Lecture 1

Introduction to the concepts of probability

If we were not ignorant there would be no probability, there could only be certainty. But our ignorance cannot be absolute, for then there would be no longer any probability at all. Thus the problems of probability may be classed according to the greater or less depth of this ignorance.

H.Poincaré

What is a **probability**?
From latin *probabilis, probare*: that can be verified, that is reasonable to assume, that can be proven.

Example: Can we predict whether the next toss of a coin will land on a *head* or on a *tail*?

Two main intepretations:

- **Objective (or physical) probability**. We toss the coin 1000 times and count how many times it lands on a tail. In this view probability is defined as:

$$P(coin\ will\ land\ on\ a\ tail) = \frac{\#\ times\ we\ obtained\ a\ tail}{\#\ times\ we\ tossed\ the\ coin}$$

- **Subjective probability**. It reflects our degree of belief that a certain event might occur.

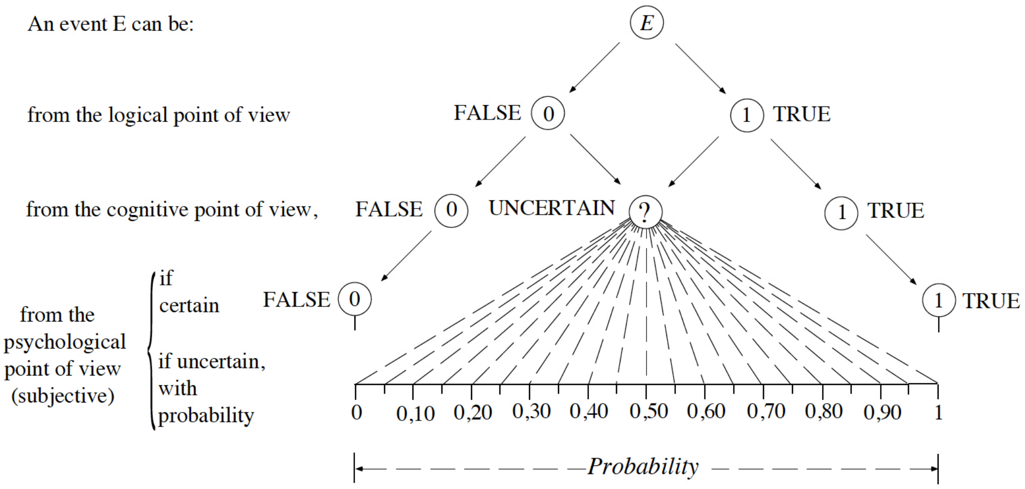
An event E can be:
E
from the logical point of view
FALSE 0
1 TRUE
from the cognitive point of view,
FALSE 0
UNCERTAIN ?
1 TRUE
from the psychological point of view (subjective)
if certain
FALSE 0
1 TRUE
if uncertain, with probability
0
0,10
0,20
0,30
0,40
0,50
0,60
0,70
0,80
0,90
1
Probability

Figure 1 Probability as a measure of the degree of belief.

A few useful definitions

- An **event** e_i is defined as any statement of which its truth can be verified, at least in principle. *Ex.: "There was a thunderstorm in Rome the day of the battle of Trafalgar."*

- **Sets** are collections of events. *Ex.: we roll a die and we want to know whether the outcome will be an even* $A = \{2, 4, 6\}$ *or an odd number* $B = \{1, 3, 5\}$.

- A **probability space** is the collection of all events that can occur in our problem $S = \{e_1, e_2, ..., e_N\}$. If we are about to toss a coin: $S = \{H, T\}$. This in an example of *discrete* probability space.

- Empirical definition of **probability**: Corresponding to an event e_i there is a definite number $\boldsymbol{P(e_i)}$, called the *probability of the event*, whose intrinsic property is that, as the number of trials (experiments) increases, the frequency of event e_i is approximated by $P(e_i)$.

Probability axioms

1. The probability sample space S must be well defined, as well as the set corresponding to the impossible event: $\emptyset$

2. Non-negativity: $P(e_i) \geq 0$ for $i = 1, ..., N$.

3. Normalization: $P(S) = 1$.

4. Take a collection of *nonoverlapping* sets $A_i (i = 1, ..., N)$. Then the probability of the union of all sets is:

$$P(\cup_i A_i) = P(A_1) + P(A_2) + ... + P(A_N) \equiv \sum_{i=1}^{N} P(A_i) \tag{1}$$

A few properties that follow:

- For any set $A \subseteq S$ we can define the complement $\bar{A}$ as the set of all events in S that are not contained in A. It follows that:

$$P(\bar{A}) = 1 - P(A) \tag{2}$$

  hence $P(\emptyset) = 0$.

- If two sets A and B are *overlapping*, i.e. they share one or more events, we can write the **joint probability** as $\boldsymbol{P(A \cap B)}$ (or, equivalently, as $\boldsymbol{P(A, B)}$), i.e. the probability that an event is contained both in A *and* in B. *Ex.: The probability that a card is a four and is red,* $P(four \ \cap \ red) = 2/52 = 1/26$. *(There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).*

- The joint probability plays a role when we want to compute the probability of the union of two overlapping sets, $P(A \cup B)$. In fact it is easy to see that we cannot just sum the probability of each set, but we need to subtract the joint probability $P(A \cap B)$:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3}$$

- For an arbitrary (but finite) number of sets:

$$P(\cup_i A_i) = \sum_i P(A_i) \; - \sum_{i<j} P(A_i \cap A_j) \; + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \\ - (-1)^N P(A_1 \cap A_2 \cap \ldots \cap A_N)$$

this is called *inclusion-exclusion principle*.

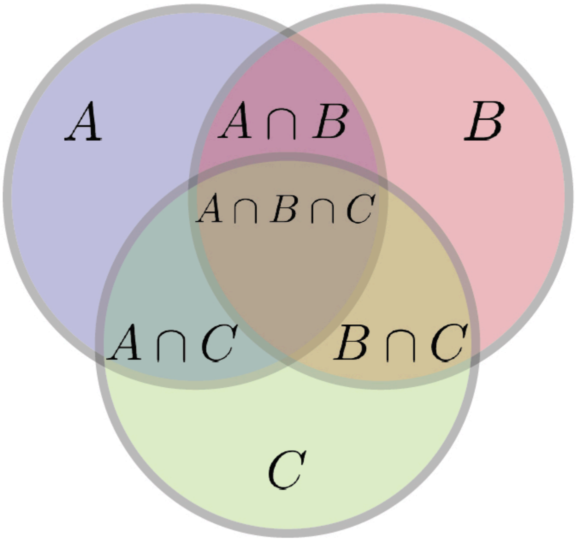

Figure 2 Illustration of the inclusion-exclusion principle for 3 sets.

The Bernoulli distribution

When we toss a coin we expect a head to appear with exactly the same probability as a tail. More generally, the Bernoulli distribution describes a binary process where one event (head, or “success”) occurs with probability p and the complementary event (tail, or “failure”) occurs with probability $(1-p)$.

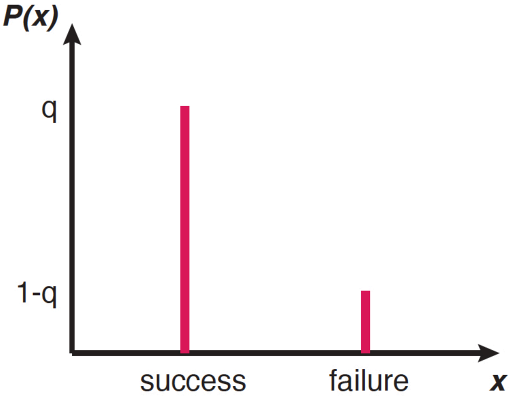

Figure 3 The Bernoulli distribution.

Combinatorics

The Bernoulli distribution describes the probability of obtaining a head or a tail after one toss. But what if we toss the coin three times, and we want to know the probability of getting two heads and one tail (set A)? Let us list all the possible outcomes and count those which contain two heads:

$$HHH, \boldsymbol{HHT}, \boldsymbol{HTH}, \boldsymbol{THH}, HTT, THT, TTH, TTT \tag{4}$$

The probability of getting exactly two heads and one tail is then:

$$P(A) = \frac{\#\ of\ events\ with\ 2\ heads\ and\ 1\ tail}{\#\ all\ events} = \frac{3}{8} \tag{5}$$

hence, we must find a systematic way to count events; luckily, combinatorics comes to our aid!

The binomial distribution - a simple model of spike generation

We consider the spiking pattern of a neuron in the interval $[0, T]$.We first divide T into N small intervals (bins) t_i of length $\Delta t = \frac{T}{N}$, so that there will be a maximum of one spike per bin:

$$P(>1\ spike\ in\ \Delta t) \backsim 0 \tag{6}$$

We know that for this neuron the probability of observing a single spike within a given bin is independent of the position of t_i, and it is equal to q, and we are interested in computing the probability of observing a certain number of spikes, k, distributed in N bins. Let us examine the problem by listing the potential outcomes in Fig. 4.

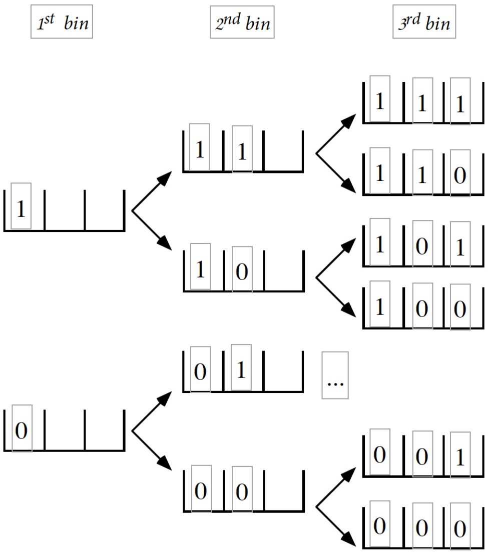

Figure 4

We notice two things: i) order is not important and ii) there is no replacement (since we set to k the number of spikes we are interested in). Hence:

$$\begin{aligned} P(k\ spikes\ in\ N\ bins) &= P(k\ bins\ with\ a\ spike) \times P((N-k)\ empty\ bins) \times \\ &\times (\#\ of\ ways\ in\ which\ k\ spikes\ can\ occupy\ N\ bins) = \\ &= q^k(1-q)^{N-k}\binom{N}{k} \end{aligned}$$

This probability distribution, that captures the odds of finding k independent discrete events (each occurring with probability q) in N "boxes", is called the **binomial distribution**.

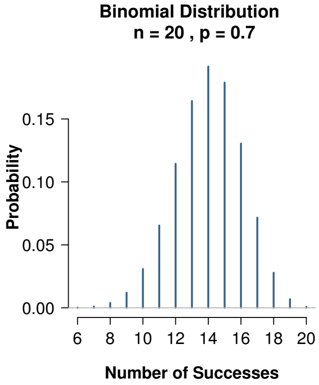

Figure 5 The binomial distribution.

Conditional probability and independence

Let us consider a (fictitious, of course!) neuron in the visual cortex (V1), that has been recorded in two conditions: either in the light, for 70% of trials, or in the dark, for the remaining 30% of trials. The probability that the neuron has emitted k spikes within a trial while the animal was in the dark can be calculated as:

$$\frac{\#\ trials\ with\ k\ spikes\ in\ the\ dark}{\#trials\ in\ the\ dark} \backsimeq P(k\ spikes \mid dark) \tag{7}$$

$$\equiv \frac{P(k\ spikes\ , dark)}{P(dark)} \tag{8}$$

The formula for total probability

We are interested in calculating the probability that our neuron spiked twice, regardless of whether the animal was in the dark or in the light. The concept of conditional probability offers us a convenient way to think about this problem: looking at Fig. 6 it is easy to convince ourselves that the probability of observing 2 spikes is equal to:

$$P(k=2) = P(k=2\ \cap\ dark) \cup P(k=2\ \cap\ light)$$

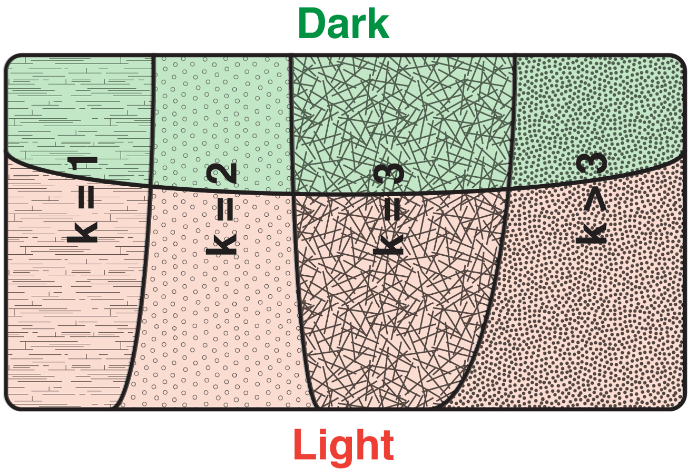

Figure 6 Venn diagram of the spike counts of an example neuron in V1.

we can thus invert 8 for each of the right hand terms, make use of the fact that the sets are nonoverlapping (see Eq. 1) and write:

$$P(k=2) = P(k=2 \mid dark)P(dark) + P(k=2 \mid light)P(light)$$

this expression is called **formula for total probability** and can be easily extended to the general case:

$$P(A) = \sum_{i=1}^{N} P(A|B_i)P(B_i)$$

the only requirement being that $\{B_1, B_2, ..., B_N\}$ is a *complete* set of *disjoint* (or mutually exclusive) events.

Independence

It is interesting to notice that, in our example, when the neuron was recorded in the light, it generated $k > 3$ spikes more often than when in the dark. So if we were to plan a new experiment where we would need the neuron to spike a lot, we would probably put the animal in the light. In other words, we gained some amount of information by conditioning the probability of spiking on whether it was dark or bright. This discussion naturally brings us to the concept of independence: when conditioning over some event does not make us gain any information about another event, then the two events are said to be independent. More formally, if A and B are independent:

$$P(A|B) = P(B).$$

An alternative formulation makes use of Eq. 8 to write:

$$P(A, B) = P(A)P(B)$$

Aside from the case $P(A|B) = P(B)$, there are two more cases to consider:

$$P(A|B) > P(A) \rightarrow \textit{then } A \textit{ and } B \textit{ are } \boldsymbol{positively\ correlated}$$
$$P(A|B) < P(A) \rightarrow \textit{then } A \textit{ and } B \textit{ are } \boldsymbol{negatively\ correlated}$$

If this seems unjustified, do not worry. We will talk more about correlation in the next lecture.

Bayes' rule

Let us go back to the relationship between joint and conditional probability in our example of Fig. 6:

$$P(k = 2, dark) = P(k = 2 \mid dark)P(dark)$$

We could as well reverse the conditioning:

$$P(k = 2, dark) = P(dark \mid k = 2)P(k = 2)$$

Hence, in general, we can write:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(B)$$

from which follows that:

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

This is known as **Bayes' rule**, and it is a way to convert one conditional probability $P(A|B)$ into the other $P(B|A)$, by reweighting it with the relative probability of the two variables (the ratio $P(B)/P(A)$). So, why is this interesting?
If we interpret B as the data we currently have and A as a certain hypothesis we have regarding how the data are generated, Bayes' rule assumes an insightful form:

$$P(hypothesis \mid data) = \frac{P(data \mid hypothesis)\ P(hypothesis)}{P(data)}$$

Bayes' rule offers a recipe to update our belief regarding a certain hypothesis (or *cause*) given the data we have. The probability that describes how likely our hypothesis is to be true, given the observed data, $P(hypothesis \mid data)$, is called *posterior*. $P(hypothesis)$ is called the *prior probability*, since it expresses our initial belief in that particular hypothesis. $P(data \mid hypothesis)$ is the *likelihood* (sometimes it can be referred to as a *generative model*) of observing the data under our hypothesis. $P(data)$ is the marginal probability, and it more often than not only serves as a normalizing factor to keep our probability bounded at 1. It is hard to calculate in general, but in many cases it can be dropped, especially when we are only interested in comparing different *posteriors*.

Example: We recorded a neuron in the prefrontal cortex (PFC) of a rat. In some trials the stimulus was a cat, in the remaining trials the stimulus was a dog. Suppose we know the probabilities that the neuron emitted more or less than 5 spikes when it had seen a cat or a dog:

$$\begin{aligned} P(sp > 5 \mid cat) &= 0.9 \\ P(sp \leq 5 \mid cat) &= 0.1 \\ P(sp > 5 \mid dog) &= 0.3 \\ P(sp \leq 5 \mid cat) &= 0.7 \end{aligned}$$

We run the experiment for one more trial and find that the neuron spiked 8 times. Which of the following sentences do you agree with most?

1. The rat saw a **cat**
2. We cannot say if the rat saw a cat or a dog, we are missing $P(sp > 5)$ and $P(sp \leq 5)$
3. The rat saw a **dog**
4. We cannot say if the rat saw a cat or a dog, we are missing $P(cat)$ and $P(dog)$

This example has taught us something very important:

> Even when we know everything about how a cause (hypothesis) generates a certain response, we cannot point to the most probable cause of an observed response if we do not know how often that cause occurs!

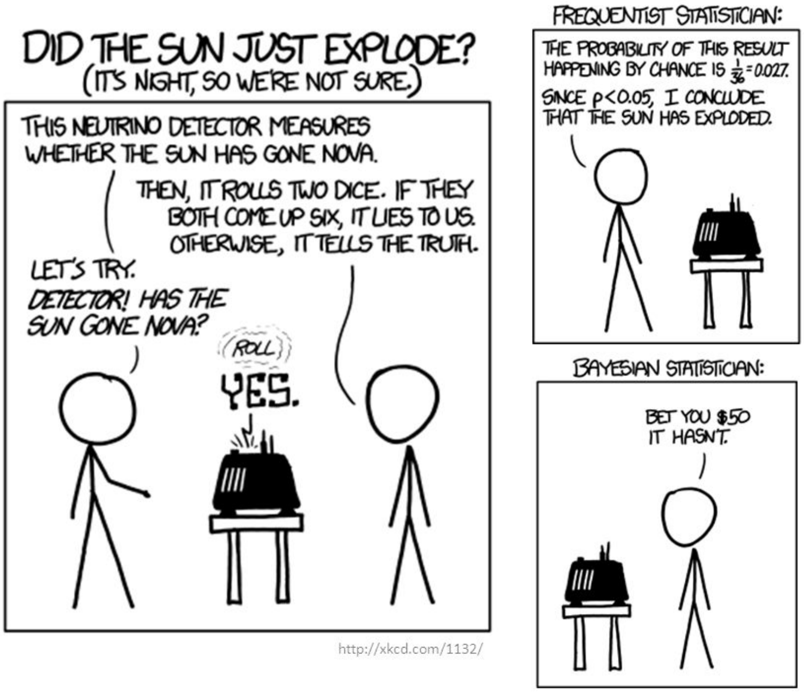
DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$=0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.
http://xkcd.com/1132/

Figure 7 The power of priors.

Note: in the context of statistical models, i.e. linear regression, hypotheses are replaced by a set of parameters, often called θ. In Bayesian regression not only is the variable y generated from a probability distribution, but the model parameters θ are assumed to come from a distribution $P(\theta)$ as well, and we can use Bayes' rule to determine $P(\theta|data)$:

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{P(data)}$$

Ex.: linear regression

$$y = ax + b$$
$$\theta = \{a, b\}$$