Basics I: Probability

Roger Barlow
The University of Huddersfield

Terascale Statistics School, DESY, Hamburg

3^{rd} July 2023

What is probability?

We all think we know, but...

From any Statistics101 Exam paper

Q1 What is meant by the probability $P(A)$ of some event A?

[1]

Please write down (or type, or whatever) your answer

4 possible answers

1. $P(A)$ is a real number between 0 and 1, obeying certain mathematical rules
2. $P(A)$ is some property of A: the larger it is, the more A happens with limits at 1 (never) and 1 (always)
3. $P(A)$ is the limit N_A/N_{Total}, as $N_{Total} \to \infty$
4. $P(A)$ expresses my belief in A, and determines the odds I will accept on a bet that A is true.

4 possible answers

1. $P(A)$ is a real number between 0 and 1, obeying certain mathematical rules
2. $P(A)$ is some property of A: the larger it is, the more A happens with limits at 1 (never) and 1 (always)
3. $P(A)$ is the limit N_A/N_{Total}, as $N_{Total} \to \infty$
4. $P(A)$ expresses my belief in A, and determines the odds I will accept on a bet that A is true.

All are possible correct answers, though only (3) and (4) are relevant

Mathematical

Kolmogorov Axioms:

A. N. Kolmogorov

For all $A \subset S$
$P_A \geq 0$
$P_S = 1$
$P_{(A \cup B)} = P_A + P_B$ if $A \cap B = \phi$ and $A, B \subset S$

From these simple axioms a complete and complicated structure can be erected. E.g. show $P_A \leq 1$....

But!!!

This says *nothing* about what P_A actually means.

Kolmogorov had frequentist probability in mind, but these axioms apply to any definition.

Classical

or Real probability

Evolved during the 18th-19th century

Developed (Pascal, Laplace and others) to serve the gambling industry.

Two sides to a coin - probability $\frac{1}{2}$ for each face

Likewise 52 cards in a pack, 6 sides to a dice...

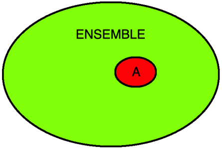
Answers questions like 'What is the probability of rolling more than 10 with 2 dice?'

Problem: can't be applied when symmetry breaks, or to continuous variables. Different answers using θ or $sin\theta$. (Bertrand's paradoxes.)

Frequentist

The usual definition taught in schools and undergrad classes

$P_A = \lim_{N\to\infty} \frac{N_A}{N}$
N is the total number of events in the ensemble (or collective)



The probability of a coin landing heads up is $\frac{1}{2}$ because if you toss a coin 1000 times, one side will come down ~ 500 times.

The lifetime of a muon is $2.2\mu s$ because if you take 1000 muons and wait $2.2\mu s$, then ~ 368 will remain.

Important

P_A is not just a property of A, but a joint property of A and the ensemble.

Problems (?) for Frequentist Probability

More than one ensemble

German life insurance companies pay out on 0.4% of 40 year old male clients.
Your friend Hans is 40 today. What is the probability that he will survive to see his 41st birthday?
99.6% is an answer (if he's insured)
But he is also a non-smoker and non-drinker - so maybe 99.8%?
He drives a Harley-Davidson - maybe 99.0%?
All these numbers are acceptable

What is the probability that a K^+ will be recognised by your PID?
Simulating lots of K^+ mesons you can count to get $P = N_{acc}/N_{tot}$
These can be divided by kaon energy, kaon angle, event complexity... and will give different probabilities ... All correct.

There may be no Ensemble

What is the probability that it will rain tomorrow?

There is only one tomorrow. It will either rain or not. P_{rain} is either 0 or 1 and we won't know which until tomorrow gets here

---

What is the probability that there is a supersymmetric particle with mass below 2 TeV?

There either is or isn't. It is either 0 or 1

From Probability to Confidence

Tomorrow it will either rain, or it will not rain.
P_{Rain} is either 0 or 1.
There is only one tomorrow

The weather forecast says it will rain
I have checked the track-record of this forecast, and they are right 90% of the time
The statement 'It will rain tomorrow' has a 90% probability of being true
I say that it will rain tomorrow, with 90% confidence

Definition

A is true at X% confidence if it is a member of an ensemble of statements of which at least X% are true

What is a measurement? What does $x = 12.3 \pm 0.4$ mean?

$x \in [11.9, 12.7]$ with 68% probabiity. NO
$x \in [11.9, 12.7]$ with 68% confidence. YES

The small print

Note that ‘at least’

1. Higher confidence statements embrace lower. If something is true at 95% confidence it is true at 90% confidence
2. It lets us handle cases with discrete data where an exact match may not be possible
3. It lets us handle composite hypotheses,

The actual fraction of cases is called the coverage
It generally depends on some parameter of the composite hypothesis - it’s a function not a simple number
Overcoverage is allowed (but inefficient)
Undercoverage is not

Bayes' theorem

Bayes' Theorem applies (and is useful) in any probability model
Conditional Probability: $P(A|B)$: probability for A, given that B is true.
Example: $P(\clubsuit A) = \frac{1}{52}$ and $P(\clubsuit A|Black) = \frac{1}{26}$

Theorem

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

Proof.

The probabiilty that A and B are both true can be written in two ways

$$P(A|B) \times P(B) = P(A\&B) = P(B|A) \times P(A)$$

Throw away middle term and divide by $P(B)$

□

Bayes' theorem

Examples

Example

$$P(\clubsuit A|Black) = \frac{P(Black|\clubsuit A)}{P(Black)} P(\clubsuit A) = \frac{1}{\frac{1}{2}} \times \frac{1}{52} = \frac{1}{26}$$

Example

Example: In a beam which is 90% π, 10% K, kaons have 95% probability of giving no Cherenkov signal; pions have 5% probability of giving none. What is the probability that a particle that gave no signal is a K?

$$P(K|no\ signal) = \frac{P(no\ signal|K)}{P(no\ signal)} \times P(K) = \frac{0.95}{0.95 \times 0.1 + 0.05 \times 0.9} \times 0.1 = 0.68$$

This uses the (often handy) breakdown:

$$P(B) = P(B|A) \times P(A) + P(B|\overline{A}) \times \overline{P(A)}$$

Bayesian Probability

Probability expresses your belief in A.
1 represents certainty, 0 represents total disbelief

Intermediate values can be calibrated by asking whether you would prefer to bet on A, or on a white ball being drawn from an urn containing a mix of white and black balls.

This avoids the limitations of frequentist probability - coins, dice, kaons, rain tomorrow, existence of SUSY can all have probabilities.

Bayesian Probability and Bayes Theorem

Re-write Bayes' theorem as

$$P(Theory|Data) = \frac{P(Data|Theory)}{P(Data)} \times P(Theory)$$

$$Posterior \propto Likelihood \times Prior$$

Works sensibly

Data predicted by theory boosts belief - moderated by probability it could happen anyway

Can be chained.

Posterior from first experiment can be prior for second experiment. And so on. (Order doesn't matter)

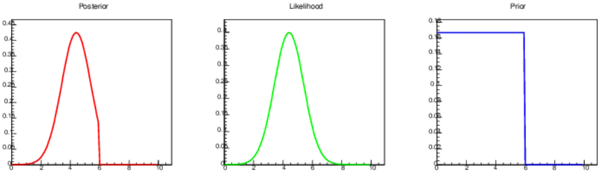
From Prior Probability to Prior Distribution

Suppose theory contains parameter a: (mass, coupling, decay rate...)

Prior probability distribution $P_0(a)$

$\int_{a_1}^{a_2} P_0(a)\, da$ is your prior belief that a lies between a_1 and a_2

$\int_{-\infty}^{\infty} P_0(a)\, da = 1$ (or: your prior belief that the theory is correct)
Generalise the number $P(data|theory)$ to the Likelihood function $L(x|a)$
Bayes' Theorem given data x the posterior is : $P_1(a) \propto L(x|a) P_0(a)$



If range of a infinite, $P_0(a)$ may be vanishingly small ('improper prior'). Not a problem. Just normalise $P_1(a)$

Your $P_0(a)$ and my $P_0(a)$ may be different. How can we compare results?

What is the right prior?

Is the wrong question.

'Principle of ignorance' - take $P(a)$ constant (uniform distribution). But then not constant in a^2 or $\sqrt{a}$ or $\ln a$, which are equally valid parameters.

Jefffreys' Objective Priors

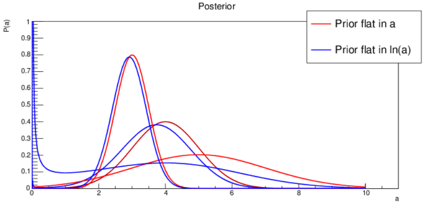
Choose a flat prior in a transformed variable a' for which the Fisher information, $-\left\langle \frac{\partial^2 L(x;a)}{\partial a^2} \right\rangle$ is flat. Not universally adopted for various reasons.

With lots of data, $P_1(a)$ decouples from $P_0(a)$. But not with little data..

Right thing to do: try several forms of prior and examine spread of results ('robustness under choice of prior')

Just an example

Measure $a = 4.0 \pm 1.0$. Likelihood is Gaussian (coming up!)



Taking a prior uniform in a gives a posterior with a mean of 4.0 and a standard deviation of 1.0 (red curve)

Taking a prior uniform in $\ln a$ (so $\propto 1/a$ for a) shifts the posterior significantly.
Also shown are 5 ± 2 (big difference) and 3 ± 0.5 (very little)

Summary

There are 4 (at least) definitions of ‘probabiity’.

The Mathematical and the Realist have severe limits. You need to use Frequentist and/or Bayesian

‘Frequentist versus Bayesian’ is the wrong language. These are not football teams!

Both have limitationss.

Frequentist statistics can lead you into some quite convoluted statements

Bayesian analyses must check for robustness under choice of prior

Be ready to use them both - but always know which you are working with