**DRAFT, DO NOT DISTRIBUTE**
**PROBABILITY**

## Probability

Probability has several different meanings and philosophers argue over them as if one must settle on the ***real*** meaning. But this is a mistake. Just like "cost" or "energy", "probability" is useful precisely because the same value has different interpretations. There are four interpretations that commonly come up.

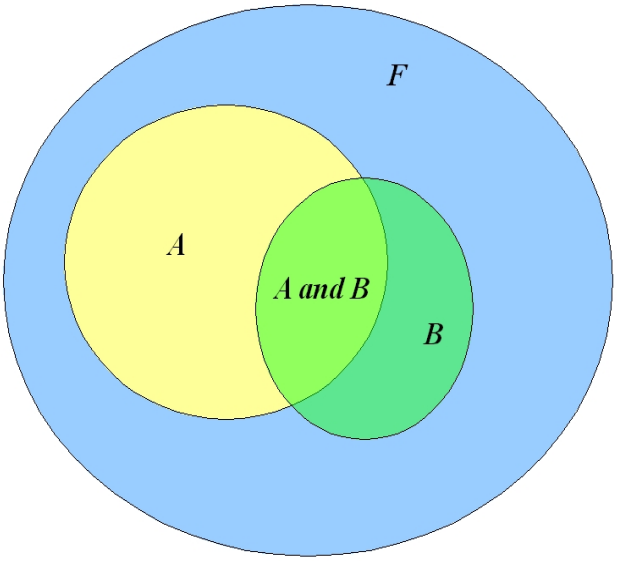
1. It has a mathematical definition that lets us manipulate it and draw inferences.
2. It has a physical interpretation as a symmetry.
3. It quantifies a degree of belief that tells us whether to act on it.
4. It has an empirical meaning that lets us measure it.

The usefulness of probability is that we can start with one of these, we can then manipulate it mathematically, and then interpret the result in one of the other ways. For example, you might observe that dice are perfectly cubical and uniform and so by (2) each face should be equally probable, i.e. P=1/6. Then you could calculate, using (1), that there are three ways of rolling a 4, . .:, : :, and .: . , out of a total of 36 possible outcomes. So the probability of a 4 on a throw is 3/36=1/12. Which tells you to only bet (3) on making a point of 4 at 12-to-1 or better odds. If you watch many game of craps and tally the results, you can approximately confirm the relative fraction of times 4 comes up (4).

## Mathematical Axioms

The mathematical definition, due to Kolmogorov, is just some rules about how a number between zero and one gets assigned to sets in a way that is consistent when you consider subsets and supersets. Here are Kolmogorov's axoims [1]:

1. Let $E$ be a set of elementary events $\{a,b,c,...\}$. Let $F$ be the set of all subsets of $E$ ; such as $A=\{\{a,b\},\{x,y,z\}\}$ or $B=\{\{b\}\}$.
2. The probability of the whole set is one. $P(F)=1$.
3. The probability of the empty set, $\emptyset$ , is zero. $P(\emptyset)=0$ .
4. If $A\cap B=\emptyset$ then $P(A\cup B)=P(A)+P(B)$
5. The probability of $B$ given $A$ is $P(B|A)=\frac{P(A\cap B)}{P(A)}$ .




The first three rules are just definitions.

The fourth says that if A and B don't have any elements in common , then the probability of A or B is the probability of A plus the probability of B. If A and B had common elements you'd be double

counting them when you add P(A)+P(B), so you need to subtract, once, the probability of the overlapping elements:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The fifth says that the probability of B given A is the probability of A and B divided by the probability of A. If you draw a Venn diagram of it, as above, it becomes obvious too. Dividing by *P(A)* is just normalizing so that *P(A | A)* =1.0.

It can also be written in an equivalent form known as Bayes' theorem after the Reverend Thomas Bayes [2]. Since $P(A \cap B)$ is symmetric in A and B we can rewrite axiom 5 as

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(B)$$

which then is rearranged into Bayes' theorem

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

This form turns up in calculating how new information should change our estimation of some hypothesis, where *B* is some hypothesis and *A* is new data. I'll come back to it in the discussion of statistics.

An important concept in probability is “independence”. It appears in the different interpretations. Physically it means there is no causal relation between the elements (including an indirect one like having a common cause). In terms of belief, it means knowledge of one doesn't change our beliefs about the other: knowing *A* doesn't change your belief about *B*. Mathematically, this is expressed as,

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

i.e. knowing *A* leaves the probability of *B* unchanged. Inserting this in (5) above

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

So when *A* and *B* are independent we have

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

Of course there is a great deal more to the mathematical theory of probability depending on the underlying sets and how the probability is assigned to subsets. In particular we will be interested in cases where the sets are intervals of values on the real line or areas in an abstract plane, as in the Venn diagram above, or volumes in a multi-dimensional space. In those cases the probabilities may be expressed by integrals. For example, we could write,

$$P(A) = \int_F f_A(s)\, ds ,$$

where $f_A(s)$ is 1.0 over the yellow circle and zero elsewhere and $ds$ is the unit of area such that

$$1.0 = \int_F ds \ .$$

$f_A(s)$ is called the probability density function of $A$. In this case it is a very simple simple function, it's either 0 or 1; but it can be a much more complicated function depicting a non-uniform distribution of probability over the elements of F. And instead of a two-dimensional region we could have many abstract dimensions. But conceptually it is all the same. Rather than try to discuss these in general, I'll leave it till we come to such examples in statistics.

**Physical Symmetry**

If we had no meaningful way to assign probability values to sets of things, the mathematical theory of probability would be useless. In practice there are two ways. One of them is based on physical symmetry. A good example is assigning probability 1/6 to each face of a die. The die is deliberately made with precise cubic symmetry and so physical interactions will not favor one face over another. This justifies modeling a throw of a die as resulting in equal probability for each face. Note that it justifies it – it doesn't necessarily make it true. Every application of a mathematical theory to the world is likely to be wrong; either in a slight degree, as an approximation, or completely off through a mistake. Very skilled persons, throwing a die onto a felt surface can do it in such a way as to bias the outcome of a throw. That's why casinos require that you use a cup and throw the dice against the end board. Also, dice can be loaded, in which case assigning equal probability could be an expensive mistake.

The use of physical symmetry to assign probability has sometimes been called “the principle of indifference”, emphasizing that because of the symmetry there is no reason to think one outcome is more likely than another and hence each of N possible outcomes should be assigned probability 1/N. But this mixes the objective physical symmetry with beliefs which only express ignorance of asymmetry. It is better to keep these concepts distinct. Which brings us to probability as a degree of belief.

**Degree of Belief**

If you are going to bet on the outcome of some process, like throwing dice, one way to do it is to assign probabilities to the possible outcomes and then bet so that your expected gain is positive. It can be proven that if you believe an assignment of the probabilities in a way inconsistent with the above axioms, then you will make bets that are certain to lose money. Only an assignment consistent with probability theory can be considered rational. Of course that doesn't mean such an assignment will guarantee you win money.

Being rational is a necessary condition for winning, but not a sufficient one. You might assume that throwing a 3 on a die has probability 1.0 and other faces have probability zero. That would still be consistent with the probability axioms. You might win, but unless the die was heavily loaded, you would most likely lose money betting on it.

On the other hand if you believed the probability of 3 was 0.7 and also that the probability of 4 was 0.7 you would make a 2-to-1 bet that 3 would appear on the next throw and also a 2-to-1 bet that 4 would appear on the next throw – a guaranteed loss, because you've violated axioms 2 and 4.

So the axioms of probability can be taken as a standard of rationality; at least within the narrow domain of believing probabilities about a given set of events or propositions. One can of course be irrational in many other ways.

**Empirical**

To say that each face of a die has probability 1/6 of coming up also implies an empirical hypothesis about the frequency of occurrence of each face in a series of throws. So we could test whether it is true by throwing a die a few thousand times. If the fraction of instances of any given face turning up was much different from 0.167 we would conclude the die was loaded and our assumption of symmetry was wrong. In this case we've taken the mathematical measure and identified it with a relative frequency. It isn't exactly the relative frequency though. The relative frequency in a finite number of throws is only an estimate of the probability; an estimate that gets better as the number of throws increases. The relative frequency will be different every time we make the throws. It's a random variable.

The last example above, illustrates the relation between statistics and probability. A statistic is just some function of a set of random variables. In the example, the average relative frequency is a statistic. It's the function that divides the number of occurrences of a given face by the number of trials. It's the average fraction of appearances by the given face. By switching over to the mathematical interpretation we can show that this statistic, the average, is a good estimator of the probability of the given face – where "good" is defined by some quantitative criteria I'll discuss later.

**Some Examples**

So there are (at least) four ways to interpret a probability assignment: A mathematical assignment of value that satisfies certain axioms. A priori – It's just based on counting possibilities or physical symmetry. Degree of belief – It quantifies a subjective judgment, as in betting. Frequency – The relative proportion observed.

Here are few examples of probability statements. Should the probability be interpreted as a mathematical assignment, a relative frequency, the result of a physical symmetry, or a degree of belief?

1. The probability of snake eyes is 1 in 36.
2. The probability of a new Honda passing it's first road test is 0.96.
3. An air-to-air missile's reliability is only 0.7, so our tactic should be to launch two at a time in order to achieve 0.91 reliability.
4. When you fly on an airliner, the chance that you will be killed in a crash is only one in 22,000.
5. We won't launch another shuttle unless the probability of success is at least 99%.
6. The probability that siblings share a particular gene is ½.
7. It's 8 to 5 the Cowboys will be in the playoffs next year.
8. The probability of a meteorite striking land is about 0.23.
9. The probability that an integer is even is 0.5.

**Random Variables**

To apply probability theory we create mathematical models of processes we're interested in and then we use these models for predictions. This is just like the application of any physical theory: the math is certain but there's always uncertainty in the modeling. When the set of events in our model is discrete,

like the faces of a die, our model may be just a simple assignment of a probability number to each event. For example P(.)=1/6, P(:)=1/6), P(.:)=1/6,... But often we're interested in results that are described by real numbers. In that case we need models that assign probabilities to continuous intervals of value; for example, the probability that a person is between 5'6" and 5'10' tall'. The values predicted by such models are called “random variables”. A model assigns probabilities to ranges or values of the random variables. The most common model of a continuous random variable is the Gaussian, aka “Normal” or “bell shaped”, distribution. It is defined by,

$$P(a<x<b)=1/\sqrt{2\pi}\int_a^b \exp(x-\mu)^2/\sigma^2\,dx$$

The factor $1/\sqrt{2\pi}$ is just to make the probability over the whole line come out to 1. The Gaussian distribution is fixed by the value of two parameters $\mu$ and $\sigma$. In general you won't know the values of these parameters, so you estimate them from sample random values. Those estimates are examples of statistics.

**Statistics**

We're going to take a sample of a random process that produces numbers. From those numbers we want to estimate the parameters in model we want to represent the random process. So we're looking for a function from a sample to one or more numbers. Such a function is called a “statistic”. The simplest example is an average. If your model is a Gaussian random distribution, then the average of a sample is the best estimating function from the sample to the parameter $\mu$ that defines the center of the distribution. It's not the only one though. For example you could take the middle value of the sample, or half-way between the lowest and highest. These are not as good, but maybe you only know the lowest and highest value.

Statistics are functions; and since they are functions of random variables their values are random variables too. So the estimate provided by a statistic has a distribution. A good statistic is one who's distribution is narrow (as little uncertainty as possible) and which converges on the true value in the limit of large samples.

Just as probability has different interpretations, so do statistics. One of the most common is the descriptive. Like the average risk of dying in an airliner crash, cited as 1 in 22,000 above, it's intended to summarize a lot of data and succinctly encapsulate what is important. This use of statistics is commonly found in newspapers; which gives rise to the warning, “Don't become a statistic”. But descriptive statistics are also used to summarize the results of scientific and engineering studies.

The other two interpretations of statistics don't coexist as peacefully as the interpretations of probability. There are contending factions that go by the name of “Frequentist” and “Bayesian”. The main area of contention is the interpretation of statistics that used to inform whether to take this or that action based on some data and some utility function.

**Frequentist**

Frequentist statistics are based on the idea of regarding whatever data we have as a random sample from some source. For example if we have a sequence of coin toss results, HHTHTHHTTH, we regard it as just a sample of the sequences we could generate. We make a decision about a model of the source by using the model to calculate the probability of the sequence we observe plus that of the other

less probable sequences. This is called a p-value. If the p-value is too low we will reject that model as not agreeing with the data. It's not the probability of the model, it's the probability of the observed result or worse, given the model. What value is too low tends to be chosen by convention of the field of application. In medicine a value of 0.05 is often used, while in particle physics a value of 0.00001 is more typical. It relates to the cost of replicating data and to the acceptability of the two different kinds of error, rejecting a true model and accepting a false model.

Frequentist statistics are those most commonly used and taught. There are tables of frequentist statistics, like Student's t-statistic, Chi-squared, and Fisher's F-statistic. But there are some fundamental problems with frequentist statistics. First is the interpretation. Decisions are based on statistics like the p-value, or Student's t, but these statistics are not probabilities; or more precisely they are not probabilities of the truth or falsity of the model. In general they can only be interpreted as the probabilities of a fictitious set of events that didn't happen. This dependence on summing up over events that didn't actually occur has the unfortunate effect of making the statistical decision depend on the sampling procedure. For example you're going to flip a coin a hundred times and decide whether it is a fair coin or not. You decide to flip it a hundred times, count the number of heads, *H*, and reject the hypothesis that it is fair if the p-value is less than 0.01. The p-value will be given by

$$p=\sum_{(h=0)}^{H}\binom{N}{h}q^{h}(1-q)^{(N-h)}$$

where sum is to *H,* the observed number of heads, assuming it's smaller than *50* or to *100-H* if it's greater, and *N=100* . So you're summing over the farther-from-fair numbers that didn't occur. But if you had decided that you would flip until you got H heads and it happened that you got that number on the 100th flip, then the above formula for p would be wrong – even though the data was exactly the same. It would be wrong because in that case you should sum over all the different numbers *N* for the fixed value of *H.* Again you're summing over the probabilities of what didn't happen.

The other problem is that the p-value is ***not*** the probability the coin isn't fair; it's the probability of the observed value or more unfair values assuming the coin ***is*** fair. So deciding to reject or accept the fairness of the coin is just based on a rule that isn't connected to the probability the coin is fair.

The Frequentist thinks all statistics should be just like descriptive statistics; they shouldn't depend on anything but the data at hand. But there's another view that this is putting blinders on your analysis.

**Bayesian**

The Bayesian points out that ignoring common sense background information can lead to silly conclusions and ignoring it is really the same as making an implicit assumption about it. Instead the background information should be explicitly included so everyone can see what for the Frequentist would be implicit. The Bayesian view is named after the Rev Thomas Bayes and his theorem,

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

It is more naturally adapted to making decisions. The left hand side, *P(B|A)*, really is the probability that *B* is true given that you've observed *A*. At least it is if you've got everything on the right hand side correct. But where the Frequentist has to make an arbitrary decision about his p-value, the Bayesian

has to know what is his prior estimate of the probability of the hypothesis he's testing, *P(B)*.

In statistics this is used to estimate the probability of a statistical model based on observed data: *B* is the event that the result came from the model with particular parameter values. *A* is the observed result. *P(A|B)* is the probability of observing *A* as computed from the model. The controversial point is the need for a "prior" probability *P(B)*, which summarizes the common sense or background knowledge. This prior summary can obviously influence the result and since it's not based on the data at hand, *A*, it can be controversial. It's a matter of judgment...but whose judgment?

Of course one is not just interested in assessing the probability of a particular model with particular parameter values; one wants to consider a range or set of hypothetical models {B} and see what the observed data implies about them. Often the models will be parameterized by real rather than integer or nominal variables. In that case Bayes theorem can be recast in terms of probability density functions,

$$f_q(q) = \frac{f_{(h|q)} f_q^0(q)}{\int_F f_{(h|q)} f_q(q)\, dq}$$

For example in evaluating the fairness of a coin, the probability of heads would be considered a variable which might take a range of values around ½. This would be represented by a kind of probability distribution for $q$, $f_{(q)}$, where the distribution quantifies our belief about $q$.

$$f_{(B|A)} = \frac{f_{(A|B)} f_{(B)}}{\int_F f_{(A|B)} f_{(B)}\, dB} = \frac{[\binom{N}{h} q^h (1-q)^{(N-h)} f_q^0(q)]}{(\int \binom{N}{h} q^h (1-q)^{(N-h)} f_q^0(q)\, dq)}$$

In general this leads to calculating some messy products of functions and integrals – and that's why the use of Bayesian inference needs computers.