Introduction to Probability

Terascale Statistics School

Roger Barlow

6th March 2017

What is Probability?

Q1: What is meant by the *probability* $P(A)$ of an event A? [1]

Possible Answers

(0) It is a number between 0 and 1 obeying certain mathematical rules.

(1) It is a property of A that makes it happen, perhaps given by symmetry.

(2) It is the limit as $N \to \infty$ of N_A/N_{total}

(3) It is my degree of belief in A, as determined by the odds I would accept as a bet

(0) Mathematical

The Kolmogorov Axioms

A. N. Kolmogorov

1) $P(A) \geq 0$

2) $P(all) = 1$

3) $P(A\&B) = P(A) + P(B)$ if A and B are disjoint

From these one can deduce $P(A) \leq 1$, $P(\overline{A}) = 1 - P(A)$, etc

But there is no definition of what $P(A)$ actually means

(1) Classical

Intrinsic property, e.g. tossing coin coming heads has $P = \frac{1}{2}$. Throwing 6 with dice $\frac{1}{6}$. Drawing a red queen from a pack of 52 cards $\frac{1}{26}$

Developed during 18th century by Laplace, Pascal etc. for the gambling industry

"Principle of Insufficient Reason" or "Principle of indifference": all outcomes equally likely

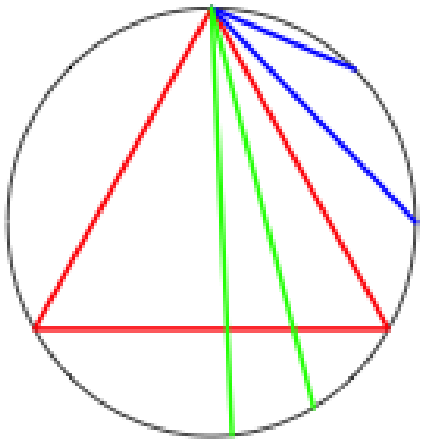
Problem: doesn't work for continuous variables

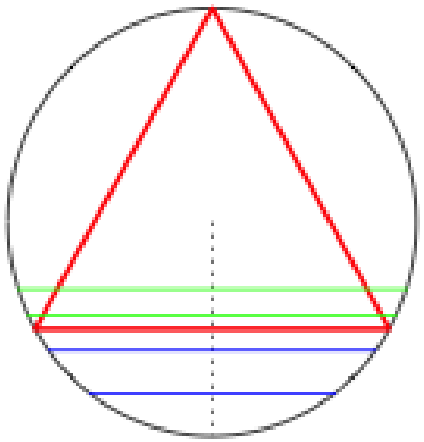
If $P(x)$ is constant, $P(f(x))$ is not, for any non-trivial transformation $f(x)$

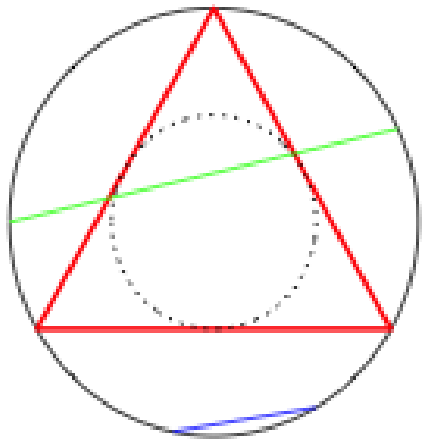
Example θ and $\cos\theta$ as 'random angles' in 2D and 3D

Bertand's Paradox

Bertand's Paradox: Inscribe an equilateral triangle in a circle. Draw a chord at random. What is the probability that the chord is longer than the side of the triangle?



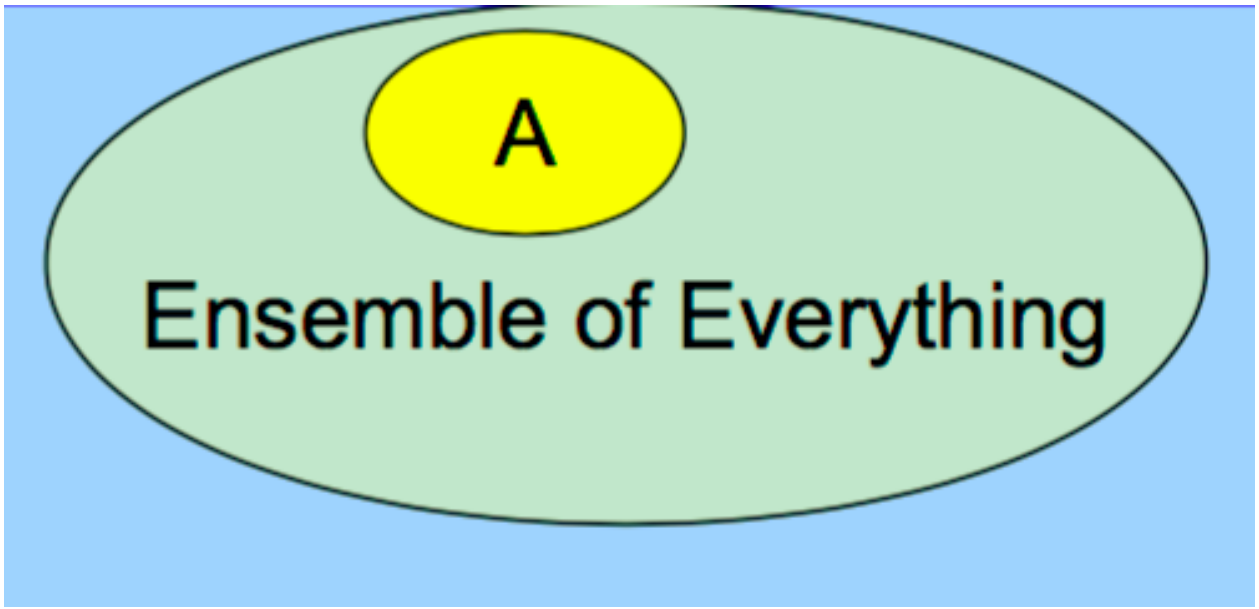




Answer is 'obviously' $\frac{1}{3}$ or $\frac{1}{2}$ or $\frac{1}{4}$

Frequentist

In an ensemble of $N \to \infty$ events, the fraction of events for which A occurs.



Ensemble can be distributed in space (toss many coins, half come up heads) or time (toss a coin many times, it comes up heads in half of them).

Empirical: no reference to internal dynamic. (Came out of Vienna Circle. von Mises and others)

P(A) depends on A and the ensemble

Example: Life insurance companies determine that 0.4% of their 40 year old male clients will die during the year.

This is a hard number: it determines their premiums and profitability

Do you tell your friend, on his 40th birthday, "Many Happy Returns - with 99.6% probability"?

No: that is his figure as a member of the ensemble 40-year old insured men. Also member of 40-year old men, 40-year old non-smokers, and/or hang-glider pilots, and/or .....

There may be several ensembles and $P(A)$ will be different $P(A)$ is a joint property of A and the ensemble

No Ensembles

Weather Forecasting

"It will probably rain tomorrow"

.

Inadmissable ('unscientific') statement, in frequentist definition There is only one tomorrow. It will either rain or not. P is 0 or 1.

Thus matters because 'The theory is probably correct' has the same problem

Workaround: suppose forecast[1] predicts rain. Track record shows it is right in 80 cases out of 100. So you can say

"The statement 'It will rain tomorrow' is probably true"

Or you say, with 80% confidence, "It will rain tomorrow"

Ensemble dependent. Maybe forecast is right 80% of the time overall , but only 70% when it predicts rain, but 90% when for rain in Hamburg. All these numbers are good.

[1] in paper or TV, or whatever

Subjective (Bayesian)

$P(A)$ is my degree of belief in A. 1 means certainty and 0 means not at all. Intermediate values calibratable against classical probability establishing which event I would rather bet on.

E.g. I will bet on rain tomorrow rather than on drawing a white ball from an urn containing 7 white and 3 black balls. If offered 9+1 I would bet on the urn. If offered 8+2 I would take either. So $P(\mathit{rain}) = 80\%$

Plus point: no restrictions on A

Minus point: your $P(A)$ and my $P(A)$ will in general be different

Bayes' Theorem

Actually valid for all probability definitions

Theorem

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

Proof: $P(A|B)P(B) = P(A\&B) = P(B|A)P(A)$

trivial example: a card drawn from the pack is black. What is the probability that it is the ace of spades?
$P(\spadesuit A|black) = \frac{P(black|\spadesuit A)}{P(black)} P(\spadesuit A) = \frac{1}{1/2}\frac{1}{52} = \frac{1}{26}$

Contrast $P(\spadesuit A|red) = \frac{0}{1/2}\frac{1}{52} = 0$

It is sometimes helpful to expand the denominator
$P(B) = P(B|A)P(A) + P(B|\overline{A})(1 - P(A))$

From prior to posterior

Bayesian Bayes: $P(Theory|data) = \frac{P(data|Theory)}{P(data)} P(Theory)$

I have some prior degree of belief $P(Theory)$ in $Theory$. An experiment reports $data$. My degree of belief is modified by a factor $\frac{P(data|Theory)}{P(data)}$ to give a posterior value $P(Theory|data)$.

if the theory predicts the data, the data supports the theory. This is moderated by the factor that the data could happen anyway.

Remember
$P(data) = P(data|Theory)P(Theory) + P(data|\overline{Theory})(1 - P(Theory))$

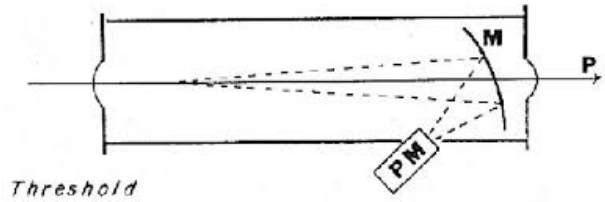
Special cases:
$P(data|Theory) = 0$: $P(Theory|data) = 0$. Forbidden event kills theory
$P(data|\overline{Theory}) = P(data)$: experiment irrelevant and belief unmodified
$P(data|\overline{Theory}) = 0$: $P(Theory|data) = 1$

Applying Bayes (1)

Cherenkov Detector

Momentum-selected beam containing π^+ and K^+ mesons.

Gas Cherenkov with threshold velocity v_C such that $\frac{pc}{\sqrt{p^2+M_K^2}} < v_C < \frac{pc}{\sqrt{p^2+M_\pi^2}}$



Ideally: Cherenkov identifies π and K by presence/absence of signal

Reality: suppose π has 90% chance of signal, K has 5% chance

What does a signal (or absence of signal) actually mean? Suppose beam is 80% π, 20% K (need to know this!)

$$P(\pi|signal) = \frac{0.9}{0.9\times0.8+0.05\times0.2} \times 0.8 = 98.6\%$$

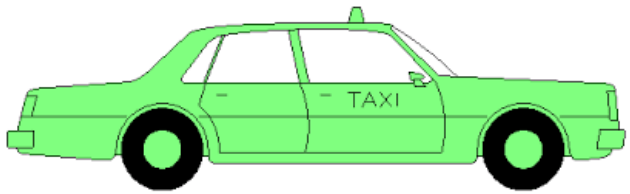
$$P(K|nosignal) = \frac{0.95}{0.1\times0.8+0.95\times0.2} \times 0.2 = 70\%$$

Applying Bayes

The taxi-cab 'paradox'

90% of cabs in a city are blue and 10% are green. A witness says a cab involved in a hit-and-run was green. Witnesses' recollections are 80% reliable. What is the most likely colour?

$$P(g) = \frac{0.8}{0.9\times0.2+0.1\times0.8} \times 0.1 = 31\%$$
$$P(b) = \frac{0.2}{0.9\times0.2+0.1\times0.8} \times 0.9 = 69\%$$



Bayes theorem combines the prior probability with the evidence, and shows that (with these numbers) the cab is still more likely to be blue

This is a parable: not really about taxi-cabs. Or blue and green

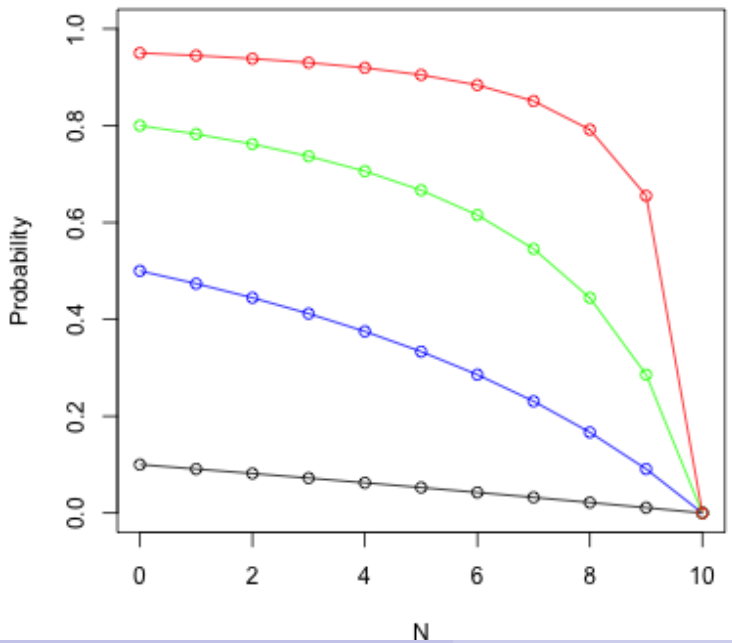
Suppose a second witness agrees with the first. Then

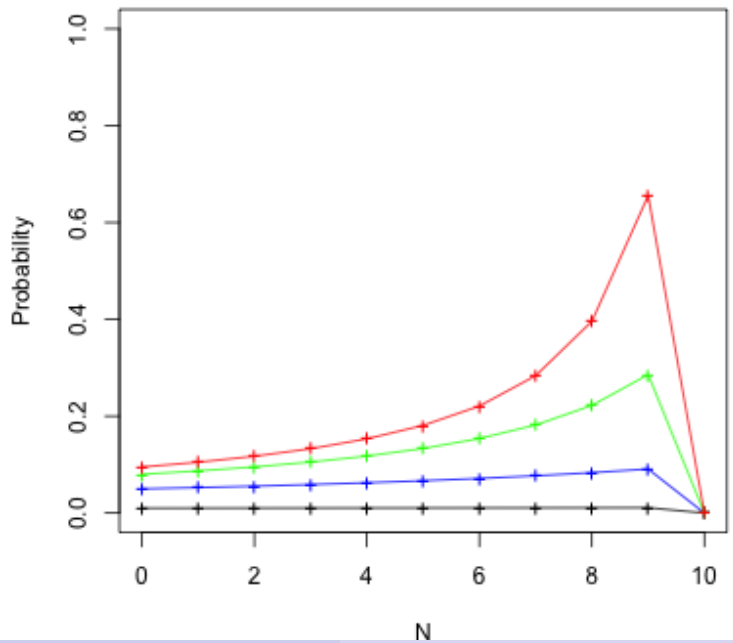
$$P'(g) = \frac{0.8}{0.69\times0.2+0.31\times0.8} \times 0.31 = 64\%$$

Everyday Bayes

The WMD effect

Suppose an object may exist - prior probability P_0 - in one of N equally-likely locations.
Search first location unsuccessfully: $P_1 = \frac{(N-1)/N}{P_0(N-1)/N+(1-P_0)} P_0$
P_n falls as n increases. For small/medium P_0 this is approx linear. For large P_0 it stays high until the final search.





Bayesian Conspiracy

Fun Fact!	**Q. What is the Bayesian Conspiracy?** **A.** The Bayesian Conspiracy is a multinational, interdisciplinary, and shadowy group of scientists that controls publication, grants, tenure, and the illicit traffic in grad students. The best way to be accepted into the Bayesian Conspiracy is to join the Campus Crusade for Bayes in high school or college, and gradually work your way up to the inner circles. It is rumored that at the upper levels of the Bayesian Conspiracy exist nine silent figures known only as the Bayes Council.

http://yudkowsky.net/bayes/bayes.html

Prior Distributions

'Theory': simple $\rightarrow$ composite

There is an X particle$\rightarrow$*there is an X particle with mass* M_X

Prior is now a function $P(M_X)$ with $\int P(M_X)\, dM_X = P$, total $P(\mathit{Theory})$

Example : M_X could be anywhere, with equal probability, up to 10 GeV

$$\begin{aligned} P(M_X) &= 0.1 \quad M_X < 10 GeV \\ &= 0 \quad\;\; M_X \geq 10 GeV \end{aligned}$$

Bayes' Theorem says $P(M_X|data) = \frac{P(data|M_X)}{P(data)} \times P(M_X)$

Note: $P(data)$ is a constant: usually determine by normalising P'

Improper priors

What if there is no upper limit?
$P(M_X) = p_0, \int_0^\infty p_0 dM_X = P(Theory) \leq 1$; p_0 vanishingly small but > 0

So statements like 'The probability of M_X lying between 2 and 4 GeV is twice the probability of it lying between 5 and 6 GeV' are still meaningful.

Bayes' Theorem says $P(M_X|data) = \frac{P(data|M_X)}{P(data)} \times p_0$

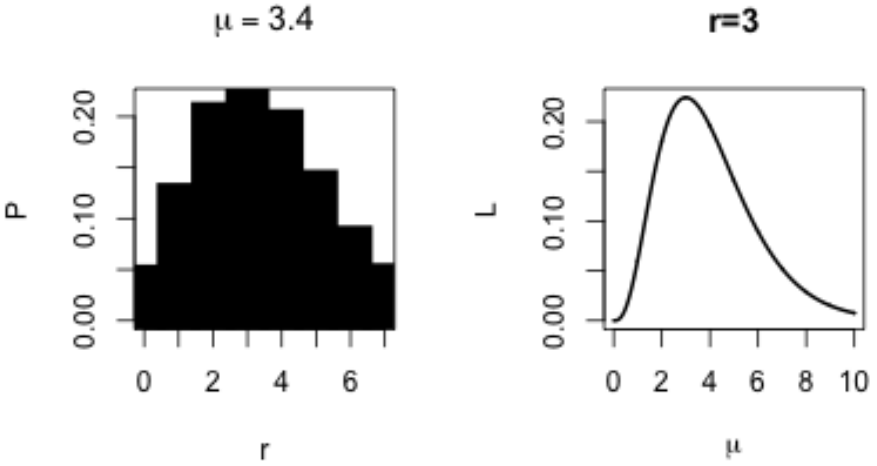
$P(data) = \int p_0 \frac{1}{\sigma\sqrt{2\pi}} e^{-(m-M_X)^2/2/\sigma^2} dM_x$ or something

so just cancel the p_0 factor, write $P(M_X|data)' \propto P(data|M_X)$ and normalise to get the posterior: you can do this only because $p_0 > 0$

Likelihood - a very useful concept

Suppose data x and theory value μ

$$P(data|theory) \equiv P(x;\mu) \equiv L(\mu;x)$$



Poisson formula $P(x;\mu) = e^{-\mu}\frac{\mu^r}{r!}$

Can ask: if $\mu = 3.4$, what are the probabilities of $x = 0, 1, 2, 3, 4...$?

Can ask: if $x = 3$, what is the likelihood of μ?

Poisson example more helpful than Gaussian as $\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$ is symmetric in x and μ

Frequentists are not allowed to integrate the likelihood, however tempting

Likelihood is not probability: this is the 'likelihood' that μ will produce x, it does not tell you about the 'likelihood' of particular μ value

To find that you need to either (i) include the prior, etc using Bayes theorem or (ii) use Frequentist techniques, discussed later.

The 'Likelihood principle' says that everything you need to know is in the likelihood function. Sounds plausible but contentious.

Prosecutor's fallacy

The criminal leaves a bloodstain at the crime sign. Police arrest a suspect whose blood type matches - say chance of a match occuring at random is 100 to 1. Prosecutor tells jury the probability of their guilt is 100 to 1.

Frequentist (etc) just says: $P(A|B) \neq P(B|A)$. Likelihood is not probability

Bayesian can go further:
$P(guilt|match) = \frac{P(match|guilt)}{P(match)} P(guilt) = \frac{1}{P(guilt)+0.01(1-P(guilt))} P(guilt)$

$0.5 \rightarrow 0.99$ Suspicion becomes certainty
$0.1 \rightarrow 0.92$ Weak suspicion becomes strong suspicion
$0.001 \rightarrow 0.09$ Rank outsider

Robustness under choice of prior

Posterior (number or distribution) depends on experiment (likelihood) and on prior

I am interested in your experiment but not in your prior - I have my own

'What is the correct prior?' is the wrong question.

'Uniform prior' is not a let-out (even if you call it 'Principle of insufficient reason' and ingenuously plead ignorance) as a prior uniform in μ is not uniform in $cos\mu$, $\sqrt{\mu}$, etc.

If there is a lot of data, posterior becomes independent of prior. Nice. But we are usually not so fortunate.

Good practice is to use several priors and check how your results change. If they are stable, all is well. If they change, you are on sticky ground and should say so.

Statisticians know this and do it. Physicists sometimes need encouragement to do the right thing.

What do you really believe?

A prior uniform in μ is convenient but may not express your true belief.

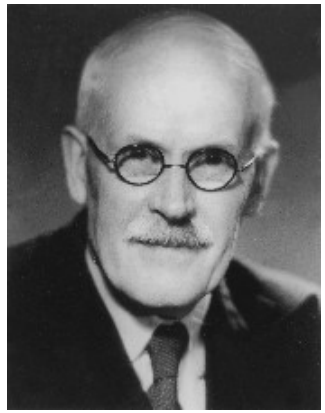
Suppose you are measuring a branching ratio B. You use a prior uniform between 0 and 1. (or between 0 and some previously established upper limit)

What would you rathe bet on: the true value lying between 0 and 0.001, or between 0.100 and 0.101?

I suspect that uniformity in $logB$ is a better description of your prior belief.

Jeffreys' Priors

Objective priors



Consider likelihood $L(\mu; x)$ plots again (often plotted on log scale, then approx parabolic)
Width of peak depends on many things. Narrow peak is clearly more 'informative' than broad peak. Define $I(\mu) = -\left\langle \frac{d^2 \ln L}{d\mu^2} \right\rangle$ as 'information' Angle brackets denote expectation value - integrating over all possible measurements at this value of μ
If $I(\mu)$ depends on μ then some values of μ will give better (more informative) results than others. Unfair?
Jeffreys proposes: transform to variable $\mu'(\mu)$ such that $I(\mu')$=constant, and then use a prior which is uniform in μ'
Simpler equivalent: Use μ but with prior $P(\mu) = \sqrt{I(\mu)}$
Common examples: for a location parameter $P(\mu) = const.$
For a scale parameter $P(\mu) \propto \frac{1}{\mu}$,
For a Poisson mean $P(\mu) \propto \frac{1}{\sqrt{\mu}}$
Not universally adopted as (i) Prior depends on experiment (ii) Not easily extendable to more than one parameter (iii) stubborn statisticians

What is probability?

(0) Mathematical: Obeys axioms.

(1)Classical: a hard('real') number

(2) Frequentist: depends on the Ensemble. Which leads to ambiguities and limits on what one can assign probabilities to

(3) Bayesian. Subjective - and subjectivism is 'unscientific'

Do we need the classical definition?

Is probability 'real'

In statistics, it's dead. Scorned by frequentists and Bayesians alike. (Bruno de Finetti: 'Probability does not exist'). A few philosophers (Popper) support it.

But Quantum Mechanics (only) predicts probabilities. Indeterminism is bad enough - if probability depends (at best) on some ensemble, at worst on some person's prior belief, what place for a 'real' world?

Is the lifetime of a muon an intrinsic property, like its charge and rest-mass, or not?

Seems plausible that a muon 'knows; about its 2.2 μsec lifetime in the same way that a coin 'knows' about it's 50% chance of landing on either side. And about the (continuous) energy spectrum of its decay electron. Just because probability distributions of continuous variables cannot be given by symmetry doesn't mean they can't be given some other way.

Do we need the frequentist definition?

Prediction and Inference are very different (though we use 'probability' for both)

The probability that a muon will give a track in the muon detector is (perhaps) a classical real property of the muon and the detector.

The probability that a track in the muon detector came from a muon depends on the ensemble (Loose muons, tight muons, high P_T muons...) The track itself is either a muon or it isn't. This is a frequentist probability (which can be adapted into Bayesian form by a sensible prior.)

So yes, in inference we have to use Frequentist or Bayesian methods - sometimes both

Conclusions and outlook

There are 4 definitions of probability.

All are correct. They are not in competition (well, a bit).

You need to know about all four, to know their strengths and weaknesses, and to be able to use the right concept on the right occasion, and to know which you are using at any time.