

## Probability — the language of randomness

The field of statistics is concerned with decision making under uncertainty. We observe a random process (generically, an *experiment*), where multiple results (*outcomes* or *events*) are possible. The set of all possible outcomes is called the *sample space*.

What do we mean by probability? The probability that an event will occur is the chance that it will happen. But what does that really mean? There are different kinds of “chance”:

- (1) What is the chance of drawing an ace from a well-shuffled deck of cards?
- (2) What is the chance that it will rain tomorrow?
- (3) What is the chance that cancer will be cured in the next fifty years?

Each of these is a valid probability question, but they clearly refer to different notions of probability. No one definition of probability can handle all of these possibilities, so different definitions have been developed as appropriate. All of these definitions have some things in common:

- (a)  $0 \leq p \leq 1$
- (b)  $p = 0$  means that the event never occurs
- (c)  $p = 1$  means that the event always occurs

### Equally likely rule

Consider a situation where you can enumerate all of the elements of the sample space. **If there are  $n$  equally likely possibilities, and  $s$  are “favorable” to the event, then the probability of the event is**

$$\frac{s}{n}.$$

- (1) The probability of drawing an ace from a well-shuffled deck:  $4/52$
- (2) Why does three-of-a-kind beat two pair?

But, what if events are not equally likely? What if they should be, but aren't?

### Frequency theory

**In situations where it is reasonable to think of a random process as being repeatable, a reasonable definition of the probability of an event is the proportion of the time (the relative frequency) that similar events occur in the long run.**

- (1) Probabilities from a loaded (unfair) die.

(2) The weather (the probability of rain at any given point in the forecast area over the specified time period).

But, what if no “long run” can be defined? How do you define “similar events”?

### Personal (subjective) probability

These definitions of probability cannot address the kinds of informal assessments of chance that we all do all the time. (Why don't you cross against a red light on a rainy day? Because you think the chance that you'll slip, fall, and get run over is too high.) **If you would bet \$a on the event occurring to get \$b, your personal probability that the event will occur is**

$$\frac{a}{a+b}.$$

(1) An “even money” bet (say, heads on the flip of a coin) says that you would bet \$1 on “heads” to get back your dollar plus one more dollar; that is, the probability is  $1/(1+1) = 1/2$ .

(2) Parimutuel betting (horse racing, greyhound racing, jai alai).

But, different people will get different probabilities for the same events; this is inevitable, since each person's personal probability is strongly affected by that person's individual life experience. And, some people are risk averse, and don't like betting, while others are risk takers, and like it (maybe too much), which affects the bets that they are willing to take and make. We all make personal probability assessments, but sometimes they are not close to objective measures that are available.

All of these probability rules must satisfy certain conditions. In all cases,  $0 \leq p \leq 1$ . In addition, any probability rule must satisfy the **addition rule: for any two events A and B,**

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

(in any probability statement, “A or B” means “either A, or B, or both”). Say you roll one fair die. What is the probability of getting an odd number? an even number? at least a 4? both an odd and an even number (*mutually exclusive* events)? either an odd or an even number? either an odd number or at least a 4?

### Conditional probability

In many situations, we are interested in several events, and how they relate to each other. For example, does training improve performance on the job? Does the weather affect

retail sales? Does seeing a commercial change the chance that someone will purchase the advertised product? We are interested in *conditional probability* — the probability of one event **given** the occurrence of another event.

(1) marginal probability:  $P(\text{receive A in class})$

conditional probability:  $P(\text{receive A in class} \mid \text{at least 90 on first exam})$

(2) marginal probability:  $P(\text{purchase product})$

conditional probability:  $P(\text{purchase product} \mid \text{see commercial})$

It is natural to think in terms of time ordering, in that a conditional probability of  $P(B|A)$  means that  $A$  occurred first, but that's not a requirement. For example, in discrimination lawsuits, an interesting probability is  $P(\text{over age 50} \mid \text{were fired})$ .

How should conditional probabilities be defined? Consider the following situation: say we observe 1000 shoppers, and notice the following pattern (after talking to them):

Saw commercial and bought product: 50

Saw commercial and didn't buy product: 150

Didn't see commercial and bought product: 100

Didn't see commercial and didn't buy product: 700

Define the events  $A$  and  $B$  as follows:

$A$ : buying the product

$B$ : seeing the commercial

If we consider these 1000 shoppers a "long run," what is

$$P(A)?$$

$$P(B)?$$

$$P(A \text{ and } B)?$$

What is

$$P(A|B),$$

the probability of buying the product *given* seeing the commercial?

It might be easier to remember what a conditional probability is referring to if you think of the event  $A|B$  ("A given B") as "A within B" or "A in B," as that makes more explicit what the mechanism of conditional probability is: by conditioning on the event  $B$ , we are reducing the sample space to only  $B$ , and then asking what "part" of  $B$  also corresponds to  $A$ . **This naturally implies the definition of conditional probability,**

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

Generally, the definition of conditional probability results in a **multiplication rule**:

$$\begin{aligned} P(A \text{ and } B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$

Note that this implies that the probability that both of two events occurs (sometimes called a *conjunction*) must be less than or equal to the probability that either occurs separately. **A rearrangement of the terms in the two equalities above gives what is sometimes called *Bayes theorem*,**

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

*Example:* The American Statistical Association (ASA) is the leading professional organization for statistics professionals in the United States. According to that organization's 2005 Career Track Survey, 22.7% of the membership is Asian, while 27.2% of the people who have been members for 6–10 years are Asian. Further, 25.0% of the membership have been members for 6–10 years. Consider a randomly selected Asian member of the ASA. What the probability that they have not been a member for 6–10 years?

It is often the case that calculations like this are done most easily using a table, and imagining a large number of replications of the process (the “hypothetical 100,000”). Say there was a total of 100,000 members of the ASA. We are classifying them on the basis of two characteristics: being Asian or not, and having been a member for 6–10 years or not:

<i>Asian</i>	<i>Member 6–10 years</i>		
	Yes	No	
Yes	(.272)(25000) = 6800		22700
No			
	25000		100000

What is  $P(\text{Not a member for 6 – 10 years} | \text{Asian})$ ?

## Independence

Sometimes, whether one event occurs doesn't depend on whether another occurs. For example, the probability of passing a statistics exam is the same, whether you wear blue socks or not; the probability of it raining is unrelated to the day of the week (although it doesn't always seem that way!). When the knowledge that one event has occurred (or will occur) gives no additional information about whether another event will occur, we say that the two events are **independent**. More formally, **two events  $A$  and  $B$  are independent if and only if**

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

( $A$  and  $B$  are arbitrary here, so we could just as well write  $P(B|A) = P(B)$ ). **Equivalently (using the multiplication rule), two events  $A$  and  $B$  are independent if and only if**

$$P(A \text{ and } B) = P(A)P(B).$$

- (1) Consider the Daily Numbers game, where an integer in the range 000 through 999 is chosen as the winning number. The way the number is chosen is as follows. There are three separate machines, each of which has 10 ping pong balls (numbered 0 through 9) rapidly bouncing around. The number is chosen by choosing one ball from each machine in a specified order. Does this make all 1000 possible numbers equally likely?
- (2) Pennsylvania lottery fraud; Arizona lottery mistake.

Note that the independence characteristics of  $A$  and  $B$  are the same as those involving the complement of  $A$  ( $A'$ ) and the complement of  $B$  ( $B'$ ); that is, for example,  $A$  and  $B$  being independent is equivalent to  $A'$  and  $B'$  being independent.

It is important to remember that the association between two random variables (or the lack of association, as when they are independent) can only be seen in the *joint* distribution of the two variables (that is, the complete set of probabilities for all pairs of outcomes from each random variable); the *marginal* distributions of each random variable provide **no** information about association.

“The more numerous the observations and the less they vary among themselves, the more their results approach the truth.”

— Pierre Simon, Marquis de Laplace

“In the long run, we are all dead.”

— John Maynard Keynes

“The theory of probability is at bottom nothing but common sense reduced to calculus.”

— Pierre Simon, Marquis de Laplace

“I shall never believe that God plays dice with the universe.”

— Albert Einstein

“All Nature is but Art, unknown to thee;  
All Chance, Direction, which thou canst see;  
All Discord, Harmony not understood.”

— Alexander Pope

“If there is a 50–50 chance that something can go wrong, then 9 times out of 10 it will.”

— Paul Harvey

“A fool must now and then be right by chance.”

— William Cowper

“If you bet on a horse, that’s gambling. If you bet you can make three spades, that’s entertainment. If you bet cotton will go up three points, that’s business. See the difference?”

— Blackie Sherrod

“It is impossible to calculate accurately events that are determined by chance.”

— Thucydides

“But to *us*, probability is the very guide of life.”

— Bishop J. Butler

“We shall have to get used to thinking of chance not as an irritating obstacle, not as an ‘inessential adjunct to phenomena’ (as is in the philosophical dictionary), but as a source of unlimited possibilities of which even the boldest imagination can have no prescience.”

— Leonard A. Rastrigin

“A probability is a generally approved proposition; what men know to happen or not to happen, to be or not to be, for the most part thus and thus is a probability. It is the nature of probability that improbable things will happen.”

— Aristotle, *Analytica Priora*

“It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge . . . The most important questions of life are, for the most part, really only problems of probability.”

— Pierre Simon, Marquis de Laplace

“The greatest advantage from gambling comes from not playing it at all.”

— Girolamo Cardano (*Book on Games of Chance*)

“Where observation is concerned, chance favours only the prepared mind.”

— Louis Pasteur

“We know that people tend to overestimate the frequency of well-publicized, spectacular events compared with more commonplace ones; this is a well-understood phenomenon in the literature of risk assessment and leads to the truism that when statistics plays folklore, folklore always wins in a rout.”

— Donald Kennedy

“This is plain truth: everyone ought to keep a sharp eye on main chance.”

— Plautus

“We are hard-wired to react to coincidences. It goes back to primitive man. You look in the bush, it looks like stripes, you’d better get out of there before you determine the odds that you’re looking at a tiger. The cost of being flattened by the tiger is high.”

— Persi Diaconis

“Probability is too important to be left to the experts.”

— Richard W. Hamming

“Chance is a nickname for Providence.”

— Sebastien Roch Nicolas Chamfort

“ . . . Chance governs all.”

— John Milton (*Paradise Lost*)

“Nay, what should mortal fear, for whom the degree of fortune are supreme, and who hath clear foresight of nothing? ’Tis best to live at random, as one may.”

— Sophocles (*Edipus the King*)

“Nothing is certain, and that’s for sure.”

— Marcel P.R. Van den Broecke

“Rowe’s Rule: the odds are six to five that the light at the end of the tunnel is the headlight of an oncoming train.”

— Paul Dickson

“The lot is cast into the lap; but the whole disposing thereof is of the Lord.”

— *Proverbs* 16:33

“The chances for the baby’s chromosomes being defective were 250 to 1, which sounds like reasonable odds. Except that all odds are, in reality, 50–50: it may happen and it may not.”

— *The London Times Magazine*, 2 August 1997

“Our brains are just not wired to do probability problems very well.”

— Persi Diaconis

“I long ago came to the conclusion that all life is 6 to 5 against.”

— Damon Runyon

“From where we stand the rain seems random. If we would stand somewhere else, we would see the order in it.”

— Tony Hillerman (*Coyote Waits*)

“Coincidences, in general, are great stumbling blocks in the way of that class of thinkers who have been educated to know nothing of the theory of probabilities — that theory to

which the most glorious of human research are indebted for the most glorious of illustration.”

— Edgar Allan Poe (*The Murders in the Rue Morgue*)

“Misunderstanding of probability may be the greatest of all impediments to scientific literacy.”

— Stephen Jay Gould

“The race is not always to the swift, nor the battle to the strong — but that’s the way to bet.”

— Damon Runyon