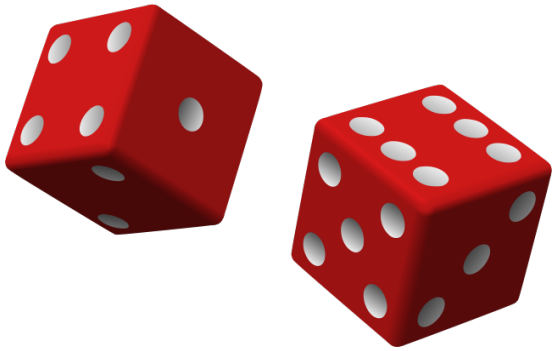
Introduction to Bayesian Inference

A natural, but perhaps unfamiliar view on probability and statistics

Michiel Botje
Nikhef, PO Box 41882, 1009DB Amsterdam
m.botje@nikhef.nl

Topical Lectures, Nikhef, 11-13 December 2013

What makes Bayesian statistics different from Frequentist statistics?



They are different views on the concept of probability

So, what then is probability?

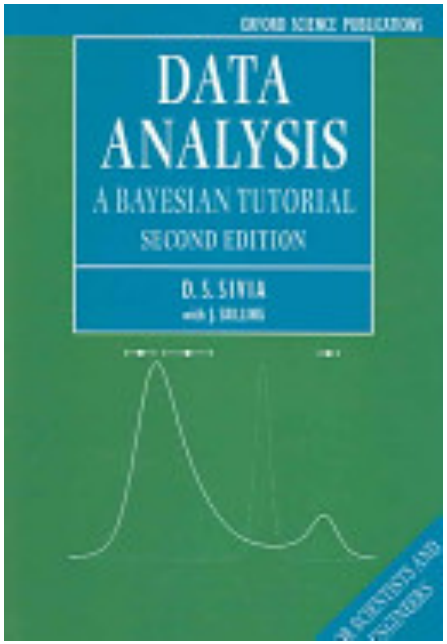
- Mathematical
  - Anything that obeys the Kolmogorov axioms (probability calculus).
- Operational
  - Frequentist

    Relative frequency of the occurrence of an event in repeated observations under identical conditions. In this view, probability is a characteristic of random events taking place in the world around us.
  - Bayesian

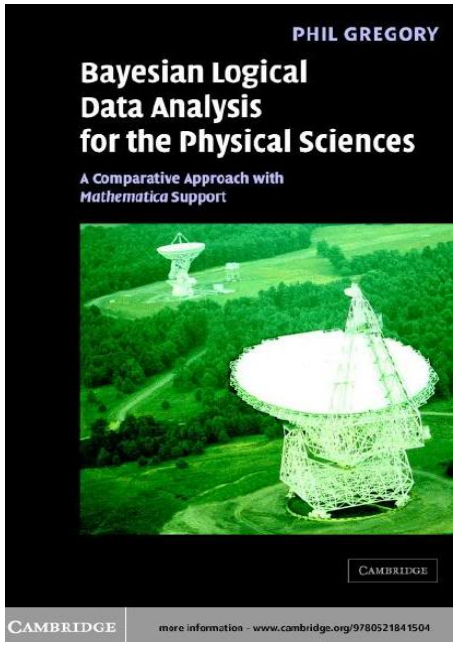
    Plausibility that a proposition is true, given the available information. This is not necessarily a property of the world around us, but more reflects what we know about this world.

This has far reaching consequences for data analysis since Bayesians can assign probabilities to propositions (hypotheses) while Frequentists cannot

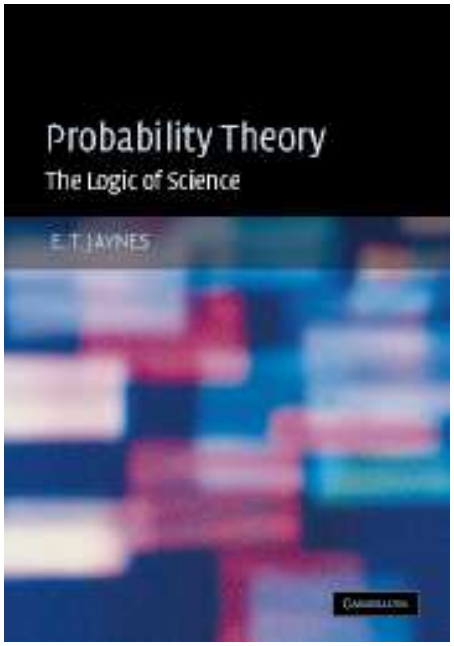
Textbooks on Bayesian statistics



D.S. Sivia
Data Analysis
A Bayesian Tutorial



P. Gregory
Bayesian Logical Data analysis for the Physical Sciences



E.T. Jaynes
Probability Theory
The logic of science
(advanced)

Plan

- Lecture 1

  Basics of logic and Bayesian probability calculus
- Lecture 2

  Probability assignment
- Lecture 3

  Parameter estimation
- Lecture 4

  Glimpse at a Bayesian network, and model selection

Lecture notes and answers to selected exercises can be downloaded from

`http://www.nikhef.nl/~h24/bayes`

1

Basics of logic and probability calculus from a Bayesian perspective

Inference is the logical process by which we draw conclusions from a set of input propositions

- Sufficient input: apply deductive logic to arrive at a definite conclusion which is either true or false
- Insufficient input: apply inductive logic which will leave us in a state of uncertainty about our conclusion.

This latter process is called plausible inference:
The art of reasoning in the presence of uncertainty

Deduction is the stuff of mathematical proofs

Plausible inference is the stuff of everyday life:

What are the chances

- that I gain from this investment …
- that I am still alive after crossing this road …
- that it will rain tomorrow …
- that I have really seen the Higgs …
- …

How to deal with such questions …

1. Try to foresee all possibilities that might arise
2. Judge how likely each is, based on everything you can see and all your past experience

Plausible Inference

3. In the light of this, judge what the probable consequences of various actions would be
4. Now make your decision

Decision theory

Deductive and inductive reasoning

Deductive reasoning

- ❑ Sunflowers are yellow
- ❑ This flower is a sunflower
- ✓ This flower is yellow

Input information is sufficient to draw a firm conclusion

Inductive reasoning

- ❑ Sunflowers are yellow
- ❑ This flower is yellow
- ✓ This flower is perhaps a sunflower

Input information is not sufficient to draw a firm conclusion

- ✓ It is more probable that this flower is a sunflower

Let us now dissect these two reasoning processes …

Elementary propositions

- In Aristotelian logic the most elementary proposition consists of a subject and a predicate, which says something about the subject

  A := 'This person is a male'

- Such a statement *A* can be either true (1) or false (0)

- The operation of negation ~*A* turns a true proposition into a false one and vice versa

Compound propositions

There are 16 ways to combine two propositions
Five are given here

$\top$	Tautology
$\bot$	Contradiction
$\vee$	Logical or
$\wedge$	Logical and
$\Rightarrow$	Implication

A	B	$A \top B$	$A \bot B$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$
0	0	1	0	0	0	1
0	1	1	0	0	1	1
1	0	1	0	0	1	0
1	1	1	0	1	1	1

Building block of inference: implication

A ⇨ *B:* If *A* is true then *B* is true.

Now let be given *A* ⇨ *B* and we observe *B*.

What can we say about *A*?

- If *B* is false we can conclude that *A* is false (in fact, ~*B* ⇨ ~*A*)
- If *B* is true we cannot conclude that *A* is true (or false) but we may conclude that it is more likely that *A* is true

A	B	$A \Rightarrow B$
0	0	1
0	1	1
1	0	0
1	1	1

Note that you cannot <u>invert</u> implication

Thus from $A \Rightarrow B$ does not follow that $B \Rightarrow A$

Instead, it follows that

$\sim B \Rightarrow \sim A$

A	B	$A \Rightarrow B$
0	0	1
0	1	1
1	0	0
1	1	1

Let us now add a minor premise to the implication and construct a syllogism

Deductive argument is a chain of strong syllogisms

Major premise	If A is true then B is true
Minor premise	A is true
Conclusion	B is true

Major premise	If A is true then B is true
Minor premise	B is false
Conclusion	A is false

This is the stuff of mathematical proofs

Inductive argument contains at least one weak syllogism

Major premise	If A is true then B is true
Minor premise	A is false
Conclusion	B is less plausible

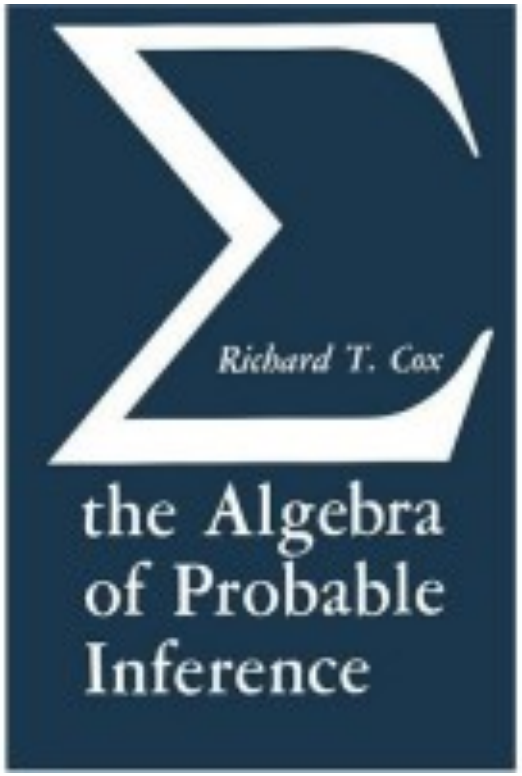
Major premise	If A is true then B is true
Minor premise	B is true
Conclusion	A is more plausible

Introduce $P(X|I)$ as a measure of the plausibility that X is true, given the information I

Cox' desiderata (1946)

1. Plausibility $P(A|I)$ is a real number
2. Plausibility $P(A|I)$ increases when more supporting evidence for the truth of A is supplied
3. Consistency
   a. The plausibility $P(A|I)$ depends only on available information and not on how the conclusion is reached
   b. All relevant information is used (not some selection) and all irrelevant information is discarded
   c. Equivalent states of knowledge lead to the same plausibility assignment

This completely fixes the algebra of plausibility



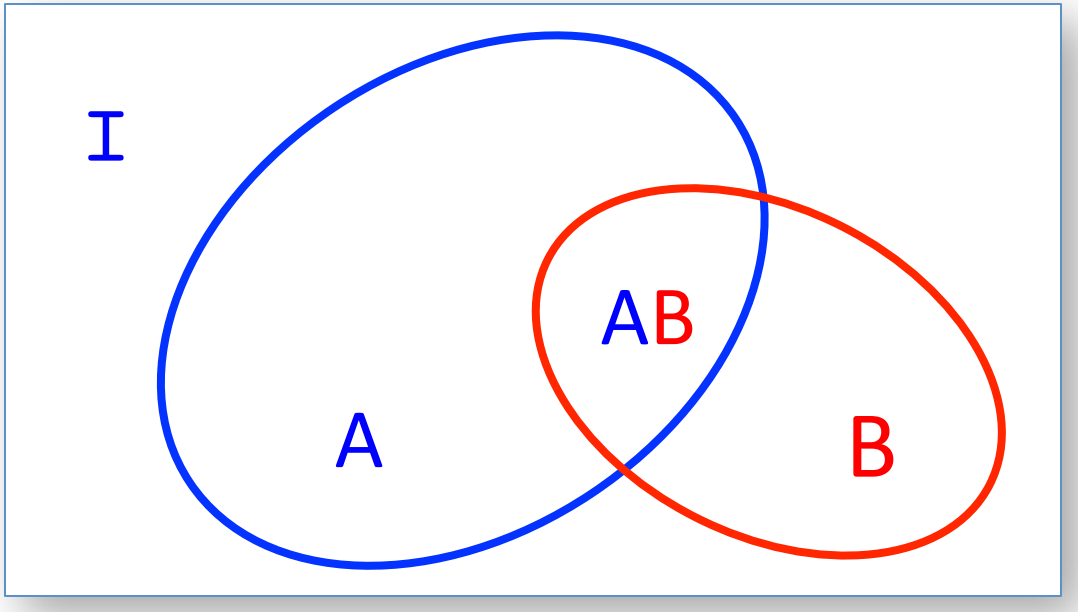
Johns Hopkins Press (1961)
Still in print and can also be found on the web

From his desiderata Cox was able to develop the algebra of plausibility and showed that this algebra is based on the Kolmogorov axioms, just like probability

This is the basis of the Bayesian view of probability as a measure of plausibility or degree of belief

Kolmogorov axiom: Sum rule

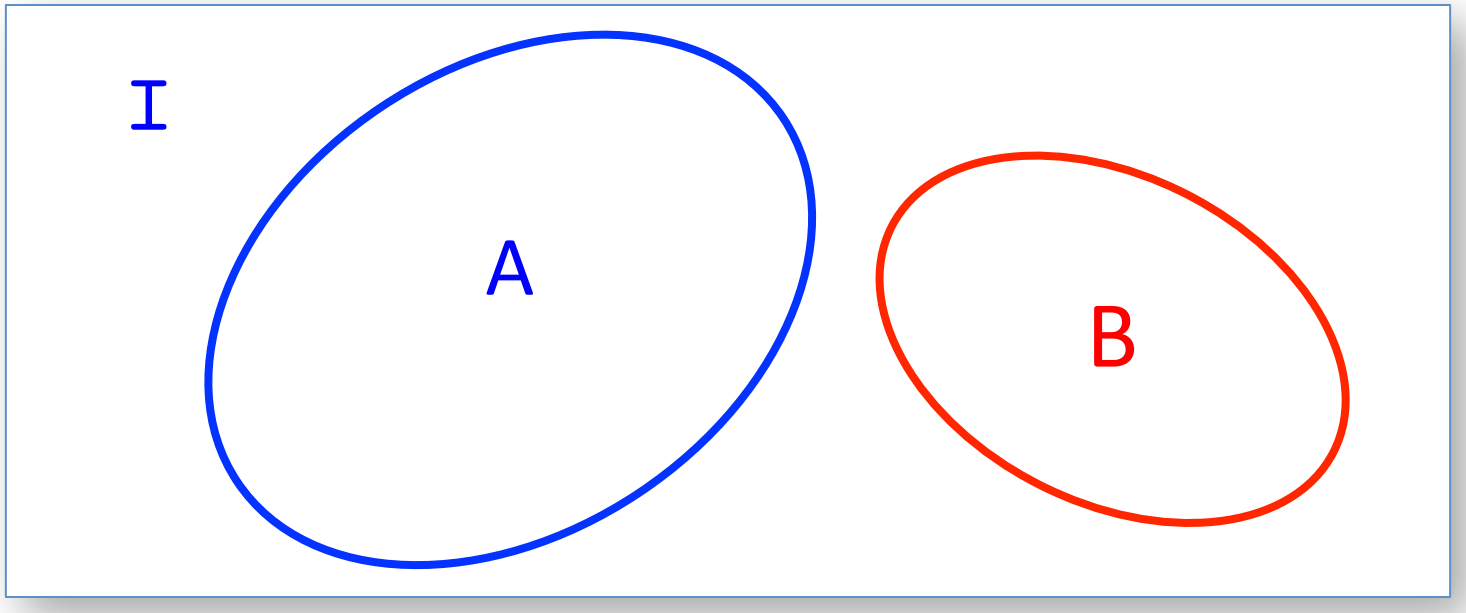
$$P(A \vee B|I) = P(A|I) + P(B|I) - P(AB|I)$$



Also stated as $P(A|I) + P(\sim A|I) = 1$

Mutually exclusive propositions

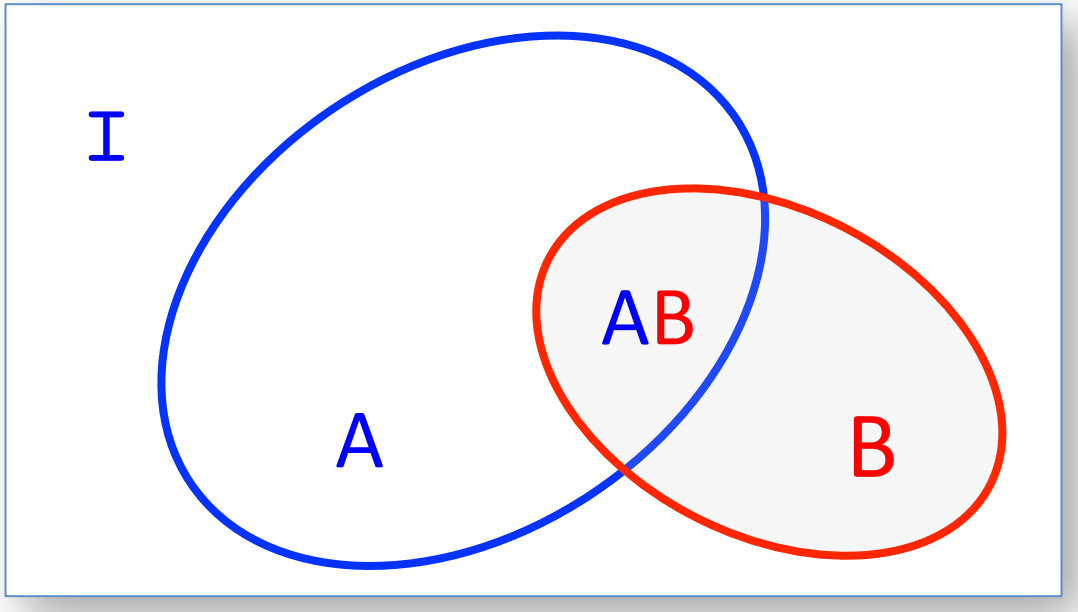
$$P(AB|I) = 0$$



$$P(A \vee B|I) = P(A|I) + P(B|I)$$

Kolmogorov axiom: Product rule

$$P(AB|I) = P(A|BI)\ P(B|I)$$



$$P(A|I) = \boxed{\text{A}} \,/\, \boxed{\text{I}}$$

$$P(B|I) = \boxed{\text{B}} \,/\, \boxed{\text{I}}$$

$$P(A|BI) = \boxed{\text{AB}} \,/\, \boxed{\text{B}}$$

Product rule terminology

$$P(AB|I) = P(A|BI)\ P(B|I)$$

$P(AB\|I)$	Joint probability
$P(A\|BI)$	Conditional probability
$P(B\|I)$	Marginal probability

Independent propositions

- *A* and *B* are dependent when learning about *A* means that you will also learn about *B*
- This is logical dependence, not to be confused with causal dependence
- For independent propositions *A* and *B*, the joint probability factorises

$$P(A|BI) = P(A|I) \quad \Rightarrow \quad P(AB|I) = P(A|I)\ P(B|I)$$

Bayes' theorem

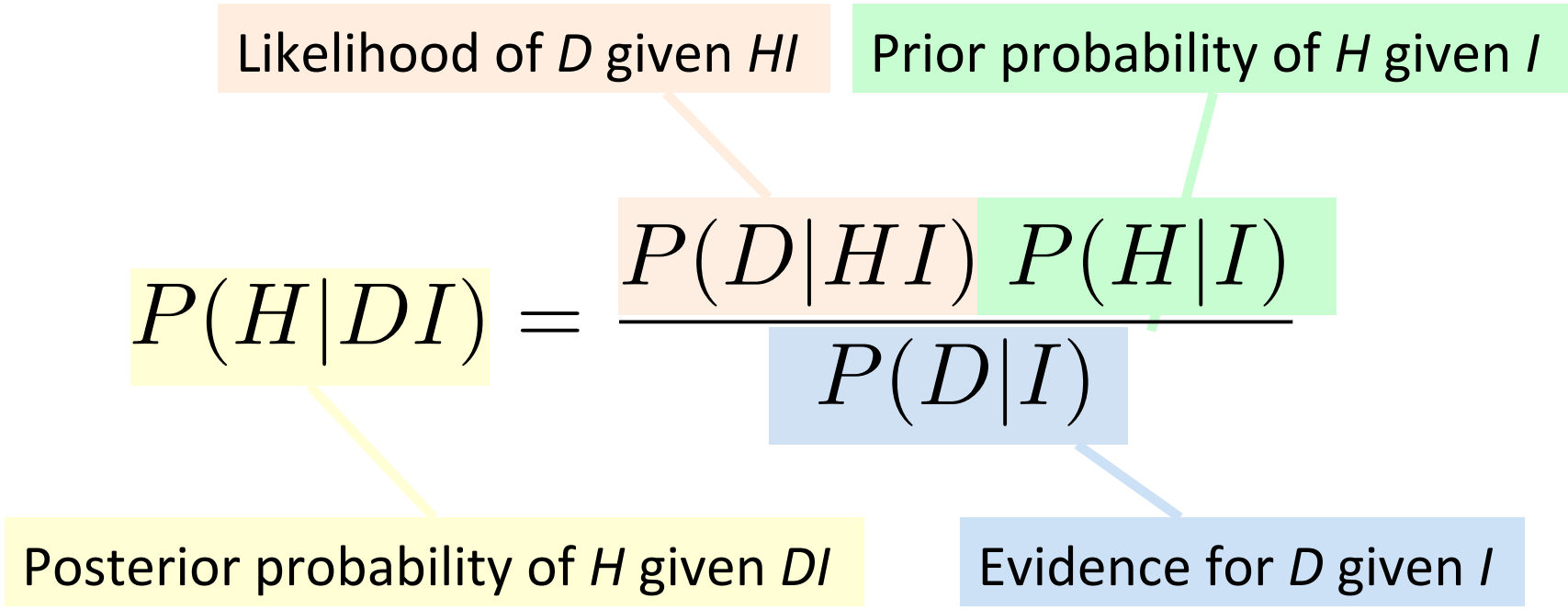
- Because *AB* = *BA*, we have

$$\begin{aligned} P(AB|I) &= P(A|BI)\ P(B|I) \\ &= P(B|AI)\ P(A|I) \end{aligned}$$

⇒ Law of conditional probability inversion

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

Bayes' theorem models a learning process



Update knowledge on hypothesis *H* with data *D*

Historical remark

- Bernoulli (1713), one of the founders of probability theory, wondered how deductive logic could be used in inductive inference
- He never solved the problem and the answer was given later (1763) by Bayes' Theorem

$$P(H|DI) = \frac{P(D|HI)\,P(H|I)}{P(D|I)}$$

- The likelihood $P(D|HI)$ represents deductive reasoning from cause (H) to effect (D)

Use of BT $P(H|DI) = \frac{P(D|HI)\ P(H|I)}{P(D|I)}$

- For a Frequentist, probabilities are properties of random variables
- *H* is a hypothesis which is either true or false for all repeated observations so it is not a RV
- Therefore *P*(*D*|*HI*) makes sense but not *P*(*H*|*DI*)

⇒ This invalidates access to *H* via Bayes' theorem

⇒ For a Bayesian *P*(*H*|*DI*) makes perfect sense since it is a measure of the plausibility that *H* is true

Lets take an example from real life…

Does Mr White have AIDS?

- Mr White goes to a doctor and tests positive on AIDS
- The test is known to be fully efficient in detecting AIDS
- Thus poor Mr White is sure to have AIDS

Right?

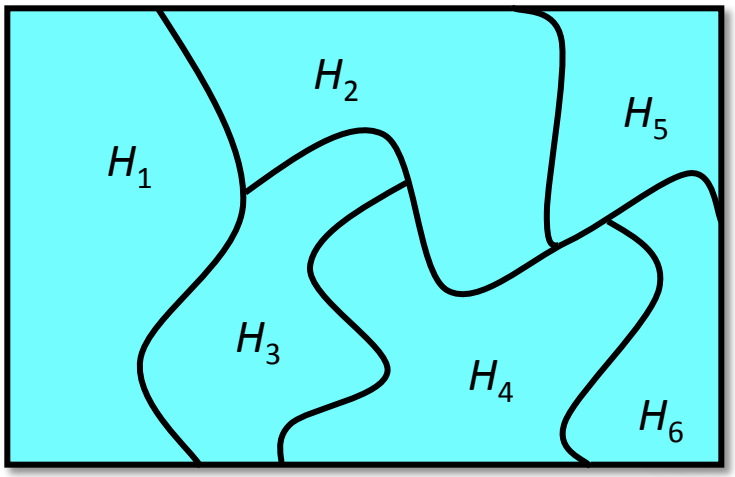
If you think this is true you have not (yet) understood that

$$P(\text{AIDS}|\text{positive}) \neq P(\text{positive}|\text{AIDS})$$
$$P(\text{rain}|\text{clouds}) \neq P(\text{clouds}|\text{rain})$$
$$P(\text{woman}|\text{pregnant}) \neq P(\text{pregant}|\text{woman})$$

To tackle mr Whites problem we first have to properly construct our hypothesis space …

Expand *H* into a set of exhaustive and exclusive hypotheses

Another way of stating this: expand *H* into a set $\{H_i\}$ such that one and only one hypothesis H_i is true

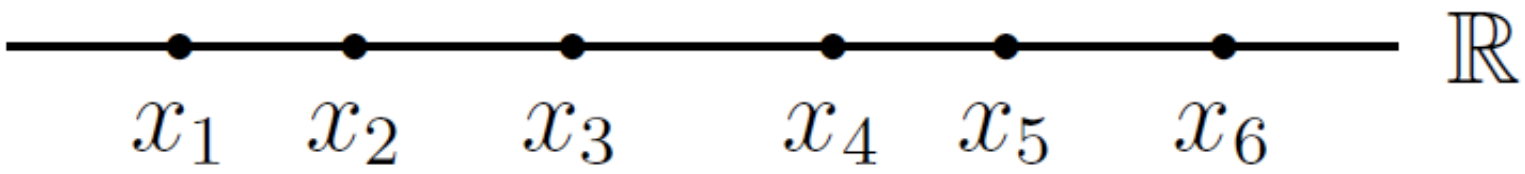


$$H = H_1 \vee H_2 \vee H_3 \cdots \qquad H_i H_j = \delta_{ij}$$

Because *H* itself is (by definition) a tautology we find from the sum rule

$$\sum_i P(H_i|I) = 1$$

Trivial example: divide an interval on the real axis into, say, 5 bins …



$$H := x_1 \leq x < x_6$$

$$H_i := x_i \leq x \leq x_{i+1}$$

- Marginalisation (integrate out the set $\{H_i\}$)

$$\sum_i P(DH_i|I) = P(D\sum_i H_i|I) = P(D|I)$$

- Expansion (is like closure relation in QM)

$$P(D|I) = \sum_i P(DH_i|I) = \sum P(D|H_iI)\ P(H_i|I)$$

$$\langle D|I\rangle = \sum \langle D|H_i\rangle\langle H_i|I\rangle$$

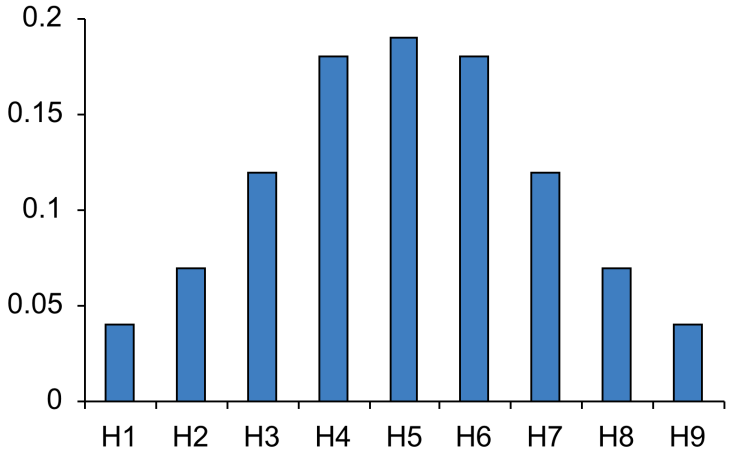
- Bayes' Theorem for the H_i in the set

$$P(H_i|DI) = \frac{P(D|H_iI)\ P(H_i|I)}{P(D|I)} = \frac{P(D|H_iI)\ P(H_i|I)}{\sum_i P(D|H_iI)\ P(H_i|I)}$$

Posterior as a function of the H_i

$$P(H_i|DI) \propto P(D|H_iI)\ P(H_i|I)$$

- If we calculate the posterior for all hypotheses H_i in the set we obtain a spectrum of probabilities
- In such a calculation the likelihood $P(D|H_iI)$ is a function of H for fixed data



- The denominator $P(D|I)$ can *post factum* be calculated as a normalization constant by integrating the posterior

Back to Mr White…

Does Mr White have AIDS?

- Mr White goes to a doctor and tests positive on AIDS
- The test is (almost) fully efficient in detecting AIDS
- What is the probability that Mr White has AIDS?

Let us now collect the Bayesian ingredients and note in passing that the problem as stated above is, in fact, ill posed

We need to specify

- Exhaustive and mutually exclusive set of hypotheses
  - A: Mr White has AIDS
  - $\sim A$: Mr White does not have AIDS
- Likelihoods for the entire hypothesis space
  - $P(T|A)$ probability for a positive test when the patient has AIDS (98%, say)
  - $P(T|\sim A)$ probability for a positive test when the patient does not have AIDS (3%, say)
- Prior probability for Mr White to have AIDS We can take the number of cases in Holland divided by the number of Dutch inhabitants, say $P(A|I)$ = 1% (I just take some round number, lets hope its less!)

Does Mr White have AIDS? (i)

- The problem can now be stated as follows
  - Before the test the prior probability for mr White to have AIDS is $P(A|I) = 1\%$ (fraction of infected Dutchies)
  - What is the posterior probability $P(A|T)$ of Mr White to have AIDS after he has tested positive?

$$P(A|TI) = \frac{P(T|AI)\,P(A|I)}{P(T|AI)\,P(A|I) + P(T|\bar{A}I)\,P(\bar{A}|I)}$$

- So, the probability for Mr White to have AIDS is only

$$P(A|TI) = \frac{0.98 \times 0.01}{0.98 \times 0.01 + 0.03 \times 0.99} = 0.25$$

Does Mr White have AIDS? (ii)

- It is illustrative to imagine a test on 10 000 individuals using the probabilities given on the previous slide
- Here is a table of outcomes (contingency table)

	T	~T	
AIDS	98	2	100
~AIDS	297	9603	9900
	395	9605	10000

$$\frac{\text{AIDS,Positive}}{\text{Positive}} = \frac{98}{395} = 25\%$$

- We leave it as an exercise to understand from this table the conditional probabilities entering Bayes' Theorem

Does Mr White have AIDS? (iii)

- Take an extreme: the test is always negative for a person without AIDS, that is, $P(T|\sim A) = 0$
- It is now certain that Mr White (if positive) has AIDS!

$$P(A|TI) = \frac{0.98 \times 0.01}{0.98 \times 0.01 + 0.0 \times 0.99} = 1$$

- We could have deduced this from Aristotelian logic since $P(T|\sim A) = 0$ encodes the implication $\sim A \Rightarrow \sim T$ or, equivalently, $T \Rightarrow A$: Mr White must have AIDS
- This example shows that the limiting case of Bayesian logic is Aristotelian logic (Boolean algebra, that is)

Does Mr White have AIDS? (iv)

- Take an extreme: the test is always positive for a person with AIDS, that is, $P(T|A) = 1$
- Does this mean that Mr White (if positive) has AIDS?

No! $$P(A|TI) = \frac{1 \times 0.01}{1 \times 0.01 + 0.03 \times 0.99} = 0.25$$

- $P(T|A) = 1$ encodes the implication $A \Rightarrow T$ but we know already from Aristotelian logic that we cannot invert to $T \Rightarrow A$
- T (positive test) does not imply A (Mr White has AIDS)
- This example shows that the limiting case of Bayesian logic is Aristotelian logic (and quantifies uncertainty as a bonus)

Does Mr White have AIDS? (v)

- Take an extreme: nobody has AIDS: $P(A|I) = 0$
- Is it now certain that Mr White has no AIDS whatever the outcome of the test

$$P(A|TI) = \frac{0.98 \times 0.0}{0.98 \times 0.0 + 0.03 \times 1.0} = 0$$

- Because $P(A|I) = 0$ encodes a prior certainty (namely that it is impossible to have AIDS) it follows that no amount of evidence to the contrary can ever change this
- This example shows that additional information can never change a prior certainty!

Mr White tests positive a second time…

- Taking the posterior of the first test (25%) as the prior for the second test gives

$$P(A|TTI) = \frac{0.98 \times 0.25}{0.98 \times 0.25 + 0.03 \times 0.75} = 0.92$$

- We could also use the likelihood for two positive tests (assuming that the tests are independent)

$$P(TT|AI) = 0.98^2 \quad \text{and} \quad P(TT|\bar{A}I) = 0.03^2$$

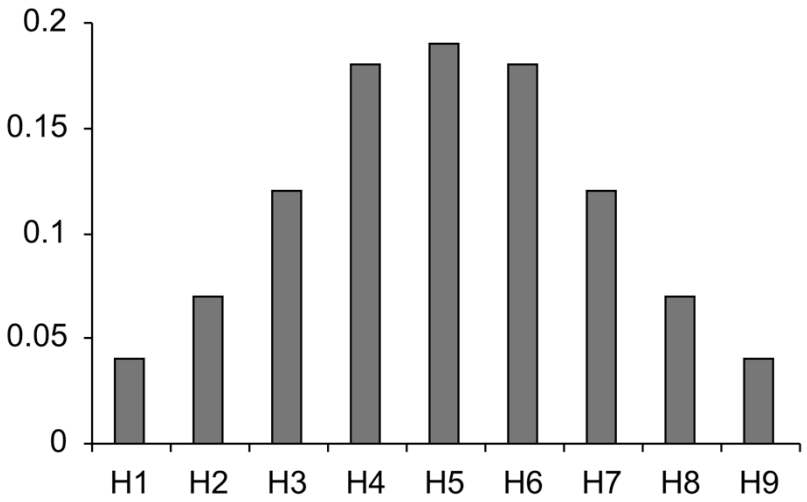
- Feeding this and $P(A|I) = 1\%$ into Bayes' Theorem gives

$$P(A|TTI) = \frac{0.98^2 \times 0.01}{0.98^2 \times 0.01 + 0.03^2 \times 0.99} = 0.92$$

- The two results are the same, as required by Cox' desideratum 3[a] of consistency: plausibility does not depend on the path of reasoning but only on information

Continuous variables

- Up to now we have worked with hypotheses or, equivalently, with discrete variables
- Continuous variables are introduced by defining a probability density through the integral relation



$$P(a \leq x < b) = \int_a^b p(x|I)\ \mathrm{d}x$$

Formulas for continuous variables

- Product rule $$p(x, y|I) = p(x|y, I)\, p(y|I)$$
- Normalization $$\int p(x|I)\, \mathrm{d}x = 1$$
- Expansion or marginalization

$$p(x|I) = \int p(x, y|I)\, \mathrm{d}y = \int p(x|y, I)\, p(y|I)\, \mathrm{d}y$$

- Bayes' Theorem

$$p(y|x, I) = \frac{p(x|y, I)\, p(y|I)}{\int p(x|y, I)\, p(y|I)\, \mathrm{d}y}$$

How to get the infinitesimals right…

- It should be remembered that probability calculus applies to probabilities and not to probability densities
- To turn a density into a probability, just add the infinitesimals in your expressions but only for the random variables (in front of the vertical bar) and not for the conditional variables (behind the vertical bar)

$$\left.\begin{array}{l} p(x,y)\mathrm{d}x\mathrm{d}y \\ p(x|y,I)\mathrm{d}x \end{array}\right\} \text{ are probabilities but not } p(x|y,I)\mathrm{d}y$$

$$p(y|x,I) = \frac{p(x|y,I)\ p(y|I)}{\int p(x|y,I)\ p(y|I)\mathrm{d}y}$$

- The result of Bayesian inference is the posterior, often summarised in terms of mean, mode, variance, covariance or quantile
- We will not present this here and refer to Section 3 of the write-up

the information entropy (Section 5.3). However, it is clear that when the likelihood is narrow compared to the prior, it does not matter very much what prior distribution we chose. On the other hand, when the likelihood is so wide that it competes with any reasonable prior then this simply means that the experiment does not carry much information on the subject. In such a case it should not come as a surprise that answers become dominated by prior knowledge or assumptions. Of course there is nothing wrong with that, as long as these prior assumptions are clearly stated (alternatively one could try to look for better data!).

The prior also plays an important role when the likelihood peaks near a physical boundary or, as very well may happen, resides in an unphysical region (likelihoods related to neutrino mass measurements are a famous example; for another example see Section 6.4 in this write-up). In such cases, the information on the boundary is mostly (or exclusively) contained in the prior and not in the data.

With these remarks we leave the Bayesian-Frequentist debate for what it is and refer to the abundant literature on the subject, see *e.g.* [13] for recent discussions.

3 Posterior Representation

The full result of Bayesian inference is the posterior distribution. However, instead of publishing this distribution in the form of a parametrisation, table, plot or computer program it is often more convenient to summarise the posterior—or any other probability distribution—in terms of a few parameters.

3.1 Measures of location and spread

The **expectation value** of a function $f(x)$ is defined by[16]

$$< f > = \int f(x)\, p(x|I)\, \mathrm{d}x. \tag{3.1}$$

Here the integration domain is understood to be the definition range of the distribution $p(x|I)$. The ***k*-th moment** of a distribution is the expectation value $< x^k >$. From (2.19) it immediately follows that the zeroth moment $< x^0 > = 1$. The first moment is called the **mean** of the distribution and is a location measure

$$\mu = \bar{x} = < x > = \int x\, p(x|I)\, \mathrm{d}x. \tag{3.2}$$

The **variance** σ^2 is the second moment about the mean

$$\sigma^2 = < \Delta x^2 > = < (x-\mu)^2 > = \int (x-\mu)^2\, p(x|I)\, \mathrm{d}x. \tag{3.3}$$

The square root of the variance is called the **standard deviation** and is a measure of the width of the distribution.

[16] We discuss here only continuous variables; the expressions for discrete variables are obtained by replacing the integrals with sums.

16

From Bayesian to Frequentist …

- Probability taken as a degree of belief is not new because it was the view of the founders of probability theory (Bernoulli 1713, Bayes 1763, Laplace 1812) and very successful
- So why was it abandoned at the beginning of the 20th century in favor of Frequentist theory?

Three main objections to Bayesian theory

1. Not clear why probability calculus would apply to probability as degree of belief

   ⇒ Solved by Cox algebra of plausibility (1946)

2. Plausibility is subjective

   ⇒ Yes, in the sense that it depends on the available information which I may have, but not you

3. Not clear how to assign the prior

   ⇒ Still a large field of research; we use symmetry arguments or MAXENT (see later)

Frequentists overcome these problems by defining probability as the limiting frequency of occurrence

1. Rules of probability calculus do apply

2. Objective because probability is a property of random processes in the world around us

3. Cannot access a hypothesis via BT because a hypothesis is not a random variable ⇨ no prior probability to worry about

But how can you access a hypothesis without using Bayes' theorem?

The Frequentist answer is to construct a statistic

- Function of the data and thus a random variable with (presumably) known probability distribution
- Estimator : estimate the value of a parameter, e.g. sample mean to estimate the mean of a sampling distribution
- Test statistic : test validity of hypothesis or discriminate between hypotheses, e.g. chi-squared, *F*-statistic, *t*-statistic, *z*-statistic, etc.
- No general rule how to construct a statistic but base choice on properties like consistency, bias, efficiency, robustness (estimator) or power, type-1, type-2 error probabilities (test statistic).

OK, but probability theory now has to be supplemented with the theory of statistics!

What we have learned

- In Bayesian inference plausibility is identified with probability so that we can use probability calculus
- The three important ingredients are
  - Expansion (express compound probability in terms of elementary probabilities)
  - Bayes theorem (conditional probability inversion)
  - Marginalisation (integrate out nuisance parameters)
- Bayesians can assign probabilities to hypotheses and use BT to invert likelihood $P(D|HI)$ to posterior $P(H|DI)$
- Frequentists cannot do this and have to use a statistic to access the hypothesis H

- ✓ Lecture 1

  Basics of logic and Bayesian probability calculus
- Lecture 2

  Probability assignment
- Lecture 3

  Parameter estimation
- Lecture 4

  Glimpse at a Bayesian network, and model selection