, Bayes rule dictates that



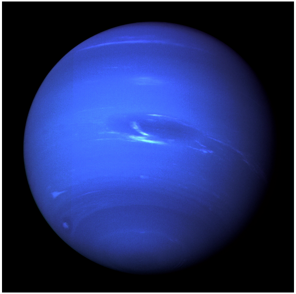
$$\frac{P(N|\text{data})}{P(\bar{N}|\text{data})} = \frac{\phi((1.98 - 0.875)/0.16)}{\phi((1.98 - 1.75)/0.16)} \times \frac{P(N)}{P(\bar{N})} = 0.81 \times 10^{-10} \times 1$$

Evaluating Eddington observations using Bayes rule leads to a strong belief that Newton theory is wrong.

Science: Neptune

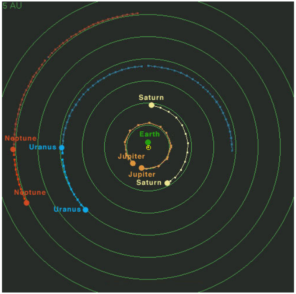
At the beginning of 19th century observations of the orbit of Uranus showed that it was not following the path predicted by Newtonian theory.

A naïve reasoning may lead to looking at these observations as a falsification of Newton.



Science: Neptune

At the beginning of 19th century observations of the orbit of Uranus showed that it was not following the path predicted by Newtonian theory.



A naïve reasoning may lead to looking at these observations as a falsification of Newton.

This, however, was considered very unlikely at that time, so other explanations were sought, including the existence of a further planet in the solar system: the astronomers Leverrier and Adams then computed mass and orbit of a planet which, if present, could explain the observations on Uranus.

Based on their prediction the planet Neptune was discovered in 1846.

Science and Bayesian reasoning

The first example (kind of) works because there are two clear alternative theories. Assuming that reality is either Einstenian or Newtonian, Bayesian updating is a reasonable (the only reasonable) description of the thought process which leads to the scientific conclusion.

The second example does not work because there is a theory–Newton–and no precise alternative; if we used the pseudo-alternative “Newton is false” we would reach a conclusion (incidentally we would wrongly reach a correct conclusion), but this misses the actual thought process which involved looking for alternatives (however unlikely).

The tenet is that Bayesian reasoning would work if we could precisely define all alternative theories a priori, which is unrealistic in general: *“[...] because it is very hard to be sufficiently imaginative and because life is short.”*.

Law: Regina v DJA

The question is whether the defendant is guilty of a rape (G).

Prior to any evidence, he is one of 200 000 possible culprits (male population of the area in a suitable age range), so $P(G) = 1/200\,000$

Evidence:

- DNA match (M)
- not recognized by the victim (neither during a parade, nor after, the victim described the attacker as a man in his twenties, DJA was 37)
- alibi from his girlfriend

Law: Regina v DJA

The question is whether the defendant is guilty of a rape (G).

Prior to any evidence, he is one of 200 000 possible culprits (male population of the area in a suitable age range), so $P(G) = 1/200\,000$

Evidence:

- DNA match (M)
  prosecutor: $P(M|\bar{G}) = 1/2 \times 10^8$, $P(M|G) = 1$
  defence: $P(M|\bar{G}) = 1/2 \times 10^6$, $P(M|G) = 1$
- not recognized by the victim (neither during a parade, nor after, the victim described the attacker as a man in his twenties, DJA was 37)
  $P(\bar{R}|G) = 0.1$, $P(\bar{R}|\bar{G}) = 0.9$
- alibi from his girlfriend
  $P(A|G) = 0.25$, $P(\bar{A}|\bar{G}) = 0.5$

Law: Regina v Denis John Adams

Combining the evidence according to

$$\frac{P(G|M\bar{R}A)}{P(\bar{G}|M\bar{R}A)} = \frac{P(G)}{P(\bar{G})}\frac{P(M|G)}{P(M|\bar{G})}\frac{P(\bar{R}|G)}{P(\bar{R}|\bar{G})}\frac{P(A|G)}{P(A|\bar{G})}$$

leads, depending on which probability is chosen for $P(M|\bar{G})$, to different guilt probabilities.

- For the prosecutor: $P(G|M\bar{R}A) \approx 0.98$,
- for the defence: $P(G|M\bar{R}A) \approx 0.36$.

Whether 0.98 is high enough to convict is dubious, 0.36 is certainly not.

Here the reasoning works, the point is that most if not all the probabilities which are involved have to be elicited and are debatable.

Diagnostic: Hanahaki disease

Hanahaki diseases has prevalence of 1% in a population.

A test is available such that the probability that it comes out positive is

- 90% if the tested individual is affected by Hanahaki: $P(T|H) = 0.9$
- 10% if the tested individual is not affected by Hanahaki: $P(T|\bar{H}) = 0.1$

John Smith is tested and is positive, should he worry seriously?

How likely is he to be affected?

$$P(H|T) = 0.08.$$

Moral of the story

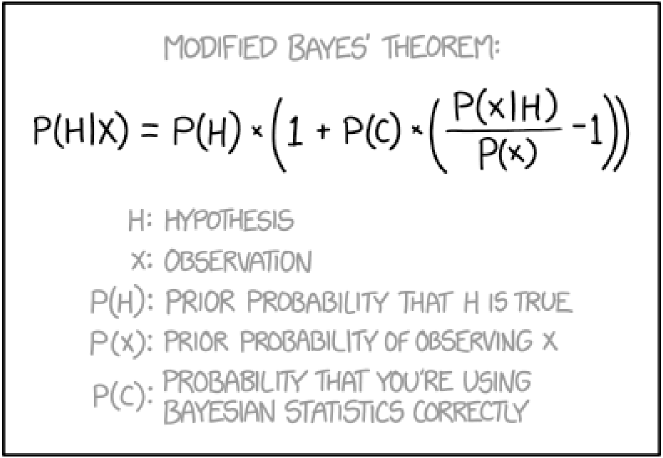
Bayesian reasoning can not describe all human reasoning.

Bayesian reasoning is a compelling framework but only

- limited to the hypothesis under consideration (and limited by the reasonableness of such hypotheses),
- conditional on the likelihood given to the evidence under the various hypothesis.

In statistical terms this translates in conditional on the model specification, hence the importance of evaluating the fit to check on model adequacy.

Moral in pictures



Further readings

For the history of Bayesian statistics, with examples, see McGrayne (2011).

For a modern presentation of the subjective Bayes approach see Jaynes (2003); Lindley (2013), further readings include De Finetti (1974); Jeffreys (1998); Lindley (1970); Savage (1972)

The classical approach to inference is described in Cox (2006), its principles are discussed in Mayo and Cox (2006); Mayo (2011). The works which originated the approach are also readable although with some difficulty: Fisher (1922, 1925); Neyman et al. (1933).

A modern approach to Bayesian inference is in Gelman et al. (2013), see also Gelman et al. (2011) and Gelman and Shalizi (2013) for the role of Bayesian inference in science.

References

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge university press.

De Finetti, B. (1974). *Theory of probability*. John Wiley & Sons.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A 222*(594-604), 309–368.

Fisher, R. (1925). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh.

Gelman, A. et al. (2011). Induction and deduction in bayesian data analysis. *Rationality, Markets and Morals 2*(67-78), 1999.

Gelman, A. and C. R. Shalizi (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology 66*(1), 8–38.

Gelman, A., H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Good, I. (1988). The interface between statistics and philosophy of science. *Statistical Science 3*(4), 386–397.

Goodman, S. N. (2016). Aligning statistical and scientific reasoning. *Science 352*, 1180–1181.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Lindley, D. V. (1970). Introduction to probability and statistics from a bayesian viewpoint. Technical report.

Lindley, D. V. (2013). *Understanding uncertainty*. John Wiley & Sons.

Mayo, D. G. (2011). Statistical science and philosophy of science: Where do/should they meet in 2011 (and beyond)?

Mayo, D. G. and D. R. Cox (2006). Frequentist statistics as a theory of inductive inference. In *Optimality*, pp. 77–97. Institute of Mathematical Statistics.

Mayo, D. G. and A. Spanos (2006). Severe testing as a basic concept in a neyman–pearson philosophy of induction. *The British Journal for the Philosophy of Science 57*(2), 323–357.

References (cont.)

McGrayne, S. B. (2011). *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press.

Neyman, J., E. S. Pearson, and K. Pearson (1933). IX. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231*(694-706), 289–337.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature 506*(7487), 150–152.

Pauli, F. (2018). The p-value case, a review of the debate: Issues and plausible remedies. In C. Perna, M. Pratesi, and A. Ruiz-Gazen (Eds.), *Studies in Theoretical and Applied Statistics*, Cham, pp. 95–104. Springer International Publishing.

Royall, R. (2004). The likelihood paradigm for statistical evidence. *The nature of scientific evidence: Statistical, philosophical, and empirical considerations*, 119–152.

Savage, L. J. (1972). *The foundations of statistics*. Dover Publications.

Senn, S. (2011). You may believe you are a bayesian but you are probably wrong. *Rationality, Markets and Morals 2*(48-66), 27.

Final note of caution