Notes: Practical Statistics for Physics & Astronomy

R.B. Metcalf

Alma Mater Studiorum - Universitá di Bologna

February 19, 2018

Contents

1 What is Probability?

1.1 Frequentist interpretation of probability

Imagine there is some event, instance or outcome of an experiment or observation called A. The probability of A is the fraction of times A occurs when the experiment or observation repeated in the same way or circumstances an *infinite* number of times.

$$P(A) = \lim_{N\to\infty} \frac{\text{number of trials where A is true}}{N(\text{total number of trials})} \tag{1.1}$$

This is the traditional definition of probability as formally stated by Laplace in 1774 and almost universally used for centuries despite no one ever having done anything *exactly* the same twice let along an *infinite* number of times.

Applying this definition to any physical phenomenon requires a partitioning of the world into things that are known and fixed on each repatriation of the observation and those things that are not known and change every repetition. If nature is deterministic and an experiment could be set up *exactly* the same way in all respects than the outcome would always be the same and probability would not apply. Of course even in classical physics it is not possible to know the state of every atom and photon that might possibly influence your measurement apparatus (or brain). It is these things that change when repeating the observation.

This partitioning between known and unknown factors seems reasonable when we talk about the positions and momenta of particles in a gas or flipping a coin, but in many other common situations where probability is used it seems less well defined. Say someone tells you that there is a 30% probability that candidate A will win an election tomorrow. Of course an identical election will never be run again and was never run in the past. There are many factors, known and unknown, that could affect an election. This statement was probably based on polling data. By the above definition of probability, this means that if the election were held an infinite number of times in which the polling data were exactly the same the candidate would win a 30% of them. This seems like a completely unverifiable claim. If scientific knowledge must be reproducible to be considered true then it would seem that any such argument should be considered unscientific. And yet probability through statistics is at the foundation of all quantitative measurements.

Lets be a bit more practical. Lets say we don't need an infinite number of trials, but just a very *large number* of them. Lets say we flip a coin a *very large number* of times. If we did it say one billion times we would not expect that *exactly* 500 million times it would be heads. We would expect that roughly half, but not exactly half of the times it would be heads even if the probability of getting heads in each flip is 1/2. We might try to quantify how close the number of heads should be to 500 million, but in doing so we would need to use a probabilistic argument that would use the very concept we are trying to define.

Many statisticians and philosophers have found this definition of probability problematic. Despite this it is the definition usually used by scientists when they are forced to addressing this subject.

1.2 classical interpretation of probability

The classical interpretation of probability relies on identifying events that are equally likely or probable. This is often the argumentation used in classical statistical mechanics where each micro-state of the system is taken to be equally probable. If one then says that the probability of being in either of two mutually exclusive states is the sum of their probabilities and that the some of the probabilities of being in all possible states is one then you can find a numerical value for the

probability of each state. A macro-state (one with temperature equal to some value or total energy equal to some value) corresponds to many micro-states so by adding up their probabilities you can find the probability of macro states which will not necessarily be equal.

The biggest criticism of this interpretation is that it doesn't really say what probability is, it just tells you how to calculate it in a restricted class of problems. What does it mean that two states are equally probable? What does the probability of a macro-state mean? Another problem is that not all events that we commonly apply probability to can reduced in this way to a collections of equally probable mutually exclusive events.

1.3 Subjective or Bayesian interpretation of probability

Thomas Bayes (1701 - 1761) (and Jacob Bernoulli 1655-1705) had a different conception of what probability is although the idea was not put on a firm theoretical foundation until the 1940's and 50's by G. Polya, R.T. Cox and E.T. Jaynes. It did not make its way into common use in science, in the form of Bayesian statistics, until relatively recently (80s and 90s).

In this school of thought, probability theory is an extension of formal logic to situations where the truth or falsehood of a proposition (e.g. "It will rain tomorrow." or "The mass of the Earth is between 5.972×10^{24} kg and 5.978×10^{24} kg.") cannot be deduced conclusively by deductive reasoning. A proposition has a probability function that depends on the evidence for and against its truth. When deductive reasoning can be applied conclusively this function is either zero (false) or one (true). In this way Boolean logic is a limiting case of probability theory. Surprisingly from just the following requirements (or *desiderata*) on the probability function of a proposition you can deduce the rules of probability and show that they are complete without ever mentioning randomness or repetition of experiments.

1. Degrees of plausibility are represented by real numbers.
2. The measure of plausibility must exhibit qualitative agreement with rationality. This means that as new information supporting the truth of a proposition is supplied, the number which represents the plausibility will increase continuously and monotonically. Also, to maintain rationality, the deductive limit must be obtained where appropriate.
3. Consistency
   (a) *Structured consistency* : If the conclusion can be reasoned out in more than one way, every possible way must lead to the same result.
   (b) *Propriety*: The theory must take account of all information that is relevant to the question.
   (c) *Jaynes consistency*: Equivalent states of knowledge must be represented by equivalent plausibility assignments. For example, if $A, B|C = B|C$, then the plausibility of $A, B|C$ must equal the plausibility of $B|C$

(taken from Gregory (2006)).

These foundational proofs are very interesting, but outside the scope of this course (for those that are interested see chapter 2 of Gregory (2006) or, more comprehensively, Jaynes (2003)). One thing that is of importance here is that this definition allows one to define the probability of something that would not usually be considered a *random variable* or a repeated even. This is central to the Bayesian method of parameter estimation and model selection that we will get to later.

A	B	A, B	$\overline{A, B}$	$\overline{A} \cup \overline{B}$	$A \cup B$	$\overline{A \cup B}$	$\overline{A}, \overline{B}$
F	T	F	T	T	T	F	F
F	F	F	T	T	F	T	T
T	T	T	F	F	T	F	F
T	F	F	T	T	T	F	F

Table 1: The truth table for binary logical expressions.

1.4 Quantum mechanical probability

The interpretation of probability as a measure certainty, or conversely ignorance,

1.5 the rules of probability

Suppose the A, B, ... are events that either occur or don't occur , that is they have values true or false (or 0 and 1 if you prefer). $P(A)$ is the probability of A occurring or being true. We can combine events in one of two ways. (A, B) means "A *and* B". It is true if both of them are true and false if both are false. $(A \cup B)$ means "A *or* B" it is true if either A or B is true. It is true if both are true. $\overline{A}$ means "not A". Note that $\overline{A \cup B} = \overline{A}, \overline{B}$ and $\overline{A, B} = \overline{A} \cup \overline{B}$ in the sense that there are no combinations of trues and falses for A and B that give different answers on either side of the equality. See table 1. In the language of Boolean algebra, they have the same truth table and are therefor equivalent statements. There probabilities must also be the same.

$P(A, B)$ is called the **joint probability** of events A and B. $P(A \cup B)$ is often called the **disjoint probability** of events A and B.

$P(A|B)$ is called a **conditional probability**. It means the probability of A *given* that B is true. You can imagine every probability being a conditional probability where it is "conditioned" on everything that you assume about the state of the Universe. Some of these things are assumed to be irrelavent and are left out . Some might be relevant but it is taken for granted so they are left out. The probability that a coin comes up heads does not depend on the time of day. It does depend on the assumption that it is a fair coin - no more likely to be heads than tails - although it might not always be stated. This is a simple example of a **statistical model** for the experiment, in this case flipping a coin.

The two fundamental rules of probability theory are

$$\begin{array}{cc} P(A, B) = P(A)P(B|A) & \textbf{product rule} \\ P(A) + P(\overline{A}) = 1 & \textbf{sum rule} \end{array} \tag{1.2}$$

These rules are actually derivable from some basic requirements or "desidariata" of how probabilities should behave, but for our purposes we can take them to be axioms. From these two rules and logic rules we can derive all the necessary properties of probability.

There are several particularly useful results that follow from these rules. From the logical requirement that (A, B) is the same as (B, A) and the product rule we get

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \textbf{Bayes' theorem} \tag{1.3}$$

Applying the sum rule to $(A \cup B)$ gives

$$\begin{aligned}
P(A \cup B) &= 1 - P(\overline{A \cup B}) && (1.4)\\
&= 1 - P(\overline{A}, \overline{B}) && (1.5)\\
&= 1 - P(\overline{A})P(\overline{B}|\overline{A}) && (1.6)\\
&= 1 - P(\overline{A})\left[1 - P(B|\overline{A})\right] && (1.7)\\
&= 1 - P(\overline{A}) - P(\overline{A})P(B|\overline{A}) && (1.8)\\
&= P(A) + P(\overline{A})P(B|\overline{A}) && (1.9)\\
&= P(A) + P(\overline{A}, B) && (1.10)\\
&= P(A) + P(B)P(\overline{A}|B) && (1.11)\\
&= P(A) + P(B)\left[1 - P(A|B)\right] && (1.12)\\
&= P(A) + P(B) - P(B)P(A|B) && (1.13)\\
P(A \cup B) &= P(A) + P(B) - P(B, A) \qquad \textbf{extended sum rule} && (1.14)
\end{aligned}$$

In words, the disjoint probability of two events is equal to the sum of their probabilities minus their joint probability.

If A and B are **independent** then the probability of A occurring does not depend on whether B has occurred so $P(A|B) = P(A)$ through the product rule this implies $P(B|A) = P(B)$ and

$$P(A, B) = P(A)P(B) \quad \text{independent events} \tag{1.15}$$

If two events are **mutually exclusive**, that is they cannot occur at the same time (the first flip of a coin cannot be both heads and tails) then $P(A, B) = 0$ and the extended sum rule becomes

$$P(A \cup B) = P(A) + P(B) \quad \text{mutually exclusive events} \tag{1.16}$$

Example: If you roll a die once the probability of getting a 6 *or* a 5 is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. If you roll a die twice the probability of getting a 6 *and then* a 5 is $\left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}$. The probability of getting a 6 *and* a 5 is twice this because, $\frac{1}{18}$, because there are two ways of doing this, a 6 first or a 5 first.

This second case can be calculated in an alternative way. In the first roll we must get a 5 or a 6. We have calculated that the probability of this is $\frac{1}{3}$. Once this is done in the second roll we most get whichever number we didn't get in the first roll, one number out of 6, probability $\frac{1}{6}$. The probability of these two independent events happening is then given by the product rule $\left(\frac{1}{3}\right)\left(\frac{1}{6}\right) = \frac{1}{18}$.

Now say we have a set of observations $\{A_I\}$ that are all mutually exclusive and together they include all possible outcome then

$$\begin{aligned}
1 &= P(A_1 \cup A_2 \cup A_3 \cup \ldots|B) + P(\overline{A_1 \cup A_2 \cup A_3 \cup \ldots}|B) && (1.17)\\
&= P(A_1|B) + P(A_2 \cup A_3 \cup \ldots|B) + 0 && (1.18)\\
&= P(A_1|B) + P(A_2|B) + P(A_3 \cup \ldots|B) && (1.19)\\
&= \sum_i P(A_i|B) && (1.20)
\end{aligned}$$

This is the origin of the normalization requirement on any probability distribution function (PDF). Note that I have put a B in as a condition on all the probabilities, but this would hold without them.

Another important result along these lines is

$$\sum_i P(B|A_i)P(A_i) = \sum_i P(B, A_i) = \sum_i P(A_i|B)P(B) = P(B)\sum_i P(A_i|B) = P(B) \tag{1.21}$$

with the same requirements on the set $\{A_i\}$. This is the origin of what we will later call marginalization.

A Matrix basics

$$(\boldsymbol{ABC}\dots)^T = \dots \boldsymbol{C}^T\boldsymbol{B}^T\boldsymbol{A}^T \tag{A.1}$$
$$(\boldsymbol{ABC}\dots)^{-1} = \dots \boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \tag{A.2}$$
$$(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T \tag{A.3}$$
$$(\boldsymbol{A}+\boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T \tag{A.4}$$

Some properties of the determinant

$$|\boldsymbol{A}| = \prod_i \lambda_i \tag{A.5}$$
$$|\boldsymbol{A}^{-1}| = 1/|\boldsymbol{A}| \tag{A.6}$$
$$|\boldsymbol{BA}| = |\boldsymbol{B}||\boldsymbol{A}| \tag{A.7}$$
$$|c\boldsymbol{A}| = c^n|\boldsymbol{A}| \tag{A.8}$$
$$|\boldsymbol{A}^T| = |\boldsymbol{A}| \tag{A.9}$$

Some properties of the trace

$$\mathrm{tr}(\boldsymbol{A}) = \sum_i A_{ii} \tag{A.10}$$
$$\mathrm{tr}(\boldsymbol{A}) = \sum_i \lambda_{ii} \tag{A.11}$$
$$\mathrm{tr}(\boldsymbol{A}^T) = \mathrm{tr}(\boldsymbol{A}) \tag{A.12}$$
$$\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA}) \tag{A.13}$$
$$\mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B}) \tag{A.14}$$

$\boldsymbol{A}$ is an **orthogonal matrix** if and only if

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I} \tag{A.15}$$

An orthogonal matrix has the following properties

$$\boldsymbol{A}^T = \boldsymbol{A}^{-1} \tag{A.16}$$
$$|\boldsymbol{A}| = \pm 1 \tag{A.17}$$

The $|\lambda_i| = 1$ for all eigenvalues and the magnitude of all eigenvactors are 1.

$\boldsymbol{C}$ is a **positive definite matrix** if

$$\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x} > 0 \quad \forall \boldsymbol{x}. \tag{A.18}$$

It has the following properties

- all eigenvalues at positive
- $\mathrm{tr}(\boldsymbol{C}) > 0$
- all diagonal elements are positive, $\boldsymbol{C}_{ii} > 0, \forall i$
- $\boldsymbol{C}$ is invertible

The covariance matrix is always positive definite.

"A and B"	A, B
"A or B"	$A \cup B$
continuous random variables	x,y,x_i,y_i
vector of random variables	$\boldsymbol{x}$ or $\vec{x}$
discrete random numbers	n,m
parameters	α , β
estimator of parameter α	θ_α or $\hat{\alpha}$
data	D or d_i
indexes data or for multiple random numbers	i, j
statistical and/or theoretical model	M
Gaussian or Normal pdf	$\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu},\boldsymbol{C}\right)$
$\boldsymbol{x}$ is normally distributed	$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\sigma})$
x is χ^2 distributed with n degrees of freedom	$x \sim \chi^2_n$
arithmetic mean of N samples	$\bar{x}_N$
likelihood of data given model	$\mathcal{L}(\boldsymbol{D}\|M_i)$ or $P(\boldsymbol{D}\|M_i)$
Bayesian evidence of data	$\mathcal{E}(\boldsymbol{D})$
Heaviside function, 1 when B is true, 0 otherwise	$\Theta(B)$

Table 2: notation

B notation

Notation may vary but in general I will follow the guide in table 2

C Some Useful Integrals and mathematical definitions

$$\int_{-\infty}^{\infty} dx\; e^{-\frac{x^2}{2}} = \sqrt{2\pi} \tag{C.1}$$

$$\begin{aligned}\int_{-\infty}^{\infty} dx\; e^{-(ax^2+bx+c)} &= e^{-c}\int_{-\infty}^{\infty} dx\; e^{-\left(\sqrt{a}x+\frac{b}{2\sqrt{a}}\right)^2+\frac{b^2}{4a}} = e^{-c+\frac{b^2}{4a}}\int_{-\infty}^{\infty}\frac{dy}{\sqrt{a}}\; e^{y^2} \\ &= \sqrt{\frac{\pi}{a}}e^{-c+\frac{b^2}{4a}}\end{aligned} \tag{C.2}$$

$$\int_0^{\infty} dx\; x^n e^{-\frac{1}{2}Ax^2} = 2^{\frac{n-1}{2}}A^{-\frac{n+1}{2}}\Gamma\left(\frac{n+1}{2}\right) \qquad n > -1 \tag{C.3}$$

The Gamma function

$$\begin{array}{lll}\int_0^{\infty} dx\; x^n e^{x^2} & = \frac{1}{2}\Gamma\left(\frac{n+1}{2}\right) & \\ \Gamma(n) & = (n-1)! & n = 1, 2, \ldots \\ \Gamma\left(\frac{1}{2}+n\right) & = \frac{(2n)!}{4^n n!}\sqrt{\pi} & n = 0, 1, 2, \ldots\end{array} \tag{C.4}$$

Stirling's approximation

$$\ln N! \simeq N\ln N - N \text{ for } N \gg 1 \tag{C.5}$$

or more accurately

$$N! \simeq \sqrt{2\pi N}\left(\frac{N}{e}\right)^N \text{ for } N \gg 1 \tag{C.6}$$

Index

References

Gregory P., 2006, Bayesian Logical Data Analysis for the Physical Sciences. Cambridge University Press

Jaynes E., 2003, Probability Theory - The Logic of Science