Caltech Department of Mathematics

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2017

Lecture 2: Random Experiments; Probability Spaces; Random Variables; Independence

Relevant textbook passages:

Pitman [4]: Sections 1.3–1.4., pp. 26–46.

Larsen–Marx [3]: Sections 2.2–2.5, pp. 18–66.

The great coin-flipping experiment

This year there were 194 submissions of 128 flips, for a total of 24,832 tosses! You can find the data at http://www.math.caltech.edu/~2016-17/2term/ma003/Data/FlipsMaster.txt

Recall that I put predictions into a sealed envelope. Here are the predictions of the average number of runs, by length, compared to the experimental results.

Run length	Theoretical average	Predicted range	Total runs	Average runs	How well did I do?
1	32.5	31.3667 –33.6417	6340	32.680412	Nailed it.
2	16.125	15.4583 –16.8000	3148	16.226804	Nailed it.
3	8	7.5500 – 8.4583	1578	8.134021	Nailed it.
4	3.96875	3.6417 – 4.3000	725	3.737113	Nailed it.
5	1.96875	1.7333 – 2.2083	388	2.000000	Nailed it.
6	0.976563	0.8083 – 1.1500	187	0.963918	Nailed it.
7	0.484375	0.3667 – 0.6083	101	0.520619	Nailed it.
8	0.240234	0.1583 – 0.3333	49	0.252577	Nailed it.
9	0.119141	0.0583 – 0.1833	16	0.082474	Nailed it.
10	0.059082	0.0167 – 0.1083	12	0.061856	Nailed it.
11	0.0292969	0.0000 – 0.0667	9	0.046392	Nailed it.
12	0.0145264	0.0000 – 0.0417	2	0.010309	Nailed it.
13	0.00720215	0.0000 – 0.0250	0	0.000000	Nailed it.
14	0.00357056	0.0000 – 0.0167	1	0.005155	Nailed it.
15	0.00177002	0.0000 – 0.0083	0	0.000000	Nailed it.

[a] The formula for the theoretical average is the object of an optional Exercise.
[b] This is based on a Monte Carlo simulation of the 95% confidence interval for a sample size of 120, not 194.

Yes! There are Laws of Chance.

How did we do on Heads versus Tails? Out of 24,832 there were:

	Number	Percent
Tails	12,507	50.366
Heads	12,325	49.634

How close to 50/50 is this? We'll see in a bit.

2.1 Probability measures

Recall from last time that a **probability measure** or **probability distribution** (as in Pitman [4]) or simply a **probability** (although this usage can be confusing) is a **set function** $P\colon \mathcal{E} \to [0,1]$ that satisfies:

Pitman [4]: § 1.3

Larsen–Marx [3]: § 2.3

Normalization $P(\varnothing) = 0$; and $P(S) = 1$.

Nonnegativity For each event E, we have $P(E) \geqslant 0$.

Additivity If $EF = \varnothing$, then $P(\cup F) = P(E) + P(F)$.

Note that while the domain of P is technically $\mathcal{E}$, the set of events, we may also refer to P as a probability (measure) on S, the set of samples.

2.1.1 Remark To cut down on the number of delimiters in our notation, we may omit the some of them simply write something like $P\left(f(s) = 1\right)$ or $P\{s \in S : f(s) = 1\}$ instead of $P\big(\{s \in S : f(s) = 1\}\big)$ and we may write $P(s)$ instead of $P\big(\{s\}\big)$. You will come to appreciate this.

2.1.1 Elementary Probability Identities

Larsen–Marx [3]: § 2.3

1.
$$P(E^{\mathrm{c}}) = 1 - P(E)$$

2. If $F \subset E$, then
$$P(E \setminus F) = P(E) - P(F)$$

3. If $F \subset E$, then
$$P(F) \leqslant P(E)$$

4. If $E_1, \ldots, E_n$ are **pairwise disjoint**, i.e., $i \neq j \implies E_i E_j = \varnothing$, then
$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i).$$

Proof of last: Let $\mathbb{P}(n)$ stand for the proposition for n. Then $\mathbb{P}(2)$ is just Additivity. Assume $\mathbb{P}(n-1)$. Write
$$\bigcup_{i=1}^{n} E_i = \underbrace{\bigcup_{i=1}^{n-1} E_i}_{=B} \cup E_n$$

Then $BE_n = \varnothing$, so
$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) &= P\left(B \cup E_n\right) \\ &= P(B) + P(E_n) \qquad \text{by Additivity} \\ &= \sum_{i=1}^{n-1} P(E_i) + P(E_n) \quad \text{by } \mathbb{P}(n-1) \\ &= \sum_{i=1}^{n} P(E_i). \end{aligned}$$

2.1.2 Boole's Inequality *Even if events $E_1, \ldots, E_n$ are not pairwise disjoint,*
$$P\left(\bigcup_{i=1}^{n} E_i\right) \leqslant \sum_{i=1}^{n} P(E_i).$$

Before I demonstrate how to prove Boole's Inequality, let me describe a "trick" for "disjuncting" a sequence of sets.

2.1.3 Lemma *Let $E_1, \dots, E_n, \dots$ be a sequence (finite or infinite) of events. The there is a sequence $A_1, \dots, A_i, \dots$ of events such that:*

- *For each n, $A_n \subset E_n$.*
- *The A_i's are pairwise disjoint. That is, $i \neq j$, $A_i A_j = \varnothing$.*
- *For each n,*

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} E_i.$$

Proof: We define the sequence recursively. Set

$$A_1 = E_1.$$

Having defined $A_1, \dots, A_k$, define

$$A_{k+1} = E_{k+1} \setminus A_k.$$

A simple proof by induction completes the proof. ∎

Proof of Boole's Inequality: Let A_i be a sequence of pairwise disjoint events satisfying the conclusion of Lemma 2.1.3. That is, each $A_i \subset E_i$ and $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} E_i$. Then

$$P\left(\bigcup_{i=1}^{n} E_i\right) = P\left(\bigcup_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} P(E_i),$$

where the inequality follows from $P(A_i) \leqslant P(E_i)$ for each i. ∎

2.1.2 Odds

I find it exasperating that even generally linguistically reliable sources, such as *The New York Times*, confuse probabilities and odds. **Pitman [4]:** pp. 6–8

2.1.4 Definition *The **odds against the event E** is the ratio*

$$\frac{P(E^{\mathrm{c}})}{P(E)}.$$

That is, it is a ratio of probabilities, not a probability. It is usually spoken as "the odds against the event E are $P(E^{\mathrm{c}})/P(E)$ to one," or as a ratio of integers. (That is, we typically say "3 to 2" instead of "1 1/2 to 1.")

*The **odds in favor of the event E** is*

$$\frac{P(E)}{P(E^{\mathrm{c}})}.$$

This is to be distinguished from the **payoff odds**. The payoff odds are the ratio of the amount won to the amount wagered for a simple **bet**. For instance, in roulette, if you bet \$1 on the number 2 and 2 comes up, you get a payoff of \$35, so the payoff odds are "35 to 1." But (assuming that all numbers on a roulette wheel are equally likely) the odds against 2 are 37 to one since a roulette wheel has the "numbers" 0 and 00 in addition to the numbers 1 through 36.[1] [2] [3] (Pitman [4, p. 7] describes the outcomes and bets for a roulette wheel.)

Unfortunately, you often run across statements such as, "the odds are one in ten that X will happen," when the author probably means, "the probability that X will happen is one-tenth," so that the odds in favor of X happening are one to nine.

[1] Actually, there are (at least) two kinds of roulette wheels. In Las Vegas, roulette wheels have 0 and 00, but in Monte Carlo, the 00 is missing.

[2] The term roulette wheel is a pleonasm, since *roulette* is French for "little wheel."

[3] The word "pleonasm" is one of my favorites. Look it up.

2.2 ⋆ Countable additivity

Most probabilists assume further that $\mathcal{E}$ is a ***σ*-algebra** or ***σ*-field**, which requires in addition that

3′. If $E_1, E_2, \ldots$ belong to $\mathcal{E}$, then $\bigcap_{i=1}^{\infty} E_i$ and $\bigcup_{i=1}^{\infty} E_i$ belong to $\mathcal{E}$.

Note that if S is finite and $\mathcal{E}$ is an algebra then, it is automatically a σ-algebra. Why? Since there are only finitely many subsets of a finite set, any infinite sequence of sets has only finite may distinct sets, the rest are all copies of something else. So any infinite intersection or union is actually the same as some finite intersection or union.

Most probabilists also require the following stronger property, called **countable additivity**:

Countable additivity $P\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} P(E_i)$ provided $E_i \cap E_j = \varnothing$ for $i \neq j$.

2.2.1 Remark If the sample space S is finite, it has only finitely many subsets, so the only way an infinite sequence $E_1, E_2, \ldots$ can be pairwise disjoint, is if all but finitely many of the events E_i are equal to the empty set. In this case, since $P(\varnothing) = 0$, the infinite series $\sum_{i=1}^{\infty} P(E_i)$ reduces to a finite sum. In other words, for a finite sample space, additivity guarantees countable additivity. (Cf. Section 2.1.1, item 4.)

You need to take an advanced analysis course to understand that for infinite sample spaces, there can be probability measures that are additive, but not countably additive. So don't worry too much about it.

The next results may seem theoretical and of no practical relevance, but they are crucial to understanding the properties of cumulative distribution functions.

A sequence $E_1, \ldots, E_n \ldots$ of events is decreasing, written $E_n \downarrow$, if

$$E_1 \supset E_2 \supset \cdots \supset E_n \supset \cdots .$$

A sequence $E_1, \ldots, E_n \ldots$ of events is increasing, written $E_n \uparrow$, if

$$E_1 \subset E_2 \subset \cdots \subset E_n \subset \cdots .$$

Add the proof.

2.2.2 Proposition (Continuity and countable additivity) *If P is an additive probability, then*

1. *P is countably additive if and only if $E_n \downarrow$ implies $P\Big(\bigcap_n E_n\Big) = \lim_n P(E_n)$.*

2. *P is countably additive if and only if $E_n \uparrow$ implies $P\Big(\bigcup_n E_n\Big) = \lim_n P(E_n)$.*

2.3 Probability spaces

Our complete formal model of a **random experiment** is what we call a probability space.

2.3.1 Definition *A **probability space** is a triple $(S, \mathcal{E}, P)$, where S is a nonempty set, the **sample space** or **outcome space** of the experiment, $\mathcal{E}$ is the set of **events**, which is a σ-field of subsets of S, and P is a countably additive **probability** measure on $\mathcal{E}$.*

2.3.1 An example: Uniform probability

The **uniform probability** on a finite sample space S makes each outcome equally likely, and every subset of S is an event. This formalizes Laplace's model of probability.

2.3.2 Theorem (Uniform probability) *With a uniform probability P on a finite set S, then for any subset E of S,*

$$P(E) = \frac{|E|}{|S|}.$$

Throughout this course and in daily life, if you come across the phrase **at random** and the sample space is finite, unless otherwise specified, you should assume the probability measure is uniform.

2.3.3 Example (Coin Tossing) We usually think of a coin as being equally likely to come up H as T. That is, $P\{H\} = P\{T\}$. If our sample space is the simple $S = \{H, T\}$ and $\mathcal{E}$ is all four subsets of S, $\mathcal{E} = \big\{\varnothing, S, \{H\}, \{T\}\big\}$, then

$$\{H\}\{T\} = \varnothing \text{ and } \{H\} \cup \{T\} = S$$

so

$$1 = P(S) = P\big(\{H\} \cup \{T\}\big) = P\{H\} + P\{T\},$$

which, since $P\{H\} = P\{T\}$, implies

$$P\{H\} = P\{T\} = 1/2.$$

□

2.4 Additivity and the Inclusion–Exclusion Principle

The **Inclusion–Exclusion Principle** describes the full power of additivity of probability measures when applied to unions of not necessarily pairwise disjoint sets. Early on, we expect small children to understand the relation between sets and their cardinality—If Alex has three apples and Blair has two apples, then how many apples do they have together? The implicit assumption is that the two sets of apples are disjoint (since they belong to different children), then the measure (count) of the union is the sum of the counts. But what if Alex and Blair own some of their apples in common?

Pitman [4]: p. 22

2.4.1 Proposition (Inclusion–Exclusion Principle, I) *Even if $AB \neq \varnothing$,*

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

Proof: Now

$$A \cup B = (AB^{\mathrm{c}}) \cup (AB) \cup (A^{\mathrm{c}}B).$$

The three sets being unioned on the right-hand side are pairwise disjoint:

$$(AB^{\mathrm{c}})(AB) = \varnothing$$
$$(AB)(A^{\mathrm{c}}B) = \varnothing$$
$$(AB^{\mathrm{c}})(A^{\mathrm{c}}B) = \varnothing.$$

Therefore, by finite additivity,

$$P(A \cup B) = P(AB^{\mathrm{c}}) + P(AB) + P(A^{\mathrm{c}}B).$$

Now also by additivity,

$$\begin{aligned} P(A) &= P(AB^{\mathrm{c}}) + P(AB) \\ P(B) &= P(BA^{\mathrm{c}}) + P(AB). \end{aligned}$$

So, adding and regrouping,

$$\begin{aligned} P(A) + P(B) &= \underbrace{P(AB^{\mathrm{c}}) + P(AB) + P(BA^{\mathrm{c}})}\, + P(AB) \\ &= P(A \cup B) + P(AB). \end{aligned}$$

This implies

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

∎

Additionally,

$$\begin{aligned} P(A \cup B \cup C) = \;& P(A) + P(B) + P(C) \\ & - P(AB) - P(AC) - P(BC) \\ & + P(ABC). \end{aligned}$$

To see this refer to Figure 2.1. The events A, B, and C are represented by the three circles.

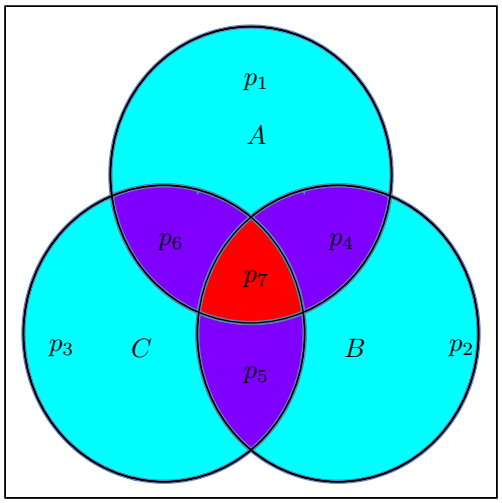

Figure 2.1. Inclusion-Exclusion for three sets.

The probability of each shaded region is designated by p_i, $i = 1, \dots, 7$. Observe that

$$\begin{aligned} P(A \cup B \cup C) &= p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 \\ P(A) &= p_1 + p_4 + p_6 + p_7 \\ P(B) &= p_2 + p_4 + p_5 + p_7 \\ P(C) &= p_3 + p_5 + p_6 + p_7 \\ P(AB) &= p_4 + p_7 \\ P(AC) &= p_6 + p_7 \\ P(BC) &= p_5 + p_7 \\ P(ABC) &= p_7. \end{aligned}$$

Thus

$$\begin{aligned} P(A) + P(B) + P(C) &= p_1 + p_2 + p_3 + 2p_4 + 2p_5 + 2p_6 + 3p_7 \\ P(AB) + P(AC) + P(BC) &= p_4 + p_5 + p_6 + 3p_7. \end{aligned}$$

So

$$\begin{aligned} \big[P(A) + P(B) + P(C)\big] - \big[P(AB) + P(AC) + P(BC)\big] + P(ABC) \\ = p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 = P(A \cup B \cup C). \end{aligned}$$

∎

The general version of the Inclusion–Exclusion Principle may be found in Pitman [4], Exercise 1.3.12, p. 31.

2.4.2 Proposition (General Inclusion–Exclusion Principle)

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) = &\sum_i P(E_i) \\ &- \sum_{i<j} P(E_i E_j) \\ &+ \sum_{i<j<k} P(E_i E_j E_k) \\ &\vdots \\ &+ (-1)^{n+1} P(E_1 E_2 \dots E_n). \end{aligned}$$

(Recall that intersection is denoted by placing sets next to each other. Note that the sign preceding a sum with the intersection of m sets is $(-1)^{m+1}$. The reason for summing over increasing indices is to avoid double counting.)

Note that if the sets are pairwise disjoint, the intersections above are all empty and so have probability zero, and this reduces to finite additivity.

While it is possible to prove this result now using induction, I will put off a proof until we learn about the expectation of random variables, which will make the proof much easier.

2.5 Random variables and random vectors

For some reason, your textbooks postpone the definition of random variables, even though they are fundamental concepts.

A **random variable** is a numerical measurement of the outcome of a random experiment.

2.5.1 Example (Some random variables) Here are some examples of random variables.

- The random experiment is to roll two dice, so the sample space is the set of ordered pairs of integers from 1 through 6. The sum of these to numbers is a random variable. The pair itself is a random vector. The difference of the numbers is another random variable.

- The experiment is to roll two dice repeatedly until boxcars (a sum of twelve) appear. The number of rolls is a random variable, which may take on the value ∞ if boxcars never appear. (This is an idealization of course.)

- The random experiment takes a sample of blood and smears it on a microscope slide with a little rectangle marked on it. The number of platelets lying in the rectangle is a random variable.

- The experiment is to record all the earthquakes in Southern California. The number of magnitude 5+ earthquakes in a year is a random variable.

- A letter is drawn at random from the alphabet. If we assign a number to each letter, then that number is a random variable. But unlike the cases above it doe not make sense to take the results of two such experiments and average them. What is the average of 'a' and 'b'? In such cases, where the result of the experiment is **categorical** and not inherently numeric, it may make more sense to take the outcome to be a random vector, indexed by the categories. This interpretation is often used in communication theory by electrical engineers, e.g., Robert Gray [2] or Thomas Cover and Joy Thomas [1].

- An experiment by Rutherford, Chadwick, and Ellis counted the number of α-particles emitted by a radioactive sample for 7.5 second time intervals. Each count is a random variable.

□

Being numerical, we can add random variables, take ratios, etc., to get new random variables. But to understand how these are related we have to go back to our formal model of random experiments as probability spaces, and define random variables in terms of a probability space.

2.5.2 Definition *A **random variable** on a probability space $(S, \mathcal{E}, P)$ is an (extended)[a] real-valued function on S which has the property that for every interval $I \subset \boldsymbol{R}$ the inverse image of I is an event.*[b]

*A **random vector** is simply a finite-dimensional vector (ordered list) of random variables.*

[a] The extended real numbers include two additional symbols, ∞ and $-\infty$. We'll have more to say about them later.

[b] Note that when the collection $\mathcal{E}$ of events consists of all subsets of S, then the requirement that inverse images of intervals be events is automatically satisfied.

So a random variable is not a variable in the usual sense of the word "variable" in mathematics. A random variable is simply an (extended) real-valued function. Traditionally, probabilists and statisticians use upper-case Latin letters near the end of the alphabet to denote random variables. This has confused generations of students, who have trouble thinking of random variables as functions. For the sake of tradition, and so that you get used to it, we follow suit. So a **random variable** X is a *function*

$$X \colon S \to \boldsymbol{R} \quad \text{such that for each interval } I, \quad \{s \in S : X(s) \in I\} \in \mathcal{E}.$$

We shall adopt the following notational convention, which I refer to as **statistician's notation**, that

$$(X \in I) \text{ means } \{s \in S : X(s) \in I\}.$$

Likewise $(X \leqslant t)$ means $\{s \in S : X(s) \leqslant t\}$, etc.

If E belongs to $\mathcal{E}$, then its **indicator function** $\mathbf{1}_E$, defined by

$$\mathbf{1}_E(s) = \begin{cases} 0 & s \notin E \\ 1 & s \in E, \end{cases}$$

is a random variable.

A random variable X, is a **mapping** from the sample space S to the real numbers, that is, X maps each point $s \in S$ to a real number $X(s)$. The function X is different from its value $X(s)$ at the point s, which is simply a real number. The value $X(s)$ is frequently referred to as a **realization** of the random variable X. A realization is just the value that X takes on for some outcome s in the sample space.

2.6 Independent events

Larsen–Marx [3]: § 2.5
Pitman [4]: § 1.4

2.6.1 Definition *Events E and F are **(stochastically) independent** if*

$$P(EF) = P(E) \cdot P(F).$$

2.6.2 Lemma *If E and F are independent, then E and F^{c} are independent; and E^{c} and F^{c} are independent; and E^{c} and F are independent.*

Proof: It suffices to prove that if E and F are independent, then E^{c} and F are independent. The other conclusions follow by symmetry. So write

$$F = (EF) \cup (E^{\mathrm{c}}F),$$

so by additivity

$$P(F) = P(EF) + P(E^{\mathrm{c}}F) = P(E)P(F) + P(E^{\mathrm{c}}F),$$

where the second equality follows from the independence of E and F. Now solve for $P(E^{\mathrm{c}}F)$ to get

$$P(E^{\mathrm{c}}F) = \big(1 - P(E)\big)P(F) = P(E^{\mathrm{c}})P(F).$$

But this is just the definition of independence of E^{c} and F. ∎

2.7 Repeated experiments and product spaces

One of the chief uses of the theory of probability is to understand long-run frequencies of outcomes of experiments. If S is the sample space of a random experiment, and we repeat the experiment, we essentially have a compound experiment whose sample space is the Cartesian product $S^2 = S \times S$, the set of ordered pairs of outcomes. Similarly, the sample space for n repetitions of the experiments is $S^n = \underbrace{S \times \cdots \times S}_{n \text{ copies of } S}$.

Any set of the form

$$E = E_1 \times \cdots \times E_n,$$

where each E_i is an event in $\mathcal{E}$, the common set of events for a single experiment is an event for the repeated experiment. Such a set is called a **rectangle**. (Geometrical rectangles are products of intervals.) But since we want the set of events in the compound experiment to be a σ-algebra we must add some more events. What we want is the smallest σ-algebra that includes all the rectangles. This σ-algebra of events is called the **product σ-algebra** and is denoted $\mathcal{E}^n$.

For example consider rolling a die, where every subset of $\{1, \dots, 6\}$ of is an event in $\mathcal{E}$. The sample space for the repeated experiment is the set of ordered pairs of the numbers one through six. The event "both tosses give the same result" is not a rectangular event, but it is the union of finitely many rectangles: $= \bigcup_{i=1}^{6} \big(\{i\} \times \{i\}\big)$, and so is an event in the product σ-algebra. So is the complementary event, "the rolls are different."

2.8 Independent repeated experiments

Our mathematical model of a random experiment is a probability space $(S, \mathcal{E}, P)$. And of the repeated experiment is $(S^2, \mathcal{E}^2, ?)$. The question mark is there because we need to decide the probabilities of events in a repeated experiment. To do this in a simple way we shall consider the case where the *experiments* are independent. That is, the outcome of the first experiment provides no information about the outcome of the second experiment.

Consider a compound event $E_1 \times E_2$, which means that the outcome of the first experiment was in $E_1 \in \mathcal{E}$ and the outcome of the second experiment was in $E_2 \in \mathcal{E}$. The event E_1 in the first experiment is really the event $\boldsymbol{E}_1 \times S$ in the compound experiment. That is, it is the set of all ordered pairs where the first coordinate belongs to E_1. Similarly the event E_2 in the second experiment corresponds to the event $S \times E_2$ in the compound experiment. Now observe that

Add a picture, and some examples.

$$(E_1 \times S) \bigcap (S \times E_2) = E_1 \times E_2.$$

Since the experiments are independent the probability of the intersection $(E_1 \times S)(S \times E_2)$ should be the probability of $(E_1 \times S)$ times the probability of $(S \times E_2)$. But these probabilities are just $P(E_1)$ and $P(E_2)$ respectively. Thus for independently repeated experiments and "rectangular events,"

$$\text{Prob}(E_1 \times E_2) = P(E_1) \times P(E_2).$$

This is enough to pin down the probability of all the events in the product algebra $\mathcal{E}^2$, and the resulting probability measure is called the product probability, and may be denoted by $P \times P$, or P^2, or by really abusing notation, simply P again.

The point to remember is that independent experiments give rise to products of probabilities.

How do we know when two experiments are independent? We rely on our knowledge of physics or biology or whatever to tell us that the outcome of one experiment yields no information on the outcome of the other. It's built into our modeling decision. I am no expert, but my understanding is that quantum entanglement implies that experiments that our intuition tells are independent are not really independent.[4] But that is an exceptional case. For coin tossing, die rolling, roulette spinning, etc., independence is probably a good modeling choice.

2.9 ⋆ A digression on infinity

Now consider an experiment with a **stopping rule**. For example, consider the experiment, "toss a coin until Heads occurs, then stop." What is the natural sample space, and set of events for this experiment. You might think the simplest sample space for this experiment is

$$\bigcup_{n=1}^{\infty} S^n,$$

[4] I asked David Politzer if this was a fair statement, and he gave his blessing.

the set of finite-length tuples of elements of S. This sample space is infinite, but at least is a nice infinity—it is countably infinite.

The event H_n that the first Head occurs on the n^{th} toss belongs $\mathcal{E}^n$, and so it should be an event in the larger experiment. Now consider the event

$$H = (\text{a Head eventually occurs}).$$

The event H is the infinite union $\cup_{n=1}^{\infty} H_n$. Is this union an event as we have defined things? No, it is not. One way to see this is to ask, what is the complement of H? It would be the event that no Head occurs so we would have to toss forever. But the infinite sequence of all Tails (while admittedly a probability zero occurrence) does not appear in our sample space. Another way to say this is that $\overset{\infty}{\underset{n=1}{\cup}} \mathcal{E}^n$ is not a σ-algebra. So if we want the set of events to include H, we need to do something drastic.

One possibility is never to consider H to be an event. After all, how could we ever "observe" such an event happening? In other words, we could say that ensuring that the set of events is a σ-algebra instead of merely an *algebra* is not worth the trouble. (Your textbooks simply ignore this difficulty, and they are still full of useful results.)

On the other hand, we might really care about the probability of the event H. If we want to do that, we have to agree that the real sample space is actually the set of all infinite sequences of outcomes of the original experiment. That is, the sample space is S^∞, and not $\bigcup_{n=1}^{\infty} S^n$. Even if S has only two points S^∞ is uncountably infinite. (Think of binary expansions of real numbers in the unit interval.)

Each approach has its advantages and disadvantages. I should discuss some of these issues in an appendix for the mathematically inclined, and will when I can find the time. Fortunately, there are still plenty of interesting things we can say without having to worry about making H an event.

2.10 Generally accepted counting principles

The Uniform Probability (or counting) model was the earliest and hence one of the most pervasive probability models. For that reason it is important to learn to count. This is the reason that probability and combinatorics are closely related.

2.10.1 Lists versus sets

I find it very useful to distinguish **lists** and **sets**. Both are collections of n objects, but two lists are different unless each object appears in the same *position* in both lists.

For instance,

$$123 \quad \text{and} \quad 213 \quad \text{are distinct lists of three elements, but the same set.}$$

A list is sometimes referred to as a **permutation** and a set is often referred to as **combination**.

2.10.2 Number of lists of length n

If I have n distinct objects, how many distinct ways can I arrange them into a list (without repetition)? Think of the objects being numbered and starting out in a bag and having to be distributed among n numbered boxes.

There are n choices for box 1, and for each such choice, there are $n-1$ for position 2, etc., so all together

there are $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$ distinct lists of n objects.

The number $n!$ is read as "***n* factorial**."

By definition,

$$0! = 1,$$

and we have the following recursion

$$n! = n \cdot (n-1)! \qquad (n > 0).$$

By convention, if $n < 0$, then $n! = 0$.

2.10.3 Number of lists of length k of n objects

How many distinct lists of length k can I make with n objects? As before, there are n choices of the first position on the lists, and then $n-1$ choices for the second position, etc., down to $n-(k-1) = n-k+1$ choices for the k^{th} position on the list. Thus there are

$$\underbrace{n \times (n-1) \times \cdots \times (n-k+1)}_{k \text{ terms}}$$

distinct lists of k items chosen from n items. There is a more compact way to write this. Observe that

$$\begin{aligned} & n \times (n-1) \times \cdots \times (n-k+1) \\ =& \frac{n \times (n-1) \times \cdots \times (n-k+1) \times (n-k) \times (n-k-1) \times \cdots \times 2 \times 1}{(n-k) \times (n-k-1) \times \cdots \times 2 \times 1} \\ =& \frac{n!}{(n-k)!} \end{aligned}$$

There are $\dfrac{n!}{(n-k)!}$ distinct lists of length k chosen from n objects.

We may write this as $(n)_k$, read "n order k." Note that when $k = n$ this reduces to $n!$ (since $0! = 1$), which agrees with the result in the previous section. When $k = 0$ this reduces to 1, and there is exactly one list of 0 objects, namely, the empty list.

2.10.4 Number of subsets of size k of n objects

How many distinct subsets of size k can I make with n objects? (A subset is sometimes referred to as a **combination** of elements.) Well there are $\frac{n!}{(n-k)!}$ distinct lists of length k chosen from n objects. But when I have a set of k objects, I can write it $k!$ different ways as a list. Thus each set appears $k!$ times in my listing of lists. So I have to take the number above and divide it by $k!$ to get the number of.

There are $\dfrac{n!}{(n-k)!\cdot k!}$ distinct subsets of size k chosen from n objects.

2.10.1 Definition *For natural numbers* $0 \leqslant k \leqslant n$

$$\binom{n}{k} = \frac{n!}{(n-k)!\cdot k!},$$

is read as

"n choose k"

For $k > n$ *define* $\binom{n}{k} = 0$.

It is the number of distinct subsets of size k chosen from a set with n elements. It is also known as the **binomial coefficient**. Note if $k > n$, there are no subsets of size k of a set of size n, so by convention we agree that in this case $\binom{n}{k} = 0$.

Other notations you may encounter include $C(n,k)$, ${}^nC_k$, and ${}_nC_k$. (These notations are easier to typeset in lines of text.)

2.10.5 Some useful identities

$$\binom{n}{0} = \binom{n}{n} = 1$$

$$\binom{n}{1} = n$$

$$\binom{n}{k} = \binom{n}{n-k}$$

$$\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k} \tag{1}$$

Here is a simple proof of (1): $\binom{n+1}{k+1}$ is the number of subsets of size $k+1$ of a set A with $n+1$ elements. So fix some element $\bar{a} \in A$ and put $B = A \setminus \{\bar{a}\}$. If E is a subset of A of size $k+1$, then either (i) $E \subset B$, or else (ii) E consists of $\bar{a}$ and k elements of B . There are $\binom{n}{k+1}$ subsets E satisfying (i), and $\binom{n}{k}$ subsets satisfying (ii).

Equation (1) gives rise to **Pascal's Triangle**, which gives $\binom{n}{k}$ as the k^{th} entry of the n^{th} row (where the numbering starts with $n = 0$ and $k = 0$). Each number is the sum of the two above it:

$$\begin{array}{c}
\binom{0}{0} \\
\binom{1}{0}\ \binom{1}{1} \\
\binom{2}{0}\ \binom{2}{1}\ \binom{2}{2} \\
\binom{3}{0}\ \binom{3}{1}\ \binom{3}{2}\ \binom{3}{4} \\
\binom{4}{0}\ \binom{4}{1}\ \binom{4}{2}\ \binom{4}{3}\ \binom{4}{4} \\
\binom{5}{0}\ \binom{5}{1}\ \binom{5}{2}\ \binom{5}{3}\ \binom{5}{4}\ \binom{5}{5} \\
\binom{6}{0}\ \binom{6}{1}\ \binom{6}{2}\ \binom{6}{3}\ \binom{6}{4}\ \binom{6}{5}\ \binom{6}{6} \\
\text{etc.}
\end{array}
=
\begin{array}{c}
1 \\
1\quad 1 \\
1\quad 2\quad 1 \\
1\quad 3\quad 3\quad 1 \\
1\quad 4\quad 6\quad 4\quad 1 \\
1\quad 5\quad 10\quad 10\quad 5\quad 1 \\
1\quad 6\quad 15\quad 20\quad 15\quad 6\quad 1 \\
\text{etc.}
\end{array}$$

Equation (1) also implies (by the telescoping method) that

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^k \binom{n}{k} = (-1)^k \binom{n-1}{k}.$$

2.10.6 Number of all subsets of a set

Given a subset A of a set X, its **indicator function** is defined by

$$\mathbf{1}_A(x) = \begin{cases} 1 & x \in A, \\ 0 & x \notin A. \end{cases}$$

There is a one-to-one correspondence between sets and indicator functions. How many different indicator functions are there? For each element the value can be either 0 or 1, and there are n elements so

there are 2^n distinct subsets of a set of n objects.

2.10.7 And so ...

If we sum the number of sets of size k from 0 to n, we get the total number of subsets, so

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

This is a special case of the following result, which you may remember from high school or Ma 1a.

2.10.2 Binomial Theorem

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Bibliography

[1] T. M. Cover and J. A. Thomas. 2006. *Elements of information theory*, 2d. ed. Hoboken, New Jersey: Wiley–Interscience.

[2] R. M. Gray. 1988. *Probability, random processes, and ergodic properties.* New York: Springer–Verlag.

[3] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[4] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.