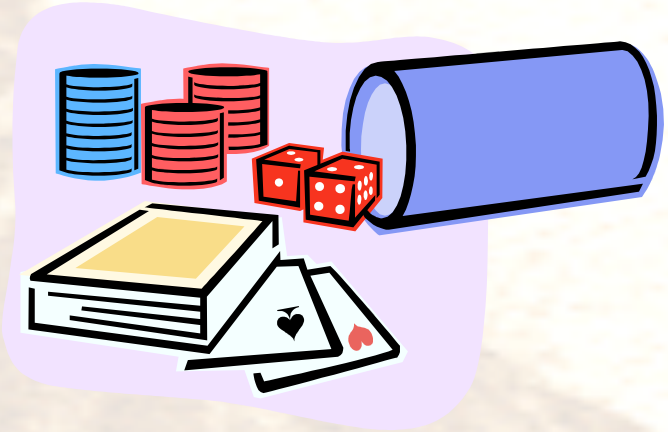
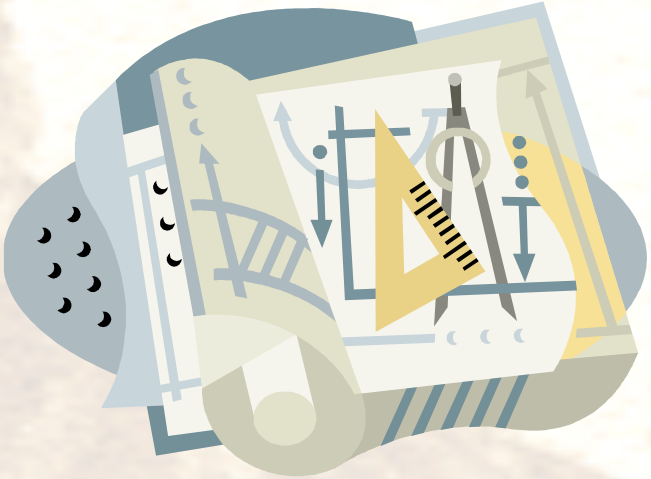
Bayesian inference: concepts

Some early history



- Bernoulli (1654-1705)
- Bayes (1701-1761)
- Laplace (1749-1827)
- Prediction problem ("forward probability"):
  - If the probability of an outcome in a single trial is *p*, what is the relative frequency of occurrence of this outcome in a series of trials?
- Learning problem ("inverse probability"):
  - Given a number of observations in a series of trials, what are the probabilities of the different possible outcomes?

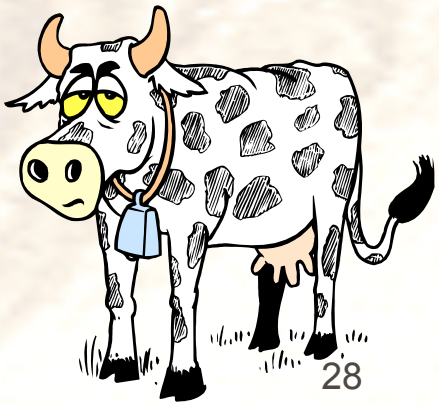
The Bayes rule



- Axioms of probability theory:
  - The sum rule:
    - ✓ $P(A \mid C) + P(\bar{A} \mid C) = 1$
  - The product rule:
    - ✓ $P(AB \mid C) = P(A \mid BC)\, P(B \mid C)$
- The Bayes rule:
  - $P(A \mid BC) = P(A \mid C)\, P(B \mid AC) / P(B \mid C)$
- A rule for updating our beliefs after obtaining new information
- H = hypothesis (model), I = background information, D = data (observations):
  - $P(H \mid D\, I) = P(H \mid I)\, P(D \mid H\, I) / P(D \mid I)$

On plausible reasoning

- "*The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, non of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind*" (James Clerk Maxwell)
- Probabilistic reasoning is intuitively easy to understand, but on the other hand intuition may be a poor guide when facing probabilistic evidence
- "*Inside every non-Bayesian there is a Bayesian struggling to get out*" (Dennis V. Lindley)



Do I have a good test?

- A new home HIV test is assumed to have "95% sensitivity and 98% specificity"
- a population has HIV prevalence of 1/1000. If you use the test, what is the chance that someone testing positive actually has HIV?

Test continued ...

- P(HIV + | test HIV +) = ?
- We know that
  - P(test HIV + | HIV +) = .95
  - P(test HIV + | HIV -) = .02
- from Bayes we have learned that we can calculate the probability of having HIV given a positive test result by

$$= \frac{.95 \times .001}{.95 \times .001 + .02 \times .999} = .045$$

Thus finally

- thus over 95% of those testing positive will, in fact, not have HIV
- the right question is:

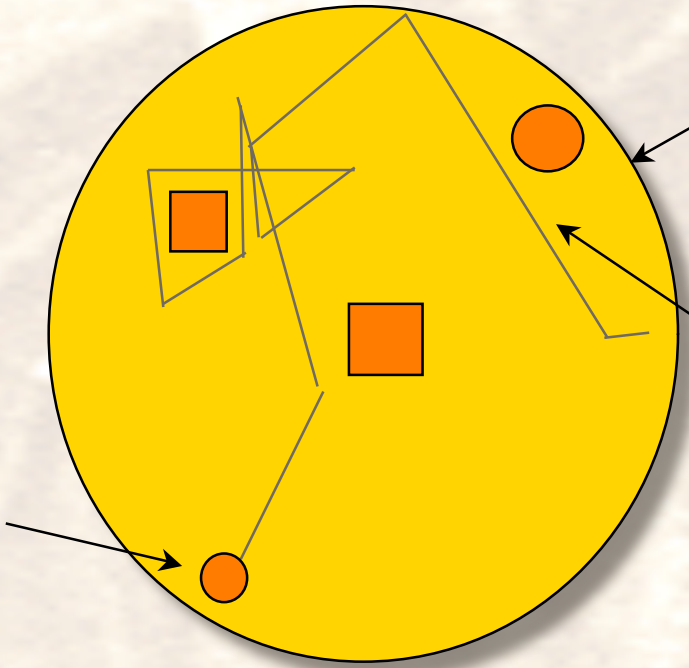
> How should the test result change our belief that we are HIV positive?

Fundamental questions

What is the model space?

How do we compare models?

How do we search?



Bayesian answers

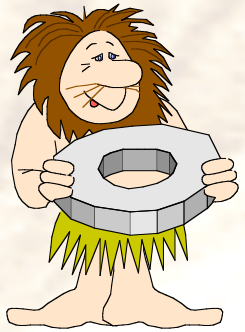
- Model family (space) is made explicit
- Comparison criteria is a probability
- No restrictions on the search algorithm

Classical statistics answers

- Model family is implicit (normal distributions)
- Comparison criteria is fit to data, deviation from "random" behavior, "model index"
- Simple deterministic "greedy" algorithms

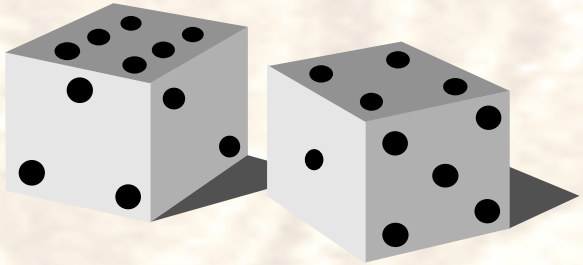
Bayesian?

- Probabilities can be interpreted in various ways:
- Frequentist interpretation (Fisher,Neyman, Cramer)
- "Degree of belief" interpretation (Bernoulli, Bayes, Laplace, Jeffreys, Lindley, Jaynes)



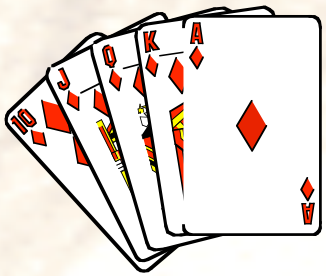
Frequentist says ...

- The long-run frequency of an event is the proportion of the time it occurs in a long sequence of trials - probability is this frequency
- probability can only be attached to "random variables" - not to individual events



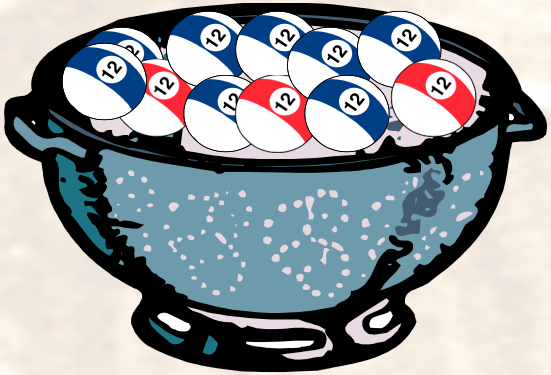
Bayesian says ...

- an event x = state of some part of the universe
- probability of x is the degree of belief that event x will occur
- probability will always depend on the state of knowledge
- p(x|y,C) means probability of event x given that event y is true and background knowledge C



Frequentist language for solving problems

- P(data | model)
- sampling distributions



Model

Data

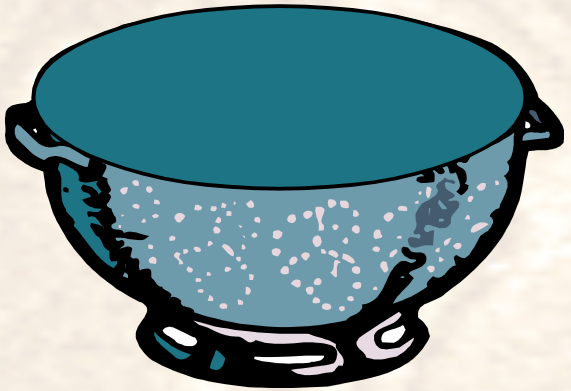
?

Bayesian language for solving problems

- Bayesian: P(data | model) & P(model | data)

Prior knowledge 

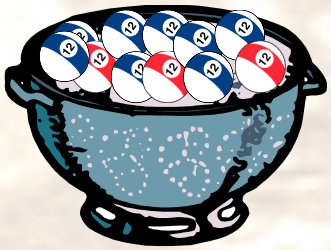
Data 



?

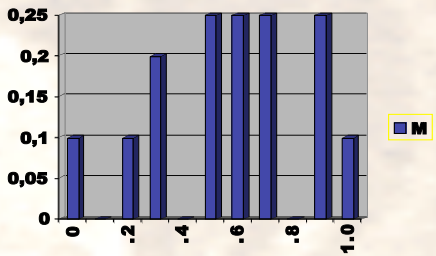
Isn't this what I already do? No.

Model



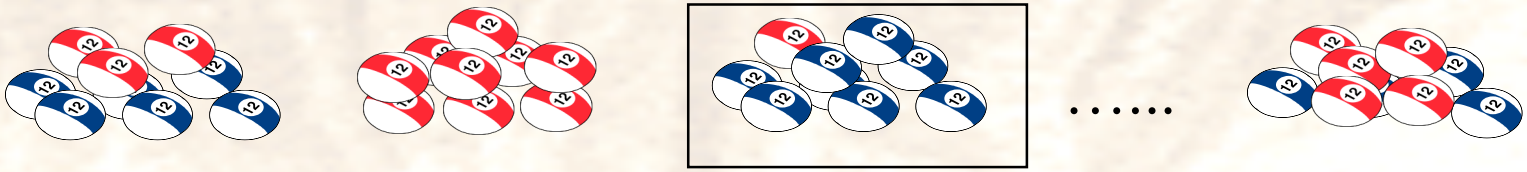
Parameter

"Sampling distribution of the estimator"



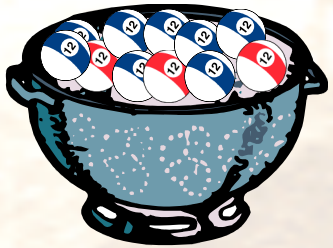
Estimator (function of data)

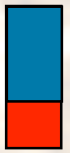
Data



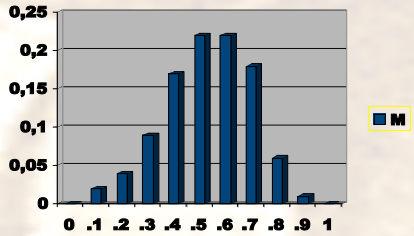
“The Bayesian way”

Model



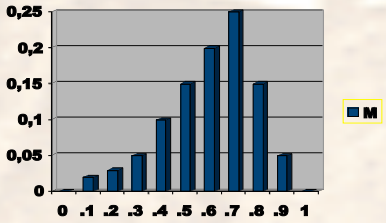


Parameter

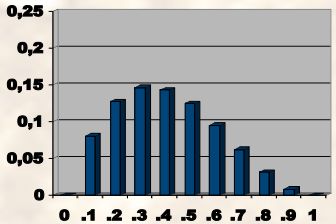


Posterior distribution of the parameter

Likelihood

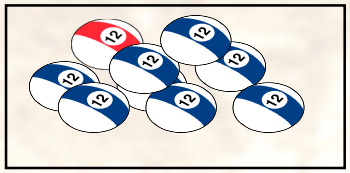


×



Prior distribution of the parameter

Data



Reasons for using probability theory

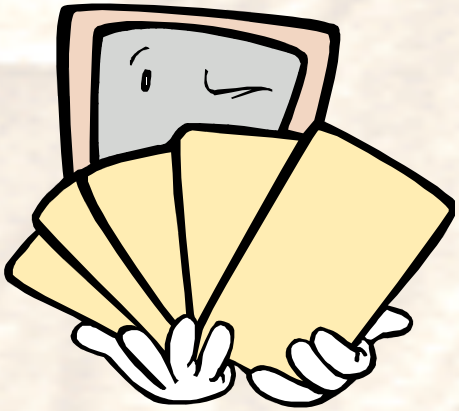
- Cox/Jaynes argument: probability is an appealing choice as the language for plausible inference
- Berger argument: Decision theory offers a theoretical framework for optimal decision making, and decision theory needs probabilities
- Pragmatic argument: it is a very general framework and it works

Qualitative properties of p.r.

- D1. Degrees of plausibility are represented by real numbers
- D2. Direction of inference has a qualitative correspondence with common sense
  - For example: if Plaus(A | C´) > Plaus(A | C) and Plaus(B |C') = Plaus(B | C), then Plaus(AB | C') > Plaus(AB | C)
  - Ensures consistency in the limit (with perfect certainty) with deductive logic
- D3. If a conclusion can be inferred in more than one way, every possible way should lead to the same result
- D4. All relevant information is always taken into account
- D5. Equivalent states of knowledge must be represented by equivalent plausibility assignments

Real questions

- Q1: Given plausibilities Plaus(A) and Plaus(B), what is Plaus(AB)?
- Q2: How is Plaus(~A) related to Plaus(A)?



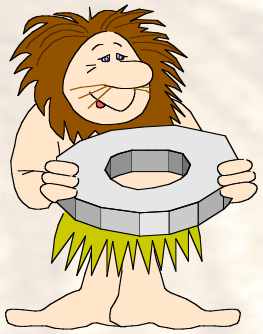
Cox/Jaynes/Cheeseman argument

- Every allowed extension of Aristotelian logic to plausibility theory is isomorphic to Bayesian probability theory
- Product rule (answers question Q1)
  - $P(AB \mid C) = P(A \mid BC)\, P(B \mid C)$
- Sum rule (answers question Q2)
  - $P(A \mid C) + P(\bar{A} \mid C) = 1$

Bayesian inference: How to update beliefs?

- Select the model space
- Use Bayes theorem to obtain the posterior probability of models (given data)

$$P(M_i \mid D) = \frac{P(M_i) \, P(D \mid M_i)}{P(D)}$$



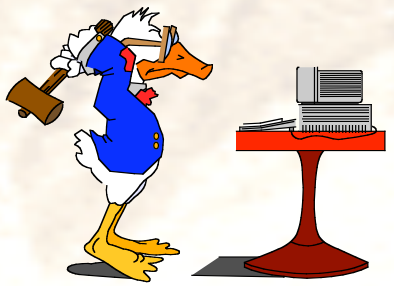
Posterior distribution is “the result” of the inference; what one needs from the posterior depends on what decisions are to be made

The Bayesian modeling viewpoint

- Explicitly include prediction (and intervention) in modeling

Models are a means (a language) to describe interesting properties of the phenomenon to be studied, but they are not intrinsic to the phenomenon itself.

"All models are false, but some are useful."



(Being predictive ...)

- True prediction performance is a function of future data, not a model fit to current data

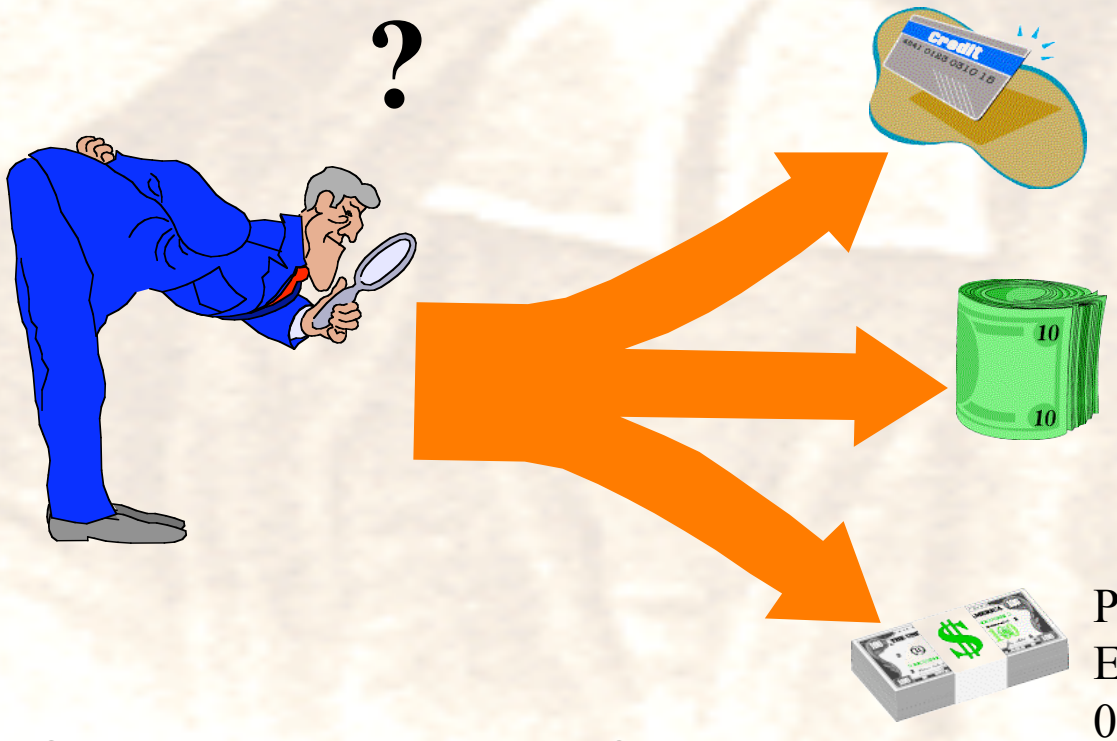
Good predictive models describe useful regularities of the data generating mechanism, models that give a high probability to the *data* have learnt to memorize it

Bayesian decision making for kids

- assign a benefit for every possible outcome (for every possible decision)
- assign a probability to every possible outcome given every possible decision
- what is the best decision?

Decision theory argument

- Decision theory offers a theoretical framework for optimal decision making



P($100)=0.1, P($50)=0.9
Expected utility:
0.1*$100+0.9*$50=$55

P($200)=0.2, P($5)=0.8
Expected utility:
0.2*$200+0.8*$5=$44

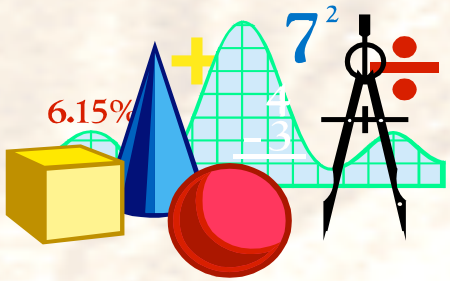
P($80)=0.5, P($50)=0.5
Expected utility:
0.5*$80+0.5*$50=$65

Optimal actions

- Optimal policy: choose the action with **maximal expected utility**
- The Dutch book argument: betting agencies **must be** Bayesians
- Where to get the utilities? (decision theory)

“Pragmatic” reasons for using probability theory

- The predictor and predicted variables (the inference task) do not have to be determined in advance
  - probabilistic models can be used for solving both classification (discriminative tasks), and configuration problems and prediction (regression problems)
  - predictions can also be used as a criteria for Data mining (explorative structures)



More pragmatic reasons for using probability theory

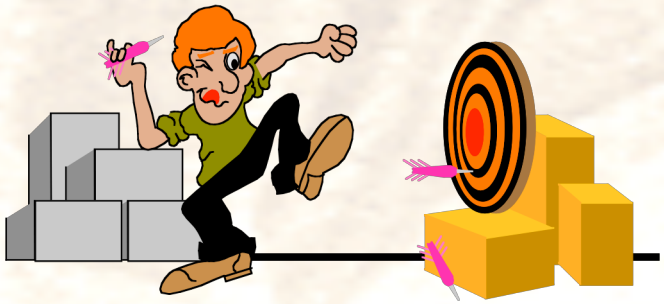
- consistent calculus
  - creating a consistent calculus for uncertain inference is not easy (the Cox theorem)
  - cf. fuzzy logic
- Probabilistic models can handle both discrete and continuous variables at the same time
- Various approaches for handling missing data (both in model building and in reasoning)

Nice theory, but…

- "isn't probabilistic reasoning counter-intuitive, something totally different from human reasoning?"
  - Cause for confusion: the old frequentist interpretation. But probabilities do NOT have to be thought of as frequencies, but as measures of belief
  - The so called paradoxes are often misleading
    - ✓A: P($1.000.000)=1.0
    - ✓B: P($1.000.000)=0.25, P($4.000.000)=0.25, P($0)=0.5
  - Even if that were true, maybe that would be a good thing!

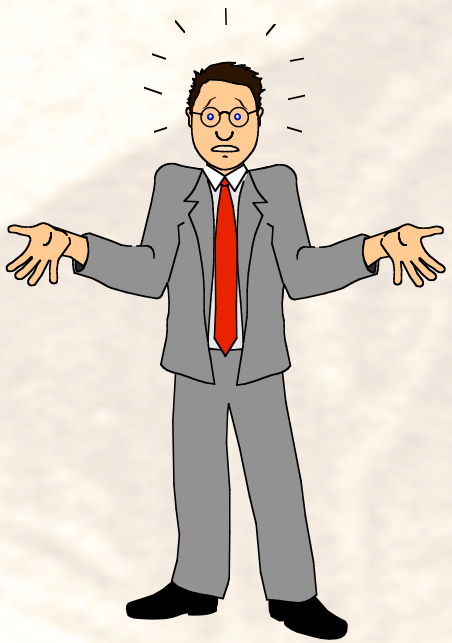
Nice theory, but...

- "Where do all the numbers come from?"
  - Bayesian networks: small number of parameters
  - the numbers do not have to be accurate
  - probability theory offers a framework for constructing models from sample data, from domain knowledge, or from their combination



We can learn from Bayesians :-)

"I rest my case"



- Bayesian approaches never overfit (in principle)
- Bayesian approaches infer only from observed data (not possible data)
- Bayesian inference is always relative to a model family
- Does all this semi-philosophical debate really matter in practice?
  - **YES**!!
  - (see e.g. "The great health hoax" by Robert Matthews. The Sunday Telegraph, September 13, 1998.)