Counting and Sets
Class 1, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definitions and notation for sets, intersection, union, complement.
2. Be able to visualize set operations using Venn diagrams.
3. Understand how counting is used computing probabilities.
4. Be able to use the rule of product, inclusion-exclusion principle, permutations and combinations to count the elements in a set.

2 Counting

2.1 Motivating questions

Example 1. A coin is *fair* if it comes up heads or tails with equal probability.
You flip a fair coin three times. What is the probability that exactly one of the flips results in a head?

answer: With three flips, we can easily list the eight possible outcomes:

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

Three of these outcomes have exactly one head:

$$\{TTH, THT, HTT\}$$

Since all outcomes are equally probable, we have

$$P(\text{1 head in 3 flips}) = \frac{\text{number of outcomes with 1 head}}{\text{total number of outcomes}} = \frac{3}{8}.$$

Think: Would listing the outcomes be practical with 10 flips?

A deck of 52 cards has 13 ranks (2, 3, ..., 9, 10, J, Q, K, A) and 4 suits ($\heartsuit, \spadesuit, \diamondsuit, \clubsuit$,). A poker hand consists of 5 cards. A *one-pair* hand consists of two cards having one rank and three cards having three other ranks, e.g., $\{2\heartsuit, 2\spadesuit, 5\heartsuit, 8\clubsuit, \text{K}\diamondsuit\}$

Test your intuition: the probability of a one-pair hand is:
(a) less than 5%
(b) between 5% and 10%
(c) between 10% and 20%

(d) between 20% and 40%
(e) greater than 40%

At this point we can only guess the probability. One of our goals is to learn how to compute it exactly. To start, we note that since every set of five cards is equally probable, we can compute the probability of a one-pair hand as

$$P(\text{one-pair}) = \frac{\text{number of one-pair hands}}{\text{total number of hands}}$$

So, to find the exact probability, we need to count the number of elements in each of these sets. And we have to be clever about it, because there are too many elements to simply list them all. We will come back to this problem after we have learned some counting techniques.

Several times already we have noted that all the possible outcomes were equally probable and used this to find a probability by counting. Let's state this carefully in the following principle.

Principle: Suppose there are n possible outcomes for an experiment and each is equally probable. If there are k desirable outcomes then the probability of a desirable outcome is k/n. Of course we could replace the word desirable by any other descriptor: undesirable, funny, interesting, remunerative, ...

Concept question: Can you think of a scenario where the possible outcomes are not equally probable?

Here's one scenario: on an exam you can get any score from 0 to 100. That's 101 different possible outcomes. Is the probability you get less than 50 equal to 50/101?

2.2 Sets and notation

Our goal is to learn techniques for counting the number of elements of a set, so we start with a brief review of sets. (If this is new to you, please come to office hours).

2.2.1 Definitions

A **set** S is a collection of elements. We use the following notation.

Element: We write $x \in S$ to mean the element x is in the set S.

Subset: We say the set A is a subset of S if all of its elements are in S. We write this as $A \subset S$.

Complement:: The complement of A in S is the set of elements of S that are **not** in A. We write this as A^c or $S - A$.

Union: The union of A and B is the set of all elements in A **or** B (or both). We write this as $A \cup B$.

Intersection: The intersection of A and B is the set of all elements in both A **and** B. We write this as $A \cap B$.

Empty set: The empty set is the set with no elements. We denote it $\emptyset$.

Disjoint: A and B are **disjoint** if they have no common elements. That is, if $A \cap B = \emptyset$.

Difference: The difference of A and B is the set of elements in A that are not in B. We write this as $A - B$.

Let's illustrate these operations with a simple example.

Example 2. Start with a set of 10 animals

$$S = \{\text{Antelope, Bee, Cat, Dog, Elephant, Frog, Gnat, Hyena, Iguana, Jaguar}\}.$$

Consider two subsets:

M = the animal is a mammal $= \{\text{Antelope, Cat, Dog, Elephant, Hyena, Jaguar}\}$
W = the animal lives in the wild $= \{\text{Antelope, Bee, Elephant, Frog, Gnat, Hyena, Iguana, Jaguar}\}$.

Our goal here is to look at different set operations.
Intersection: $M \cap W$ contains all wild mammals: $M \cap W = \{\text{Antelope, Elephant, Hyena, Jaguar}\}$.
Union: $M \cup W$ contains all animals that are mammals or wild (or both).
$M \cup W = \{\text{Antelope, Bee, Cat, Dog, Elephant, Frog, Gnat, Hyena, Iguana, Jaguar}\}$.
Complement: M^c means everything that is *not* in M, i.e. not a mammal. $M^c = \{\text{Bee, Frog, Gnat, Iguana}\}$.
Difference: $M - W$ means everything that's in M and not in W.
So, $M - W = \{\text{Cat, Dog}\}$.

There are often many ways to get the same set, e.g. $M^c = S - M$, $M - W = M \cap L^c$.

The relationship between union, intersection, and complement is given by DeMorgan's laws:

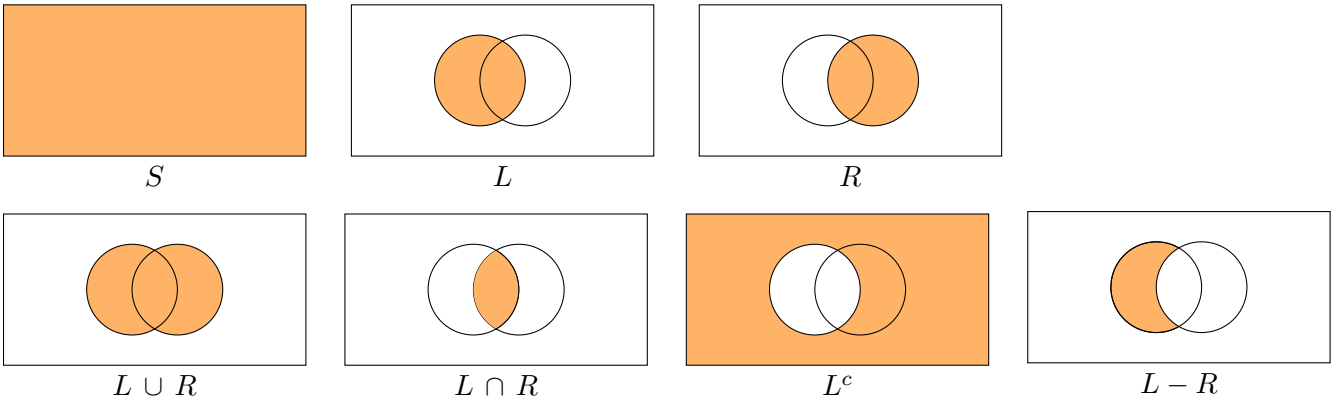
$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

In words the first law says everything not in (A or B) is the same set as everything that's (not in A) and (not in B). The second law is similar.

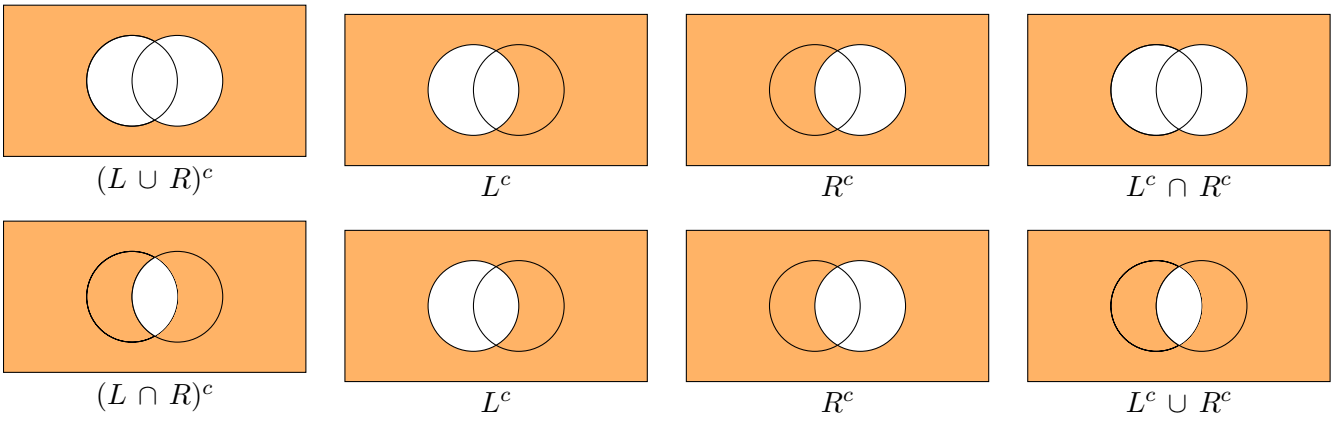
2.2.2 Venn Diagrams

Venn diagrams offer an easy way to visualize set operations.

In all the figures S is the region inside the large rectangle, L is the region inside the left circle and R is the region inside the right circle. The shaded region shows the set indicated underneath each figure.



Proof of DeMorgan's Laws



Example 3. Verify DeMorgan's laws for the subsets $A = \{1, 2, 3\}$ and $B = \{3, 4\}$ of the set $S = \{1, 2, 3, 4, 5\}$.

answer: For each law we just work through both sides of the equation and show they are the same.
1. $(A \cup B)^c = A^c \cap B^c$:
Right hand side: $A \cup B = \{1, 2, 3, 4\} \Rightarrow (A \cup B)^c = \{5\}$.
Left hand side: $A^c = \{4, 5\}$, $B^c = \{1, 2, 5\} \Rightarrow A^c \cap B^c = \{5\}$.
The two sides are equal. QED

2. $(A \cap B)^c = A^c \cup B^c$:
Right hand side: $A \cap B = \{3\} \Rightarrow (A \cap B)^c = \{1, 2, 4, 5\}$.
Left hand side: $A^c = \{4, 5\}$, $B^c = \{1, 2, 5\} \Rightarrow A^c \cup B^c = \{1, 2, 4, 5\}$.
The two sides are equal. QED

Think: Draw and label a Venn diagram with A the set of Brain and Cognitive Science majors and B the set of sophomores. Shade the region illustrating the first law. Can you express the first law in this case as a non-technical English sentence?

2.2.3 Products of sets

The product of sets S and T is the set of ordered pairs:

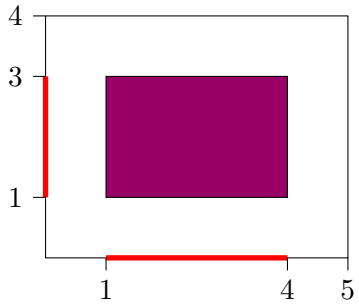
$$S \times T = \{(s, t) \mid s \in S, t \in T\}.$$

In words the right-hand side reads "the set of ordered pairs (s, t) such that s is in S and t is in T.

The following diagrams show two examples of the set product.

×	1	2	3	4
1	(1,1)	(1,2)	(1,3)	(1,4)
2	(2,1)	(2,2)	(2,3)	(2,4)
3	(3,1)	(3,2)	(3,3)	(3,4)

$\{1,2,3\} \times \{1,2,3,4\}$



$[1,4] \times [1,3] \subset [0,5] \times [0,4]$

The right-hand figure also illustrates that if $A \subset S$ and $B \subset T$ then $A \times B \subset S \times T$.

2.3 Counting

If S is finite, we use $|S|$ or $\#S$ to denote the number of elements of S.

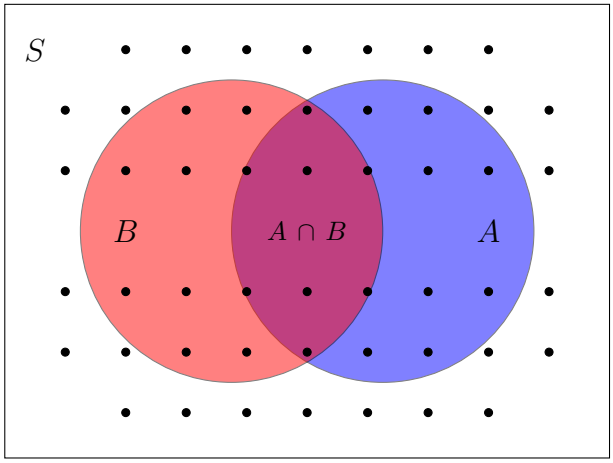
Two useful counting principles are the *inclusion-exclusion principle* and the *rule of product*.

2.3.1 Inclusion-exclusion principle

The inclusion-exclusion principle says

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

We can illustrate this with a Venn diagram. S is all the dots, A is the dots in the blue circle, and B is the dots in the red circle.



$|A|$ is the number of dots in A and likewise for the other sets. The figure shows that $|A|+|B|$ *double*-counts $|A \cap B|$, which is why $|A \cap B|$ is subtracted off in the inclusion-exclusion formula.

Example 4. In a band of singers and guitarists, seven people sing, four play the guitar, and two do both. How big is the band?

answer: Let S be the set singers and G be the set guitar players. The inclusion-exclusion principle says

$$\text{size of band} = |S \cup G| = |S| + |G| - |S \cap G| = 7 + 4 - 2 = 9.$$

2.3.2 Rule of Product

The Rule of Product says:

> If there are n ways to perform action 1 and then by m ways to perform action 2, then there are $n \cdot m$ ways to perform action 1 followed by action 2.

We will also call this the multiplication rule.

Example 5. If you have 3 shirts and 4 pants then you can make $3 \cdot 4 = 12$ outfits.

Think: An extremely important point is that the rule of product holds even if the ways to perform action 2 depend on action 1, as long as the *number* of ways to perform action 2 is independent of action 1. To illustrate this:

Example 6. There are 5 competitors in the 100m final at the Olympics. In how many ways can the gold, silver, and bronze medals be awarded?

answer: There are 5 ways to award the gold. Once that is awarded there are 4 ways to award the silver and then 3 ways to award the bronze: answer $5 \cdot 4 \cdot 3 = 60$ ways.

Note that the choice of gold medalist affects who can win the silver, but the number of possible silver medalists is always four.

2.4 Permutations and combinations

2.4.1 Permutations

A permutation of a set is a particular ordering of its elements. For example, the set $\{a, b, c\}$ has six permutations: $abc, acb, bac, bca, cab, cba$. We found the number of permutations by listing them all. We could also have found the number of permutations by using the rule of product. That is, there are 3 ways to pick the first element, then 2 ways for the second, and 1 for the first. This gives a total of $3 \cdot 2 \cdot 1 = 6$ permutations.

In general, the rule of product tells us that the number of permutations of a set of k elements is

$$k! = k \cdot (k-1) \cdots 3 \cdot 2 \cdot 1.$$

We also talk about the permutations of k things out of a set of n things. We show what this means with an example.

Example 7. List all the permutations of 3 elements out of the set $\{a, b, c, d\}$. **answer:** This is a longer list,

abc	acb	bac	bca	cab	cba
abd	adb	bad	bda	dab	dba
acd	adc	cad	cda	dac	dca
bcd	bdc	cbd	cdb	dbc	dcb

Note that *abc* and *acb* count as distinct permutations. That is, for permutations the order matters.

There are 24 permutations. Note that the rule or product would have told us there are $4 \cdot 3 \cdot 2 = 24$ permutations without bothering to list them all.

2.4.2 Combinations

In contrast to permutations, in combinations order does *not* matter: permutations are lists and combinations are sets. We show what we mean with an example

Example 8. List all the combinations of 3 elements out of the set $\{a, b, c, d\}$.

answer: Such a combination is a collection of 3 elements without regard to order. So, *abc* and *cab* both represent the same combination. We can list all the combinations by listing all the subsets of exactly 3 elements.

$$\{a, b, c\} \quad \{a, b, d\} \quad \{a, c, d\} \quad \{b, c, d\}$$

There are only 4 combinations. Contrast this with the 24 permutations in the previous example. The factor of 6 comes because every combination of 3 things can be written in 6 different orders.

2.4.3 Formulas

We'll use the following notations.
${}_nP_k$ = number of permutations (lists) of k distinct elements from a set of size n

${}_nC_k = \binom{n}{k}$ = number of combinations (subsets) of k elements from a set of size n

We emphasise that by the number of combinations of k elements we mean the number of subsets of size k.

These have the following notation and formulas:

$$\text{Permutations:} \quad {}_nP_k = \frac{n!}{(n-k)!} = n(n-1)\cdots(n-k+1)$$

$$\text{Combinations:} \quad {}_nC_k = \frac{n!}{k!(n-k)!} = \frac{{}_nP_k}{k!}$$

The notation ${}_nC_k$ is read "n choose k". The formula for ${}_nP_k$ follows from the rule of product. It also implies the formula for ${}_nC_k$ because a subset of size k can be ordered in $k!$ ways.

We can illustrate the relation between permutations and combinations by lining up the results of the previous two examples.

abc	*acb*	*bac*	*bca*	*cab*	*cba*	$\{a, b, c\}$
abd	*adb*	*bad*	*bda*	*dab*	*dba*	$\{a, b, d\}$
acd	*adc*	*cad*	*cda*	*dac*	*dca*	$\{a, c, d\}$
bcd	*bdc*	*cbd*	*cdb*	*dbc*	*dcb*	$\{b, c, d\}$
Permutations: ${}_4P_3$						Combinations: ${}_4C_3$

Notice that each row in the permutations list consists of all 3! permutations of the corresponding set in the combinations list.

2.4.4 Examples

Example 9. Count the following:
(i) The number of ways to choose 2 out of 4 things (order does not matter).
(ii) The number of ways to list 2 out of 4 things.
(iii) The number of ways to choose 3 out of 10 things.

answer: (i) This is asking for combinations: $\binom{4}{2} = \frac{4!}{2!\,2!} = 6$.
(ii) This is asking for permuations: ${}_4P_2 = \frac{4!}{2!} = 12$.
(iii) This is asking for combinations: $\binom{10}{3} = \frac{10!}{3!\,7!} = \frac{10\cdot 9\cdot 8}{3\cdot 2\cdot 1} = 120$.

Example 10. (i) Count the number of ways to get 3 heads in a sequence of 10 flips of a coin.
(ii) If the coin is fair, what is the probability of exactly 3 heads in 10 flips.

answer: (i) This asks for the number sequences of 10 flips (heads or tails) with exactly 3 heads. That is, we have to choose exactly 3 out of 10 flips to be heads. This is the same question as in the previous example.

$$\binom{10}{3} = \frac{10!}{3!\,7!} = \frac{10\cdot 9\cdot 8}{3\cdot 2\cdot 1} = 120.$$

(ii) Each flip has 2 possible outcomes (heads or tails). So the rule of product says there are $2^{10} = 1024$ sequences of 10 flips. Since the coin is fair each sequence is equally probable. So the probability of 3 heads is

$$\frac{120}{1024} = .117\,.$$

Probability: Terminology and Examples
Class 2, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definitions of sample space, event and probability function.
2. Be able to organize a scenario with randomness into an experiment and sample space.
3. Be able to make basic computations using a probability function.

2 Terminology

2.1 Probability cast list

- Experiment: a repeatable procedure with well-defined possible outcomes.
- Sample space: the set of all possible outcomes. We usually denote the sample space by Ω, sometimes by S.
- Event: a subset of the sample space.
- Probability function: a function giving the probability for each outcome.

Later in the course we will learn about

- Probability density: a continuous distribution of probabilities.
- Random variable: a random numerical outcome.

2.2 Simple examples

Example 1. Toss a fair coin.
Experiment: toss the coin, report if it lands heads or tails.
Sample space: $\Omega = \{H, T\}$.
Probability function: $P(H) = .5, \quad P(T) = .5$.

Example 2. Toss a fair coin 3 times.
Experiment: toss the coin 3 times, list the results.
Sample space: $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.
Probability function: Each outcome is equally likely with probability 1/8.

For small sample spaces we can put the set of outcomes and probabilities into a probability table.

Outcomes	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Probability	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Example 3. Measure the mass of a proton
Experiment: follow some defined procedure to measure the mass and report the result.
Sample space: $\Omega = [0, \infty)$, i.e. in principle we can get any positive value.
Probability function: since there is a continuum of possible outcomes there is no probability function. Instead we need to use a *probability density*, which we will learn about later in the course.

Example 4. Taxis (An infinite discrete sample space)
Experiment: count the number of taxis that pass 77 Mass. Ave during an 18.05 class.
Sample space: $\Omega = \{0, 1, 2, 3, 4, \dots\}$.
This is often modeled with the following probability function known as the Poisson distribution. (Do not worry about mastering the Poisson distribution just yet):

$$P(k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!},$$

where λ is the average number of taxis. We can put this in a table:

Outcome	0	1	2	3	...	k	...
Probability	$\mathrm{e}^{-\lambda}$	$\mathrm{e}^{-\lambda}\lambda$	$\mathrm{e}^{-\lambda}\lambda^2/2$	$\mathrm{e}^{-\lambda}\lambda^3/3!$	...	$\mathrm{e}^{-\lambda}\lambda^k/k!$	...

Question: Accepting that this is a valid probability function, what is $\displaystyle\sum_{k=0}^{\infty}\mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}$?

answer: This is the total probability of all possible outcomes, so the sum equals 1. (Note, this also follows from the Taylor series $\mathrm{e}^{\lambda} = \displaystyle\sum_{n=0}^{\infty}\frac{\lambda^n}{n!}$.)

In a given setup there can be more than one reasonable choice of sample space. Here is a simple example.

Example 5. Two dice (Choice of sample space)
Suppose you roll one die. Then the sample space and probability function are

Outcome	1	2	3	4	5	6
Probability:	1/6	1/6	1/6	1/6	1/6	1/6

Now suppose you roll two dice. What should be the sample space? Here are two options.
1. Record the pair of numbers showing on the dice (first die, second die).
2. Record the sum of the numbers on the dice. In this case there are 11 outcomes $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. These outcomes are not all equally likely.

As above, we can put this information in tables. For the first case, the sample space is the product of the sample spaces for each die

$$\{(1,1),\ (2,1),\ (3,1), \dots (6,6)\}$$

Each of the 36 outcomes is equally likely. (Why 36 outcomes?) For the probability function we will make a two dimensional table with the rows corresponding to the number on the first die, the columns the number on the second die and the entries the probability.

Die 1 \ Die 2	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Two dice in a two dimensional table

In the second case we can present outcomes and probabilities in our usual table.

outcome	2	3	4	5	6	7	8	9	10	11	12
probability	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

The sum of two dice

Think: What is the relationship between the two probability tables above?

We will see that the best choice of sample space depends on the context. For now, simply note that given the outcome as a pair of numbers it is easy to find the sum.

Note. Listing the experiment, sample space and probability function is a good way to start working systematically with probability. It can help you avoid some of the common pitfalls in the subject.

Events.

An event is a collection of outcomes, i.e. an event is a subset of the sample space Ω. This sounds odd, but it actually corresponds to the common meaning of the word.

Example 6. Using the setup in Example 2 we would describe the event that you get exactly two heads in words by $E =$ 'exactly 2 heads'. Written as a subset this becomes

$$E = \{HHT, HTH, THH\}.$$

You should get comfortable moving between describing events in words and as subsets of the sample space.

The probability of an event E is computed by adding up the probabilities of all of the outcomes in E. In this example each outcome has probability 1/8, so we have $P(E) = 3/8$.

2.3 Definition of a discrete sample space

Definition. A discrete sample space is one that is listable, it can be either finite or infinite.

Examples. {H, T}, {1, 2, 3}, {1, 2, 3, 4, ...}, {2, 3, 5, 7, 11, 13, 17, ...} are all discrete sets. The first two are finite and the last two are infinite.

Example. The interval $0 \le x \le 1$ is *not* discrete, rather it is *continuous*. We will deal with continuous sample spaces in a few days.

2.4 The probability function

So far we've been using a casual definition of the probability function. Let's give a more precise one.

Careful definition of the probability function.
For a discrete sample space S a probability function P assigns to each outcome ω a number $P(\omega)$ called the probability of ω. P must satisfy two rules:

- Rule 1. $0 \le P(\omega) \le 1$ (probabilities are between 0 and 1).
- Rule 2. The sum of the probabilities of all possible outcomes is 1 (something must occur)

In symbols Rule 2 says: if $S = \{\omega_1, \omega_2, \dots, \omega_n\}$ then $P(\omega_1) + P(\omega_2) + \dots + P(\omega_n) = 1$. Or, using summation notation: $\sum_{j=1}^{n} P(\omega_j) = 1$.

The probability of an event E is the sum of the probabilities of all the outcomes in E. That is,

$$P(E) = \sum_{\omega \in E} P(\omega).$$

Think: Check Rules 1 and 2 on Examples 1 and 2 above.

Example 7. Flip until heads (A classic example)
Suppose we have a coin with probability p of heads and we have the following scenario.

Experiment: Toss the coin until the first heads. Report the number of tosses.
Sample space: $\Omega = \{1, 2, 3, \dots\}$.
Probability function: $P(n) = (1-p)^{n-1}p$.

Challenge 1: show the sum of all the probabilities equals 1 (hint: geometric series).

Challenge 2: justify the formula for $P(n)$ (we will do this soon).

Stopping problems. The previous toy example is an uncluttered version of a general class of problems called stopping rule problems. A stopping rule is a rule that tells you when to end a certain process. In the toy example above the process was flipping a coin and we stopped after the first heads. A more practical example is a rule for ending a series of medical treatments. Such a rule could depend on how well the treatments are working, how the patient is tolerating them and the probability that the treatments would continue to be effective. One could ask about the probability of stopping within a certain number of treatments or the average number of treatments you should expect before stopping.

3 Some rules of probability

For events A, L and R contained in a sample space Ω.

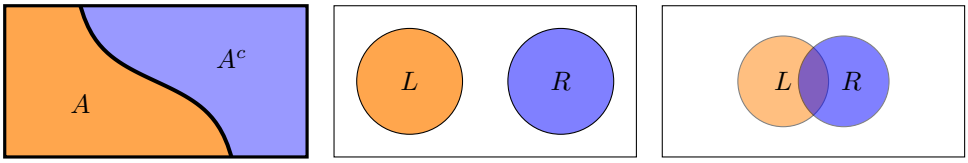
Rule 1. $P(A^c) = 1 - P(A)$.

Rule 2. If L and R are disjoint then $P(L \cup R) = P(L) + P(R)$.

Rule 3. If L and R are not disjoint, we have the **inclusion-exclusion principle**:

$$P(L \cup R) = P(L) + P(R) - P(L \cap R)$$

We visualize these rules using Venn diagrams.



$\Omega = A \cup A^c$, no overlap $L \cup R$, no overlap $L \cup R$, overlap $= L \cap R$

We can also justify them logically.

Rule 1: A and A^c split Ω into two non-overlapping regions. Since the total probability $P(\Omega) = 1$ this rule says that the probabiity of A and the probability of 'not A' are complementary, i.e. sum to 1.

Rule 2: L and R split $L \cup R$ into two non-overlapping regions. So the probability of $L \cup R$ is is split between $P(L)$ and $P(R)$

Rule 3: In the sum $P(L) + P(R)$ the overlap $P(L \cap R)$ gets counted twice. So $P(L) + P(R) - P(L \cap R)$ counts everything in the union exactly once.

Think: Rule 2 is a special case of Rule 3.

For the following examples suppose we have an experiment that produces a random integer between 1 and 20. The probabilities are not necessarily uniform, i.e., not necessarily the same for each outcome.

Example 8. If the probability of an even number is .6 what is the probability of an odd number?

answer: Since being odd is complementary to being even, the probability of being odd is 1-.6 = .4.

Let's redo this example a bit more formally, so you see how it's done. First, so we can refer to it, let's name the random integer X. Let's also name the event 'X is even' as A. Then the event 'X is odd' is A^c. We are given that $P(A) = .6$. Therefore $P(A^c) = 1 - .6 = \boxed{.4}$.

Example 9. Consider the 2 events, A: 'X is a multiple of 2'; B: 'X is odd and less than 10'. Suppose $P(A) = .6$ and $P(B) = .25$.
(i) What is $A \cap B$?
(ii) What is the probability of $A \cup B$?

answer: (i) Since all numbers in A are even and all numbers in B are odd, these events are disjoint. That is, $\boxed{A \cap B = \emptyset.}$

(ii) Since A and B are disjoint $\boxed{P(A \cup B) = P(A) + P(B) = .85.}$

Example 10. Let A, B and C be the events X is a multiple of 2, 3 and 6 respectively. If $P(A) = .6$, $P(B) = .3$ and $P(C) = .2$ what is $P(A$ or $B)$?

answer: Note two things. First we used the word 'or' which means union: 'A or B' = $A \cup B$. Second, an integer is divisible by 6 if and only if it is divisible by both 2 and 3.

This translates into $C = A \cap B$. So the inclusion-exclusion principle says

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = .6 + .3 - .2 = \boxed{.7}.$$

Conditional Probability, Independence and Bayes' Theorem
Class 3, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definitions of conditional probability and independence of events.
2. Be able to compute conditional probability directly from the definition.
3. Be able to use the multiplication rule to compute the total probability of an event.
4. Be able to check if two events are independent.
5. Be able to use Bayes' formula to 'invert' conditional probabilities.
6. Be able to organize the computation of conditional probabilities using trees and tables.
7. Understand the base rate fallacy thoroughly.

2 Conditional Probability

Conditional probability answers the question 'how does the probability of an event change if we have extra information'. We'll illustrate with an example.

Example 1. Toss a fair coin 3 times.
(a) What is the probability of 3 heads?
answer: Sample space $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.
All outcomes are equally likely, so $P(3 \text{ heads}) = 1/8$.

(b) Suppose we are told that the first toss was heads. Given this information how should we compute the probability of 3 heads?
answer: We have a new (reduced) sample space: $\Omega' = \{HHH, HHT, HTH, HTT\}$.
All outcomes are equally likely, so

$$P(3 \text{ heads given that the first toss is heads}) = 1/4.$$

This is called conditional probability, since it takes into account additional conditions. To develop the notation, we rephrase (b) in terms of *events*.

Rephrased (b) Let A be the event 'all three tosses are heads' $= \{HHH\}$.
Let B be the event 'the first toss is heads' $= \{HHH, HHT, HTH, HTT\}$.

The conditional probability of A knowing that B occurred is written

$$P(A|B)$$

This is read as

'the conditional probability of A given B'

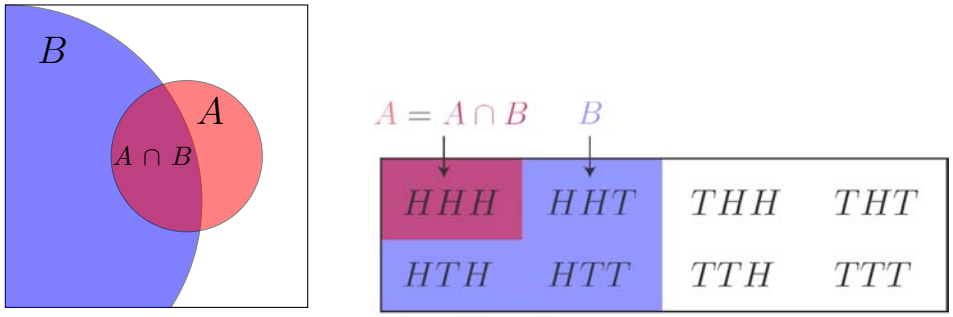
or

'the probability of A conditioned on B'

or simply

'the probability of A given B'.

We can visualize conditional probability as follows. Think of $P(A)$ as the proportion of the area of the *whole* sample space taken up by A. For $P(A|B)$ we restrict our attention to B. That is, $P(A|B)$ is the proportion of area of B taken up by A, i.e. $P(A \cap B)/P(B)$.



Conditional probability: Abstract visualization and coin example

Note, $A \subset B$ in the right-hand figure, so there are only two colors shown.

The formal definition of conditional probability catches the gist of the above example and visualization.

Formal definition of conditional probability
Let A and B be events. We define the conditional probability of A given B as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ provided } P(B) \neq 0. \tag{1}$$

Let's redo the coin tossing example using the definition in Equation (1). Recall $A =$ '3 heads' and $B =$ 'first toss is heads'. We have $P(A) = 1/8$ and $P(B) = 1/2$. Since $A \cap B = A$, we also have $P(A \cap B) = 1/8$. Now according to (1), $P(A|B) = \frac{1/8}{1/2} = 1/4$, which agrees with our answer in Example 1b.

3 Multiplication Rule

The following formula is called the multiplication rule.

$$P(A \cap B) = P(A|B) \cdot P(B). \tag{2}$$

This is simply a rewriting of the definition in Equation (1) of conditional probability. We will see that our use of the multiplication rule is very similar to our use of the rule of product in counting. In fact, the multiplication rule is just a souped up version of the rule of product.

We start with a simple example where we can check all the probabilities directly by counting.

Example 2. Draw two cards from a deck. Define the events: $S_1 =$ 'first card is a spade' and $S_2 =$ 'second card is a spade'. What is the $P(S_2|S_1)$?

answer: We can do this directly by counting: if the first card is a spade then of the 51 cards remaining, 12 are spades.

$$P(S_2|S_1) = 12/51.$$

Now, let's recompute this using formula (1). We have to compute $P(S_1)$, $P(S_2)$ and $P(S_1 \cap S_2)$: We know that $P(S_1) = 1/4$ because there are 52 equally likely ways to draw the first card and 13 of them are spades. The same logic says that there are 52 equally likely ways the second card can be drawn, so $P(S_2) = 1/4$.

Aside: The probability $P(S_2) = 1/4$ may seem surprising since the value of first card certainly affects the probabilities for the second card. However, if we look at *all* possible two card sequences we will see that every card in the deck has equal probability of being the second card. Since 13 of the 52 cards are spades we get $P(S_2) = 13/52 = 1/4$. Another way to say this is: if we are not given value of the first card then we have to consider all possibilities for the second card.

Continuing, we see that

$$P(S_1 \cap S_2) = \frac{13 \cdot 12}{52 \cdot 51} = 3/51.$$

This was found by counting the number of ways to draw a spade followed by a second spade and dividing by the number of ways to draw any card followed by any other card). Now, using (1) we get

$$P(S_2|S_1) = \frac{P(S_2 \cap S_1)}{P(S_1)} = \frac{3/51}{1/4} = 12/51.$$

Finally, we verify the multiplication rule by computing both sides of (2).

$$P(S_1 \cap S_2) = \frac{13 \cdot 12}{52 \cdot 51} = \frac{3}{51} \qquad \text{and} \qquad P(S_2|S_1) \cdot P(S_1) = \frac{12}{51} \cdot \frac{1}{4} = \frac{3}{51}. \qquad \text{QED}$$

Think: For S_1 and S_2 in the previous example, what is $P(S_2|S_1^c)$?

4 Law of Total Probability

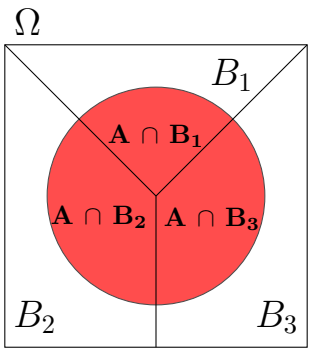
The law of total probability will allow us to use the multiplication rule to find probabilities in more interesting examples. It involves a lot of notation, but the idea is fairly simple. We state the law when the sample space is divided into 3 pieces. It is a simple matter to extend the rule when there are more than 3 pieces.

Law of Total Probability

Suppose the sample space Ω is divided into 3 disjoint events B_1, B_2, B_3 (see the figure below). Then for any event A:

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \\ P(A) &= P(A|B_1)\,P(B_1) + P(A|B_2)\,P(B_2) + P(A|B_3)\,P(B_3) \end{aligned} \tag{3}$$

The top equation says 'if A is divided into 3 pieces then $P(A)$ is the sum of the probabilities of the pieces'. The bottom equation (3) is called the law of total probability. It is just a rewriting of the top equation using the multiplication rule.



The sample space Ω and the event A are each divided into 3 disjoint pieces.

The law holds if we divide Ω into any number of events, so long as they are *disjoint* and *cover* all of Ω. Such a division is often called a *partition* of Ω.

Our first example will be one where we already know the answer and can verify the law.

Example 3. An urn contains 5 red balls and 2 green balls. Two balls are drawn one after the other. What is the probability that the second ball is red?

answer: The sample space is Ω = {rr, rg, gr, gg}.
Let R_1 be the event 'the first ball is red', G_1 = 'first ball is green', R_2 = 'second ball is red', G_2 = 'second ball is green'. We are asked to find $P(R_2)$.

The fast way to compute this is just like $P(S_2)$ in the card example above. Every ball is equally likely to be the second ball. Since 5 out of 7 balls are red, $P(R_2) = 5/7$.

Let's compute this same value using the law of total probability (3). First, we'll find the conditional probabilities. This is a simple counting exercise.

$$P(R_2|R_1) = 4/6, \qquad P(R_2|G_1) = 5/6.$$

Since R_1 and G_1 partition Ω the law of total probability says

$$\begin{aligned} P(R_2) &= P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) \qquad (4) \\ &= \frac{4}{6}\cdot\frac{5}{7} + \frac{5}{6}\cdot\frac{2}{7} \\ &= \frac{30}{42} = \frac{5}{7}. \end{aligned}$$

Probability urns

The example above used probability urns. Their use goes back to the beginning of the subject and we would be remiss not to introduce them. This toy model is very useful. We quote from Wikipedia: http://en.wikipedia.org/wiki/Urn_problem

> In probability and statistics, an urn problem is an idealized mental exercise in which some objects of real interest (such as atoms, people, cars, etc.) are represented as colored balls in an urn or other container. One pretends to draw (remove) one or more balls from the urn; the goal is to determine the probability of drawing one color or another, or some other properties. A key parameter is whether each ball is returned to the urn after each draw.

It doesn't take much to make an example where (3) is really the best way to compute the probability. Here is a game with slightly more complicated rules.

Example 4. An urn contains 5 red balls and 2 green balls. A ball is drawn. If it's green a red ball is added to the urn and if it's red a green ball is added to the urn. (The original ball is not returned to the urn.) Then a second ball is drawn. What is the probability the second ball is red?

answer: The law of total probability says that $P(R_2)$ can be computed using the expression in Equation (4). Only the values for the probabilities will change. We have

$$P(R_2|R_1) = 4/7, \qquad P(R_2|G_1) = 6/7.$$

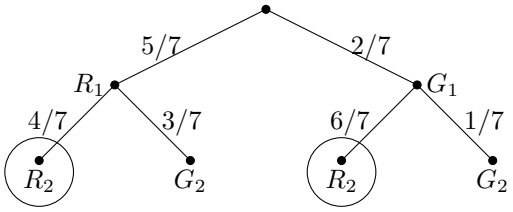
Therefore,

$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) = \frac{4}{7}\cdot\frac{5}{7} + \frac{6}{7}\cdot\frac{2}{7} = \frac{32}{49}.$$

5 Using Trees to Organize the Computation

Trees are a great way to organize computations with conditional probability and the law of total probability. The figures and examples will make clear what we mean by a tree. As with the rule of product, the key is to organize the underlying process into a sequence of actions.

We start by redoing Example 4. The sequence of actions are: first draw ball 1 (and add the appropriate ball to the urn) and then draw ball 2.



You interpret this tree as follows. Each dot is called a node. The tree is organized by levels. The top node (root node) is at level 0. The next layer down is level 1 and so on. Each level shows the outcomes at one stage of the game. Level 1 shows the possible outcomes of the first draw. Level 2 shows the possible outcomes of the second draw starting from each node in level 1.

Probabilities are written along the branches. The probability of R_1 (red on the first draw) is 5/7. It is written along the branch from the root node to the one labeled R_1. At the next level we put in conditional probabilities. The probability along the branch from R_1 to R_2 is $P(R_2|R_1) = 4/7$. It represents the probability of going to node R_2 given that you are already at R_1.

The muliplication rule says that the probability of getting to any node is just the product of the probabilities along the path to get there. For example, the node labeled R_2 at the far left really represents the event $R_1 \cap R_2$ because it comes from the R_1 node. The multiplication rule now says

$$P(R_1 \cap R_2) = P(R_1)\cdot P(R_2|R_1) = \frac{5}{7}\cdot\frac{4}{7},$$

which is exactly multiplying along the path to the node.

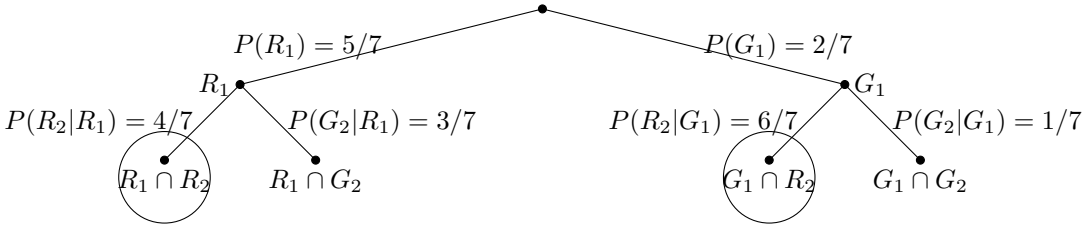
The law of total probability is just the statement that $P(R_2)$ is the sum of the probabilities of all paths leading to R_2 (the two circled nodes in the figure). In this case,

$$P(R_2) = \frac{5}{7} \cdot \frac{4}{7} + \frac{2}{7} \cdot \frac{6}{7} = \frac{32}{49},$$

exactly as in the previous example.

5.1 Shorthand vs. precise trees

The tree given above involves some shorthand. For example, the node marked R_2 at the far left really represents the event $R_1 \cap R_2$, since it ends the path from the root through R_1 to R_2. Here is the same tree with everything labeled precisely. As you can see this tree is more cumbersome to make and use. We usually use the shorthand version of trees. You should make sure you know how to interpret them precisely.



6 Independence

Two events are independent if knowledge that one occurred does not change the probability that the other occurred. Informally, events are independent if they do not influence one another.

Example 5. Toss a coin twice. We expect the outcomes of the two tosses to be independent of one another. In real experiments this always has to be checked. If my coin lands in honey and I don't bother to clean it, then the second toss might be affected by the outcome of the first toss.

More seriously, the independence of experiments can by undermined by the failure to clean or recalibrate equipment between experiments or to isolate supposedly independent observers from each other or a common influence. We've all experienced hearing the same 'fact' from different people. Hearing it from different sources tends to lend it credence until we learn that they all heard it from a common source. That is, our sources were not independent.

Translating the verbal description of independence into symbols gives

$$A \text{ is independent of } B \quad \text{if} \quad P(A|B) = P(A). \tag{5}$$

That is, knowing that B occurred does not change the probability that A occurred. In terms of events as subsets, knowing that the realized outcome is in B does not change the probability that it is in A.

If A and B are independent in the above sense, then the multiplication rule gives $P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$. This justifies the following technical definition of independence.

Formal definition of independence: Two events A and B are independent if

$$P(A \cap B) = P(A) \cdot P(B) \tag{6}$$

This is a nice symmetric definition which makes clear that A is independent of B if and only if B is independent of A. Unlike the equation with conditional probabilities, this definition makes sense even when $P(B) = 0$. In terms of conditional probabilities, we have:
1. If $P(B) \neq 0$ then A and B are independent if and only if $P(A|B) = P(A)$.
2. If $P(A) \neq 0$ then A and B are independent if and only if $P(B|A) = P(B)$.

Independent events commonly arise as different trials in an experiment, as in the following example.

Example 6. Toss a fair coin twice. Let H_1 = 'heads on first toss' and let H_2 = 'heads on second toss'. Are H_1 and H_2 independent?

answer: Since $H_1 \cap H_2$ is the event 'both tosses are heads' we have

$$P(H_1 \cap H_2) = 1/4 = P(H_1)P(H_2).$$

Therefore the events are independent.

We can ask about the independence of any two events, as in the following two examples.

Example 7. Toss a fair coin 3 times. Let H_1 = 'heads on first toss' and A = 'two heads total'. Are H_1 and A independent?

answer: We know that $P(A) = 3/8$. Since this is not 0 we can check if the formula in Equation 5 holds. Now, $H_1 = \{\text{HHH, HHT, HTH, HTT}\}$ contains exactly two outcomes (HHT, HTH) from A, so we have $P(A|H_1) = 2/4$. Since $P(A|H_1) \neq P(A)$ these events are not independent.

Example 8. Draw one card from a standard deck of playing cards. Let's examine the independence of 3 events 'the card is an ace', 'the card is a heart' and 'the card is red'.

Define the events as A = 'ace', H = 'hearts', R = 'red'.

(a) We know that $P(A) = 4/52 = 1/13$, $P(A|H) = 1/13$. Since $P(A) = P(A|H)$ we have that A is independent of H.

(b) $P(A|R) = 2/26 = 1/13$. So A is independent of R. That is, whether the card is an ace is independent of whether it's red.

(c) Finally, what about H and R? Since $P(H) = 1/4$ and $P(H|R) = 1/2$, H and R are not independent. We could also see this the other way around: $P(R) = 1/2$ and $P(R|H) = 1$, so H and R are not independent.

6.1 Paradoxes of Independence

An event A with probability 0 is independent of itself, since in this case both sides of equation (6) are 0. This appears paradoxical because knowledge that A occurred certainly

gives information about whether A occurred. We resolve the paradox by noting that since $P(A) = 0$ the statement 'A occurred' is vacuous.

Think: For what other value(s) of $P(A)$ is A independent of itself?

7 Bayes' Theorem

Bayes' theorem is a pillar of both probability and statistics and it is central to the rest of this course. For two events A and B Bayes' theorem (also called Bayes' rule and Bayes' formula) says

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}. \tag{7}$$

Comments: 1. Bayes' rule tells us how to 'invert' conditional probabilities, i.e. to find $P(B|A)$ from $P(A|B)$.
2. In practice, $P(A)$ is often computed using the law of total probability.

Proof of Bayes' rule
The key point is that $A \cap B$ is symmetric in A and B. So the multiplication rule says

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

Now divide through by $P(A)$ to get Bayes' rule.

A common mistake is to confuse $P(A|B)$ and $P(B|A)$. They can be very different. This is illustrated in the next example.

Example 9. Toss a coin 5 times. Let H_1 = 'first toss is heads' and let H_A = 'all 5 tosses are heads'. Then $P(H_1|H_A) = 1$ but $P(H_A|H_1) = 1/16$.

For practice, let's use Bayes' theorem to compute $P(H_1|H_A)$ using $P(H_A|A_1)$.The terms are $P(H_A|H_1) = 1/16$, $P(H_1) = 1/2$, $P(H_A) = 1/32$. So,

$$P(H_1|H_A) = \frac{P(H_A|H_1)P(H_1)}{P(H_A)} = \frac{(1/16) \cdot (1/2)}{1/32} = 1,$$

which agrees with our previous calculation.

7.1 The Base Rate Fallacy

The base rate fallacy is one of many examples showing that it's easy to confuse the meaning of $P(B|A)$ and $P(A|B)$ when a situation is described in words. This is one of the key examples from probability and it will inform much of our practice and interpretation of statistics. You should strive to understand it thoroughly.

Example 10. The Base Rate Fallacy
Consider a routine screening test for a disease. Suppose the frequency of the disease in the population (base rate) is 0.5%. The test is highly accurate with a 5% false positive rate and a 10% false negative rate.

You take the test and it comes back positive. What is the probability that you have the disease?

answer: We will do the computation three times: using trees, tables and symbols. We'll use the following notation for the relevant events:
$D+$ = 'you have the disease'
$D-$ = 'you do not have the disease
$T+$ = 'you tested positive'
$T-$ = 'you tested negative'.

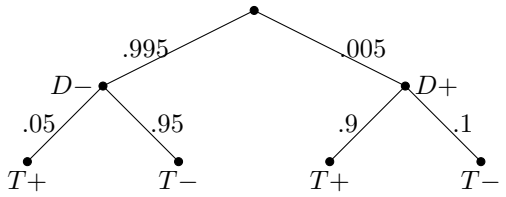
We are given $P(D+) = .005$ and therefore $P(D-) = .995$. The false positive and false negative rates are (by definition) conditional probabilities.

$$P(\text{false positive}) = P(T+|D-) = .05 \quad \text{and} \quad P(\text{false negative}) = P(T-|D+) = .1.$$

The complementary probabilities are known as the true negative and true positive rates:

$$P(T-|D-) = 1 - P(T+|D-) = .95 \quad \text{and} \quad P(T+|D+) = 1 - P(T-|D+) = .9.$$

Trees: All of these probabilities can be displayed quite nicely in a tree.



The question asks for the probability that you have the disease given that you tested positive, i.e. what is the value of $P(D+|T+)$. We aren't given this value, but we do know $P(T+|D+)$, so we can use Bayes' theorem.

$$P(D+|T+) = \frac{P(T+|D+) \cdot P(D+)}{P(T+)}.$$

The two probabilities in the numerator are given. We compute the denominator $P(T+)$ using the law of total probability. Using the tree we just have to sum the probabilities for each of the nodes marked $T+$

$$P(T+) = .995 \times .05 + .005 \times .9 = .05425$$

Thus,

$$P(D+|T+) = \frac{.9 \times .005}{.05425} = 0.082949 \approx 8.3\%.$$

Remarks: This is called the base rate fallacy because the base rate of the disease in the population is so low that the vast majority of the people taking the test are healthy, and even with an accurate test most of the positives will be healthy people. Ask your doctor for his/her guess at the odds.

To summarize the base rate fallacy with specific numbers

95% of all tests are accurate does not imply 95% of positive tests are accurate

We will refer back to this example frequently. It and similar examples are at the heart of many statistical misunderstandings.

Other ways to work Example 10

Tables: Another trick that is useful for computing probabilities is to make a table. Let's redo the previous example using a table built with 10000 total people divided according to the probabilites in this example.

We construct the table as follows. Pick a number, say 10000 people, and place it as the grand total in the lower right. Using $P(D+) = .005$ we compute that 50 out of the 10000 people are sick ($D+$). Likewise 9950 people are healthy ($D-$). At this point the table looks like:

	$D+$	$D-$	total
$T+$			
$T-$			
total	50	9950	10000

Using $P(T+|D+) = .9$ we can compute that the number of sick people who tested positive as 90% of 50 or 45. The other entries are similar. At this point the table looks like the table below on the left. Finally we sum the $T+$ and $T-$ rows to get the completed table on the right.

	$D+$	$D-$	total
$T+$	45	498	
$T-$	5	9452	
total	50	9950	10000

	$D+$	$D-$	total
$T+$	45	498	543
$T-$	5	9452	9457
total	50	9950	10000

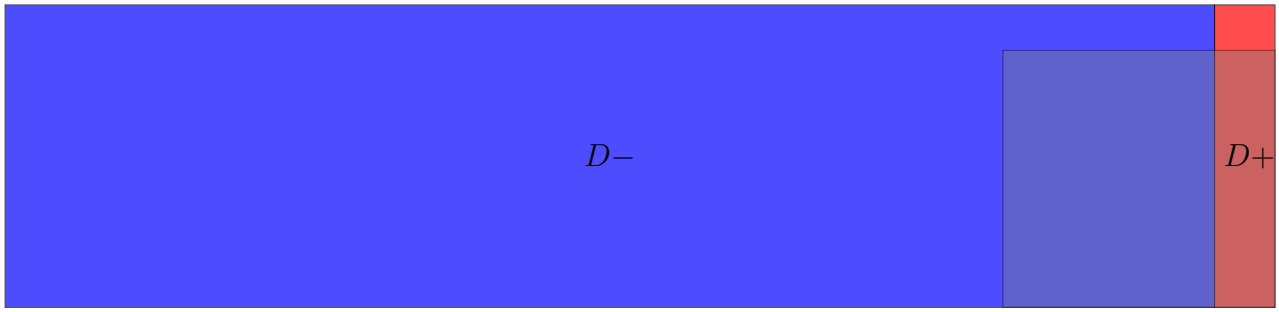
Using the complete table we can compute

$$P(D+|T+) = \frac{|D+ \cap T+|}{|T+|} = \frac{45}{543} = 8.3\%.$$

Symbols: For completeness, we show how the solution looks when written out directly in symbols.

$$\begin{aligned} P(D+|T+) &= \frac{P(T+|D+) \cdot P(D+)}{P(T+)} \\ &= \frac{P(T+|D+) \cdot P(D+)}{P(T+|D+) \cdot P(D+) + P(T+|D-) \cdot P(D-)} \\ &= \frac{.9 \times .005}{.9 \times .005 + .05 \times .995} \\ &= 8.3\% \end{aligned}$$

Visualization: The figure below illustrates the base rate fallacy. The large blue area represents all the healthy people. The much smaller red area represents the sick people. The shaded rectangle represents the the people who test positive. The shaded area covers most of the red area and only a small part of the blue area. Even so, the most of the shaded area is over the blue. That is, most of the positive tests are of healthy people.



7.2 Bayes' rule in 18.05

As we said at the start of this section, Bayes' rule is a pillar of probability and statistics. We have seen that Bayes' rule allows us to 'invert' conditional probabilities. When we learn statistics we will see that the art of statistical inference involves deciding how to proceed when one (or more) of the terms on the right side of Bayes' rule is unknown.

Discrete Random Variables
Class 4, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definition of a discrete random variable.
2. Know the Bernoulli, binomial, and geometric distributions and examples of what they model.
3. Be able to describe the probability mass function and cumulative distribution function using tables and formulas.
4. Be able to construct new random variables from old ones.
5. Know how to compute expected value (mean).

2 Random Variables

This topic is largely about introducing some useful terminology, building on the notions of sample space and probability function. The key words are

1. Random variable
2. Probability mass function (pmf)
3. Cumulative distribution function (cdf)

2.1 Recap

A discrete sample space Ω is a finite or listable set of outcomes $\{\omega_1, \omega_2 \ldots\}$. The probability of an outcome ω is denoted $P(\omega)$.

An event E is a subset of Ω. The probability of an event E is $P(E) = \sum_{\omega \in E} P(\omega)$.

2.2 Random variables as payoff functions

Example 1. A game with 2 dice.

Roll a die twice and record the outcomes as (i, j), where i is the result of the first roll and j the result of the second. We can take the sample space to be

$$\Omega = \{(1,1), (1,2), (1,3), \ldots, (6,6)\} = \{(i,j) \,|\, i, j = 1, \ldots 6\}.$$

The probability function is $P(i, j) = 1/36$.

In this game, you win \$500 if the sum is 7 and lose \$100 otherwise. We give this payoff function the name X and describe it formally by

$$X(i,j) = \begin{cases} 500 & \text{if } i+j=7 \\ -100 & \text{if } i+j \neq 7. \end{cases}$$

Example 2. We can change the game by using a different payoff function. For example

$$Y(i,j) = ij - 10.$$

In this example if you roll $(6,2)$ then you win \$2. If you roll $(2,3)$ then you win -\$4 (i.e., lose \$4).

Question: Which game is the better bet?

answer: We will come back to this once we learn about expectation.

These payoff functions are examples of random variables. A random variable assigns a number to each outcome in a sample space. More formally:

Definition: Let Ω be a sample space. A discrete random variable is a function

$$X : \Omega \to \mathbf{R}$$

that takes a discrete set of values. (Recall that $\mathbf{R}$ stands for the real numbers.)

Why is X called a random variable? It's 'random' because its value depends on a random outcome of an experiment. And we treat X like we would a usual variable: we can add it to other random variables, square it, and so on.

2.3 Events and random variables

For any value a we write $X = a$ to mean the event consisting of all outcomes ω with $X(\omega) = a$.

Example 3. In Example 1 we rolled two dice and X was the random variable

$$X(i,j) = \begin{cases} 500 & \text{if } i+j=7 \\ -100 & \text{if } i+j \neq 7. \end{cases}$$

The event $X = 500$ is the set {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}, i.e. the set of all outcomes that sum to 7. So $P(X = 500) = 1/6$.

We allow a to be any value, even values that X never takes. In Example 1, we could look at the event $X = 1000$. Since X never equals 1000 this is just the empty event (or empty set)

$$\text{'}X = 1000\text{'} \;=\; \{\} \;= \emptyset \qquad P(X = 1000) = 0.$$

2.4 Probability mass function and cumulative distribution function

It gets tiring and hard to read and write $P(X = a)$ for the probability that $X = a$. When we know we're talking about X we will simply write $p(a)$. If we want to make X explicit we will write $p_X(a)$. We spell this out in a definition.

Definition: The probability mass function (pmf) of a discrete random variable is the function $p(a) = P(X = a)$.

Note:
1. We always have $0 \le p(a) \le 1$.
2. We allow a to be any number. If a is a value that X never takes, then $p(a) = 0$.

Example 4. Let Ω be our earlier sample space for rolling 2 dice. Define the random variable M to be the maximum value of the two dice:

$$M(i, j) = \max(i, j).$$

For example, the roll (3,5) has maximum 5, i.e. $M(3, 5) = 5$.

We can describe a random variable by listing its possible values and the probabilities associated to these values. For the above example we have:

value	a:	1	2	3	4	5	6
pmf	$p(a)$:	1/36	3/36	5/36	7/36	9/36	11/36

For example, $p(2) = 3/36$.

Question: What is $p(8)$? **answer:** $p(8) = 0$.

Think: What is the pmf for $Z(i, j) = i + j$? Does it look familiar?

2.5 Events and inequalities

Inequalities with random variables describe events. For example $X \le a$ is the set of all outcomes ω such that $X(w) \le a$.

Example 5. If our sample space is the set of all pairs of (i, j) coming from rolling two dice and $Z(i, j) = i + j$ is the sum of the dice then

$$Z \le 4 = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)\}$$

2.6 The cumulative distribution function (cdf)

Definition: The cumulative distribution function (cdf) of a random variable X is the function F given by $F(a) = P(X \le a)$. We will often shorten this to distribution function.

Note well that the definition of $F(a)$ uses the symbol less than or equal. This will be important for getting your calculations exactly right.

Example. Continuing with the example M, we have

value	a:	1	2	3	4	5	6
pmf	$p(a)$:	1/36	3/36	5/36	7/36	9/36	11/36
cdf	$F(a)$:	1/36	4/36	9/36	16/36	25/36	36/36

$F(a)$ is called the cumulative distribution function because $F(a)$ gives the total probability that accumulates by adding up the probabilities $p(b)$ as b runs from $-\infty$ to a. For example, in the table above, the entry 16/36 in column 4 for the cdf is the sum of the values of the pmf from column 1 to column 4. In notation:

As events: '$M \leq 4$' $= \{1, 2, 3, 4\}$; $\quad F(4) = P(M \leq 4) = 1/36+3/36+5/36+7/36 = 16/36$.

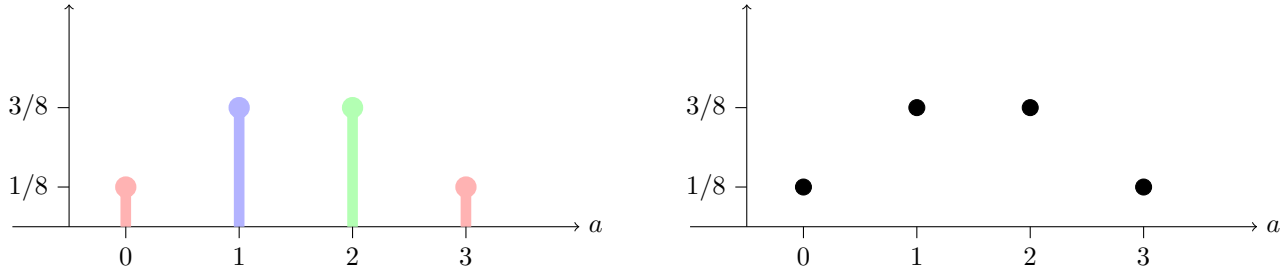
Just like the probability mass function, $F(a)$ is defined for all values a. In the above example, $F(8) = 1$, $F(-2) = 0$, $F(2.5) = 4/36$, and $F(\pi) = 9/36$.

2.7 Graphs of $p(a)$ and $F(a)$

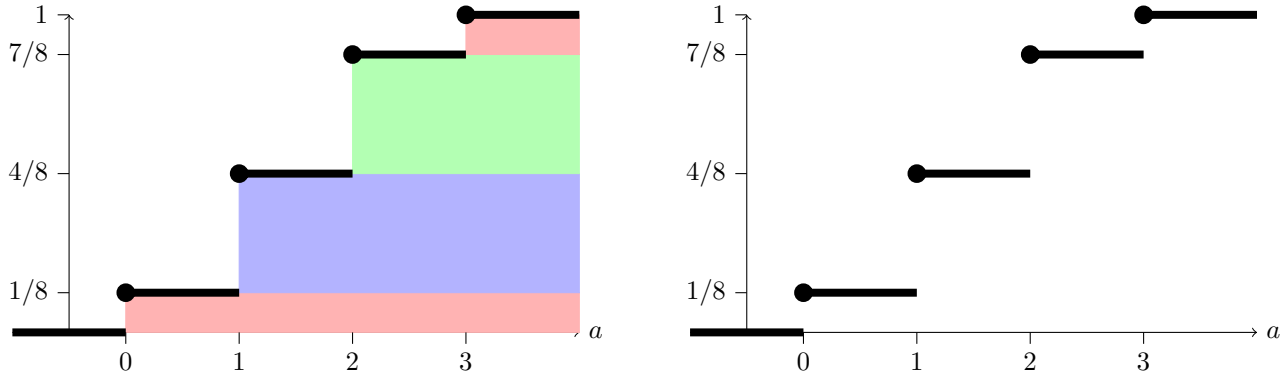
We can visualize the pmf and cdf with graphs. For example, let X be the number of heads in 3 tosses of a fair coin:

value a:	0	1	2	3
pmf $p(a)$:	1/8	3/8	3/8	1/8
cdf $F(a)$:	1/8	4/8	7/8	1

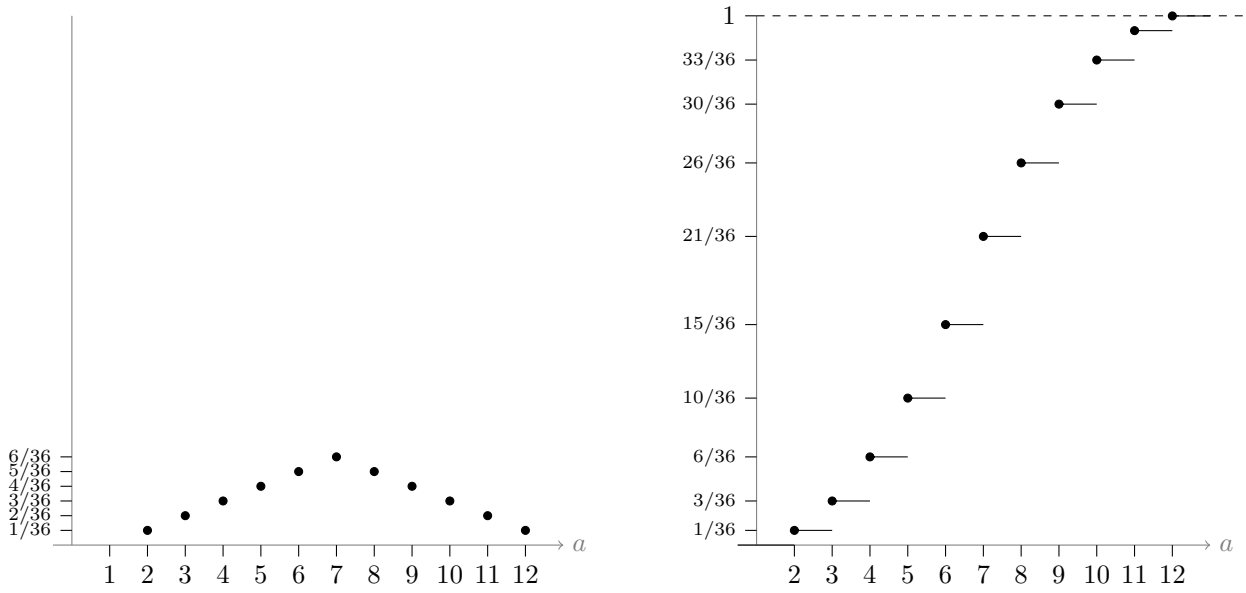
The colored graphs show how the cumulative distribution function is built by accumulating probability as a increases. The black and white graphs are the more standard presentations.



Probability mass function for X



Cumulative distribution function for X



pmf and cdf for the maximum of two dice (Example 4)

Histograms: Later we will see another way to visualize the pmf using histograms. These require some care to do right, so we will wait until we need them.

2.8 Properties of the cdf F

The cdf F of a random variable satisfies several properties:

1. F is non-decreasing. That is, its graph never goes down, or symbolically if $a \le b$ then $F(a) \le F(b)$.
2. $0 \le F(a) \le 1$.
3. $\lim_{a\to\infty} F(a) = 1, \quad \lim_{a\to-\infty} F(a) = 0.$

In words, (1) says the cumulative probability $F(a)$ increases or remains constant as a increases, but never decreases; (2) says the accumulated probability is always between 0 and 1; (3) says that as a gets very large, it becomes more and more certain that $X \le a$ and as a gets very negative it becomes more and more certain that $X > a$.

Think: Why does a cdf satisfy each of these properties?

3 Specific Distributions

3.1 Bernoulli Distributions

Model: The Bernoulli distribution models one trial in an experiment that can result in either success or failure This is the most important distribution is also the simplest. A random variable X has a Bernoulli distribution with parameter p if:

1. X takes the values 0 and 1.

2. $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

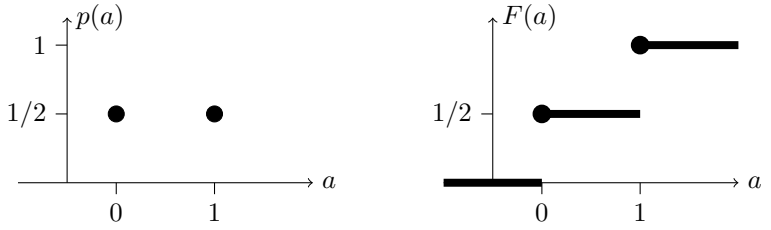
We will write $X \sim$ Bernoulli(p) or Ber(p), which is read "X follows a Bernoulli distribution with parameter p" or "X is drawn from a Bernoulli distribution with parameter p".

A simple model for the Bernoulli distribution is to flip a coin with probability p of heads, with $X = 1$ on heads and $X = 0$ on tails. The general terminology is to say X is 1 on success and 0 on failure, with success and failure defined by the context.

Many decisions can be modeled as a binary choice, such as votes for or against a proposal. If p is the proportion of the voting population that favors the proposal, than the vote of a random individual is modeled by a Bernoulli(p).

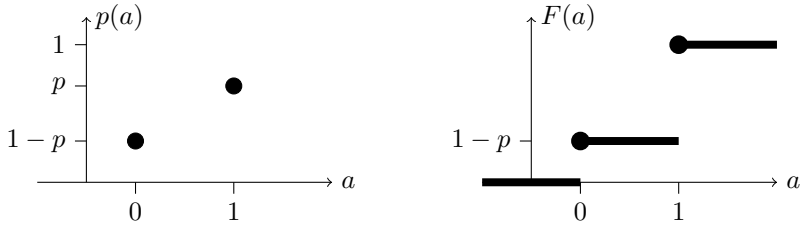
Here are the table and graphs of the pmf and cdf for the Bernoulli(1/2) distribution and below that for the general Bernoulli(p) distribution.

value a:	0	1
pmf $p(a)$:	1/2	1/2
cdf $F(a)$:	1/2	1



Table, pmf and cmf for the Bernoulli(1/2) distribution

values a:	0	1
pmf $p(a)$:	1-p	p
cdf $F(a)$:	1-p	1



Table, pmf and cmf for the Bernoulli(p) distribution

3.2 Binomial Distributions

The binomial distribution Binomial(n,p), or Bin(n,p), models the number of successes in n independent Bernoulli(p) trials.

There is a hierarchy here. A single Bernoulli trial is, say, one toss of a coin. A single binomial trial consists of n Bernoulli trials. For coin flips the sample space for a Bernoulli trial is $\{H, T\}$. The sample space for a binomial trial is all sequences of heads and tails of length n. Likewise a Bernoulli random variable takes values 0 and 1 and a binomial random variables takes values 0, 1, 2, ..., n.

Example 6. Binomial(1,p) is the same as Bernoulli(p).

Example 7. The number of heads in n flips of a coin with probability p of heads follows a Binomial(n, p) distribution.

We describe $X \sim \text{Binomial}(n, p)$ by giving its values and probabilities. For notation we will use k to mean an arbitrary number between 0 and n.

We remind you that 'n choose k $= \binom{n}{k} = {}_nC_k$ is the number of ways to choose k things out of a collection of n things and it has the formula

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}. \tag{1}$$

(It is also called a binomial coefficient.) Here is a table for the pmf of a Binomial(n, k) random variable. We will explain how the binomial coefficients enter the pmf for the binomial distribution after a simple example.

values a:	0	1	2	$\cdots$	k	$\cdots$	n
pmf $p(a)$:	$(1-p)^n$	$\binom{n}{1}p^1(1-p)^{n-1}$	$\binom{n}{2}p^2(1-p)^{n-2}$	$\cdots$	$\binom{n}{k}p^k(1-p)^{n-k}$	$\cdots$	p^n

Example 8. What is the probability of 3 or more heads in 5 tosses of a fair coin?

answer: The binomial coefficients associated with $n = 5$ are

$$\binom{5}{0} = 1, \quad \binom{5}{1} = \frac{5!}{1!\,4!} = \frac{5\cdot4\cdot3\cdot2\cdot1}{4\cdot3\cdot2\cdot1} = 5, \quad \binom{5}{2} = \frac{5!}{2!\,3!} = \frac{5\cdot4\cdot3\cdot2\cdot1}{2\cdot1\cdot3\cdot2\cdot1} = \frac{5\cdot4}{2} = 10,$$

and similarly

$$\binom{5}{3} = 10, \quad \binom{5}{4} = 5, \quad \binom{5}{5} = 1.$$

Using these values we get the following table for $X \sim$ Binomial(5,p).

values a:	0	1	2	3	4	5
pmf $p(a)$:	$(1-p)^5$	$5p(1-p)^4$	$10p^2(1-p)^3$	$10p^3(1-p)^2$	$5p^4(1-p)$	p^5

We were told $p = 1/2$ so

$$P(X \geq 3) = 10\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 + 5\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^5 = \frac{16}{32} = \frac{1}{2}.$$

Think: Why is the value of 1/2 not surprising?

3.3 Explanation of the binomial probabilities

For concreteness, let $n = 5$ and $k = 2$ (the argument for arbitrary n and k is identical.) So $X \sim$ binomial$(5, p)$ and we want to compute $p(2)$. The long way to compute $p(2)$ is to list all the ways to get exactly 2 heads in 5 coin flips and add up their probabilities. The list has 10 entries:

HHTTT, HTHTT, HTTHT, HTTTH, THHTT, THTHT, THTTH, TTHHT, TTHTH, TTTHH

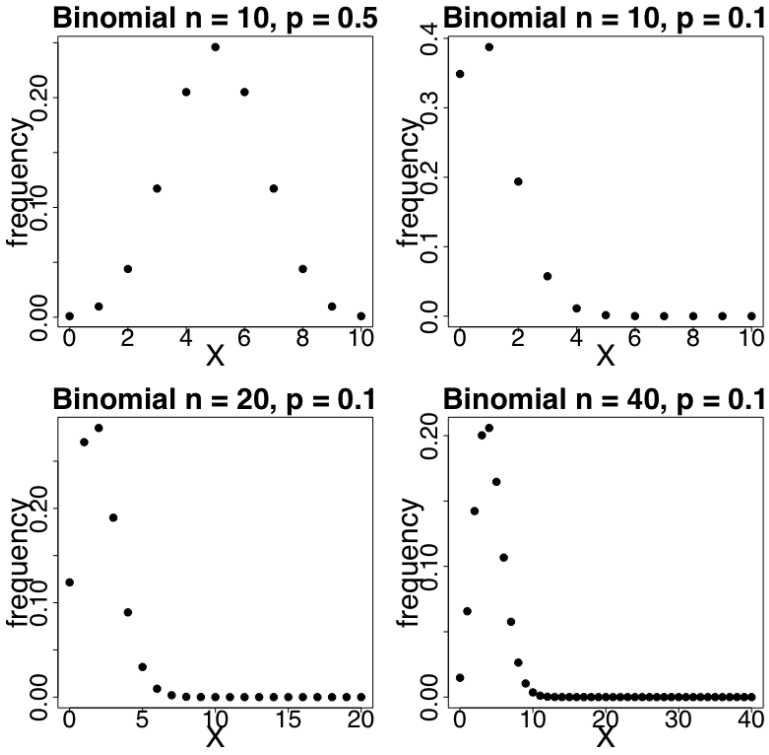
Each entry has the same probability of occurring, namely

$$p^2(1-p)^3.$$

This is because each of the two heads has probability p and each of the 3 tails has probability $1-p$. Because the individual tosses are independent we can multiply probabilities. Therefore, the total probability of exactly 2 heads is the sum of 10 identical probabilities, i.e. $p(2) = 10p^2(1-p)^3$, as shown in the table.

This guides us to the shorter way to do the computation. We have to count the number of sequences with exactly 2 heads. To do this we need to choose 2 of the tosses to be heads and the remaining 3 to be tails. The number of such sequences is the number of ways to choose 2 out of 5 things, that is $\binom{5}{2}$. Since each such sequence has the same probability, $p^2(1-p)^3$, we get the probability of exactly 2 heads $p(2) = \binom{5}{2}p^2(1-p)^3$.

Here are some binomial probability mass function (here, frequency is the same as probability).



3.4 Geometric Distributions

A geometric distribution models the number of tails before the first head in a sequence of coin flips (Bernoulli trials).

Example 9. (a) Flip a coin repeatedly. Let X be the number of tails before the first heads. So, X can equal 0, i.e. the first flip is heads, 1, 2, ... In principle it take any nonnegative integer value.

(b) Give a flip of tails the value 0, and heads the value 1. In this case, X is the number of 0's before the first 1.

(c) Give a flip of tails the value 1, and heads the value 0. In this case, X is the number of 1's before the first 0.

(d) Call a flip of tails a success and heads a failure. So, X is the number of successes before the first failure.

(e) Call a flip of tails a failure and heads a success. So, X is the number of failures before the first success.

You can see this models many different scenarios of this type. The most neutral language is the number of tails before the first head.

Formal definition. The random variable X follows a geometric distribution with parameter p if

- X takes the values 0, 1, 2, 3, ...
- its pmf is given by $p(k) = P(X = k) = (1-p)^k p$.

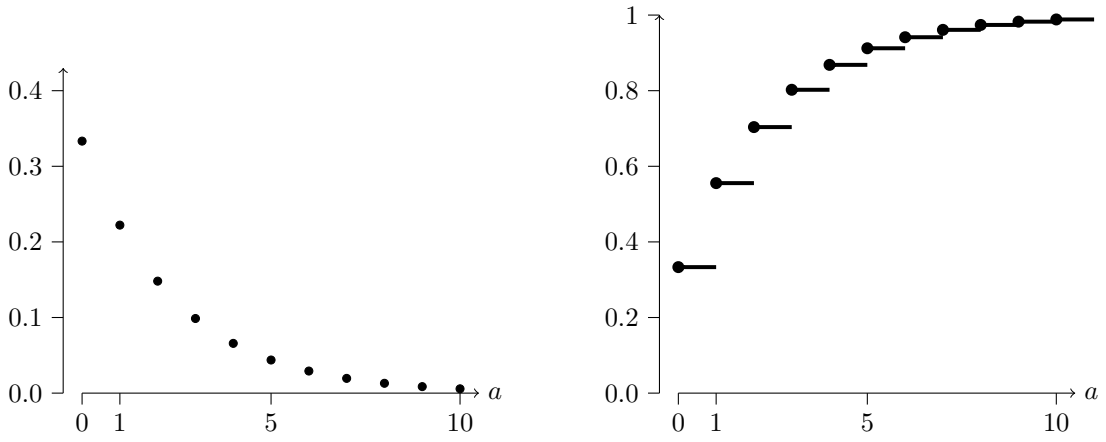
We denote this by $X \sim$ geometric(p) or geo(p). In table form we have:

value	a:	0	1	2	3	...	k	...
pmf	$p(a)$:	p	$(1-p)p$	$(1-p)^2p$	$(1-p)^3p$	...	$(1-p)^kp$	...

Table: $X \sim$ geometric(p): $X =$ the number of 0s before the first 1.

We will show how this table was computed in an example below.

The geometric distribution is an example of a discrete distribution that takes an infinite number of possible values. Things can get confusing when we work with successes and failure since we might want to model the number of successes before the first failure or we might want the number of failures before the first success. To keep straight things straight you can translate to the neutral language of the number of tails before the first heads.



pmf and cdf for the geometric(1/3) distribution

Example 10. Computing geometric probabilities. Suppose that the inhabitants of an island plan their families by having babies until the first girl is born. Assume the probability of having a girl with each pregnancy is 0.5 independent of other pregnancies, that all babies survive and there are no multiple births. What is the probability that a family has k boys?

answer: In neutral language we can think of boys as tails and girls as heads. Then the number of boys in a family is the number of tails before the first heads.

Let's practice using standard notation to present this. So, let X be the number of boys in a (randomly-chosen) family. So, X is a geometric random variable. We are asked to find $p(k) = P(X = k)$. A family has k boys if the sequence of children in the family from oldest to youngest is

$$BBB \dots BG$$

with the first k children being boys. The probability of this sequence is just the product of the probability for each child, i.e. $(1/2)^k \cdot (1/2) = (1/2)^{k+1}$. (Note: The assumptions of equal probability and independence are simplifications of reality.)

Think: What is the ratio of boys to girls on the island?

More geometric confusion. Another common definition for the geometric distribution is the number of tosses until the first heads. In this case X can take the values 1, i.e. the first flip is heads, 2, 3, ... This is just our geometric random variable plus 1. The methods of computing with it are just like the ones we used above.

3.5 Uniform Distribution

The uniform distribution models any situation where all the outcomes are equally likely.

$$X \sim \text{uniform}(N).$$

X takes values $1, 2, 3, \dots, N$, each with probability $1/N$. We have already seen this distribution many times when modeling to fair coins ($N = 2$), dice ($N = 6$), birthdays ($N = 365$), and poker hands ($N = \binom{52}{5}$).

3.6 Discrete Distributions Applet

The applet at http://mathlets.org/mathlets/probability-distributions/ gives a dynamic view of some discrete distributions. The graphs will change smoothly as you move the various sliders. Try playing with the different distributions and parameters.

This applet is carefully color-coded. Two things with the same color represent the same or closely related notions. By understanding the color-coding and other details of the applet, you will acquire a stronger intuition for the distributions shown.

3.7 Other Distributions

There are a million other named distributions arising is various contexts. We don't expect you to memorize them (we certainly have not!), but you should be comfortable using a resource like Wikipedia to look up a pmf. For example, take a look at the info box at the top right of http://en.wikipedia.org/wiki/Hypergeometric_distribution. The info box lists many (surely unfamiliar) properties in addition to the pmf.

4 Arithmetic with Random Variables

We can do arithmetic with random variables. For example, we can add subtract, multiply or square them.

There is a simple, but extremely important idea for counting. It says that if we have a sequence of numbers that are either 0 or 1 then the sum of the sequence is the number of 1s.

Example 11. Consider the sequence with five 1s

$$1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0.$$

It is easy to see that the sum of this sequence is 5 the number of 1s.

We illustrates this idea by counting the number of heads in n tosses of a coin.

Example 12. Toss a fair coin n times. Let X_j be 1 if the jth toss is heads and 0 if it's tails. So, X_j is a Bernoulli(1/2) random variable. Let X be the total number of heads in the n tosses. Assuming the tosses are independence we know $X \sim$ binomial$(n, 1/2)$. We can also write

$$X = X_1 + X_2 + X_3 + \ldots + X_n.$$

Again, this is because the terms in the sum on the right are all either 0 or 1. So, the sum is exactly the number of X_j that are 1, i.e. the number of heads.

The important thing to see in the example above is that we've written the more complicated binomial random variable X as the sum of extremely simple random variables X_j. This will allow us to manipulate X algebraically.

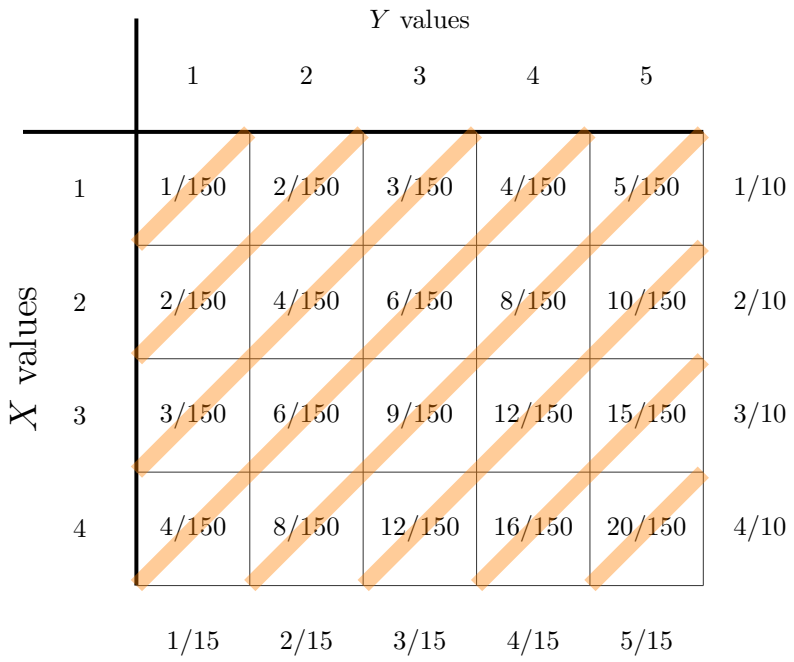
Think: Suppose X and Y are independent and $X \sim$ binomial$(n, 1/2)$ and $Y \sim$ binomial$(m, 1/2)$. What kind of distribution does $X + Y$ follow? (**Answer:** binomial$(n + m, 1/2)$. Why?)

Example 13. Suppose X and Y are independent random variables with the following tables.

Values of X	x:	1	2	3	4	
pmf	$p_X(x)$:	1/10	2/10	3/10	4/10	
Values of Y	y:	1	2	3	4	5
pmf	$p_Y(y)$:	1/15	2/15	3/15	4/15	5/15

Check that the total probability for each random variable is 1. Make a table for the random variable $X + Y$.

answer: The first thing to do is make a two-dimensional table for the product sample space consisting of pairs (x, y), where x is a possible value of X and y one of Y. To help do the computation, the probabilities for the X values are put in the far right column and those for Y are in the bottom row. Because X and Y are independent the probability for (x, y) pair is just the product of the individual probabilities.



The diagonal stripes show sets of squares where $X + Y$ is the same. All we have to do to compute the probability table for $X + Y$ is sum the probabilities for each stripe.

$X + Y$ values:	2	3	4	5	6	7	8	9
pmf:	1/150	4/150	10/150	20/150	30/150	34/150	31/150	20/150

When the tables are too big to write down we'll need to use purely algebraic techniques to compute the probabilities of a sum. We will learn how to do this in due course.

Discrete Random Variables: Expected Value
Class 4, 18.05
Jeremy Orloff and Jonathan Bloom

1 Expected Value

In the R reading questions for this lecture, you simulated the average value of rolling a die many times. You should have gotten a value close to the exact answer of 3.5. To motivate the formal definition of the average, or expected value, we first consider some examples.

Example 1. Suppose we have a six-sided die marked with five 5 3's and one 6. (This was the red one from our non-transitive dice.) What would you expect the average of 6000 rolls to be?

answer: If we knew the value of each roll, we could compute the average by summing the 6000 values and dividing by 6000. Without knowing the values, we can compute the expected average as follows.

Since there are five 3's and one six we expect roughly 5/6 of the rolls will give 3 and 1/6 will give 6. Assuming this to be exactly true, we have the following table of values and counts:

value:	3	6
expected counts:	5000	1000

The average of these 6000 values is then

$$\frac{5000 \cdot 3 + 1000 \cdot 6}{6000} = \frac{5}{6} \cdot 3 + \frac{1}{6} \cdot 6 = 3.5$$

We consider this the expected average in the sense that we 'expect' each of the possible values to occur with the given frequencies.

Example 2. We roll two standard 6-sided dice. You win \$1000 if the sum is 2 and lose \$100 otherwise. How much do you expect to win on average per trial?

answer: The probability of a 2 is 1/36. If you play N times, you can 'expect' $\frac{1}{36} \cdot N$ of the trials to give a 2 and $\frac{35}{36} \cdot N$ of the trials to give something else. Thus your total expected winnings are

$$1000 \cdot \frac{N}{36} - 100 \cdot \frac{35N}{36}.$$

To get the expected average per trial we divide the total by N:

$$\text{expected average} \;=\; 1000 \cdot \frac{1}{36} - 100 \cdot \frac{35}{36} = -69.44.$$

Think: Would you be willing to play this game one time? Multiple times?

Notice that in both examples the sum for the expected average consists of terms which are a value of the random variable times its probabilitiy. This leads to the following definition.

Definition: Suppose X is a discrete random variable that takes values $x_1, x_2, \ldots, x_n$ with probabilities $p(x_1), p(x_2), \ldots, p(x_n)$. The expected value of X is denoted $E(X)$ and defined

by

$$E(X) = \sum_{j=1}^{n} p(x_j)\, x_j = p(x_1)x_1 + p(x_2)x_2 + \ldots + p(x_n)x_n.$$

Notes:

1. The expected value is also called the mean or average of X and often denoted by μ ("mu").
2. As seen in the above examples, the expected value need not be a possible value of the random variable. Rather it is a weighted average of the possible values.
3. Expected value is a summary statistic, providing a measure of the location or central tendency of a random variable.
4. If all the values are equally probable then the expected value is just the usual average of the values.

Example 3. Find $E(X)$ for the random variable X with table:

values of X:	1	3	5
pmf:	1/6	1/6	2/3

answer: $E(X) = \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 3 + \frac{2}{3}\cdot 5 = \frac{24}{6} = 4$

Example 4. Let X be a Bernoulli(p) random variable. Find $E(X)$.

answer: X takes values 1 and 0 with probabilities p and $1-p$, so

$$E(X) = p\cdot 1 + (1-p)\cdot 0 = p.$$

Important: This is an important example. Be sure to remember that the expected value of a Bernoulli(p) random variable is p.

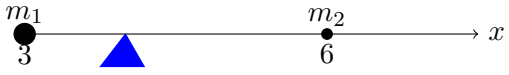
Think: What is the expected value of the sum of two dice?

1.1 Mean and center or mass

You may have wondered why we use the name 'probability mass function'. Here's the reason: if we place an object of mass $p(x_j)$ at position x_j for each j, then $E(X)$ is the position of the center of mass. Let's recall the latter notion via an example.

Example 5. Suppose we have two masses along the x-axis, mass $m_1 = 500$ at position $x_1 = 3$ and mass $m_2 = 100$ at position $x_2 = 6$. Where is the center of mass?

answer: Intuitively we know that the center of mass is closer to the larger mass.



From physics we know the center of mass is

$$\overline{x} = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2} = \frac{500 \cdot 3 + 100 \cdot 6}{600} = 3.5.$$

We call this formula a 'weighted' average of the x_1 and x_2. Here x_1 is weighted more heavily because it has more mass.

Now look at the definition of expected value $E(X)$. It is a weighted average of the values of X with the weights being probabilities $p(x_i)$ rather than masses! We might say that "The expected value is the point at which the distribution would balance". Note the similarity between the physics example and Example 1.

1.2 Algebraic properties of $E(X)$

When we add, scale or shift random variables the expected values do the same. The shorthand mathematical way of saying this is that $E(X)$ is linear.

1. If X and Y are random variables on a sample space Ω then

$$E(X + Y) = E(X) + E(Y)$$

2. If a and b are constants then

$$E(aX + b) = aE(X) + b.$$

We will think of $aX + b$ as scaling X by a and shifting it by b.

Before proving these properties, let's consider a few examples.

Example 6. Roll two dice and let X be the sum. Find $E(X)$.

answer: Let X_1 be the value on the first die and let X_2 be the value on the second die. Since $X = X_1 + X_2$ we have $E(X) = E(X_1) + E(X_2)$. Earlier we computed that $E(X_1) = E(X_2) = 3.5$, therefore $E(X) = 7$.

Example 7. Let $X \sim \text{binomial}(n, p)$. Find $E(X)$.

answer: Recall that X models the number of successes in n Bernoulli(p) random variables, which we'll call $X_1, \ldots X_n$. The key fact, which we highlighted in the previous reading for this class, is that

$$X = \sum_{j=1}^{n} X_j.$$

Now we can use the Algebraic Property (1) to make the calculation simple.

$$X = \sum_{j=1}^{n} X_j \;\Rightarrow\; E(X) = \sum_j E(X_j) = \sum_j p = \boxed{np}.$$

We could have computed $E(X)$ directly as

$$E(X) = \sum_{k=0}^{n} kp(k) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}.$$

It is possible to show that the sum of this series is indeed np. We think you'll agree that the method using Property (1) is much easier.

Example 8. (For infinite random variables the mean does not always exist.) Suppose X has an infinite number of values according to the following table

values x:	2	2^2	2^3	...	2^k	...
pmf $p(x)$:	$1/2$	$1/2^2$	$1/2^3$	...	$1/2^k$	...

Try to compute the mean.

answer: The mean is

$$E(X) = \sum_{k=1}^{\infty} 2^k \frac{1}{2^k} = \sum_{k=1}^{\infty} 1 = \infty.$$

The mean does not exist! This can happen with infinite series.

1.3 Proofs of the algebraic properties of $E(X)$

The proof of Property (1) is simple, but there is some subtlety in even understanding what it means to add two random variables. Recall that the value of random variable is a number determined by the outcome of an experiment. To add X and Y means to add the values of X and Y for the same outcome. In table form this looks like:

outcome ω:	ω_1	ω_2	ω_3	...	ω_n
value of X:	x_1	x_2	x_3	...	x_n
value of Y:	y_1	y_2	y_3	...	y_n
value of $X+Y$:	x_1+y_1	x_2+y_2	x_3+y_3	...	x_n+y_n
prob. $P(\omega)$:	$P(\omega_1)$	$P(\omega_2)$	$P(\omega_3)$	...	$P(\omega_n)$

The proof of (1) follows immediately:

$$E(X+Y) = \sum (x_i + y_i)P(\omega_i) = \sum x_i P(\omega_i) + \sum y_i P(\omega_i) = E(X) + E(Y).$$

The proof of Property (2) only takes one line.

$$E(aX+b) = \sum p(x_i)(ax_i + b) = a\sum p(x_i)x_i + b\sum p(x_i) = aE(X) + b.$$

The b term in the last expression follows because $\sum p(x_i) = 1$.

Example 9. Mean of a geometric distribution

Let $X \sim$ geo(p). Recall this means X takes values $k = 0, 1, 2, \ldots$ with probabilities $p(k) = (1-p)^k p$. (X models the number of tails before the first heads in a sequence of Bernoulli trials.) The mean is given by

$$E(X) = \frac{1-p}{p}.$$

To see this requires a clever trick. Mathematicians love this sort of thing and we hope you are able to follow the logic. In this class we will not ask you to come up with something like this on an exam.

Here's the trick.: to compute $E(X)$ we have to sum the infinite series

$$E(X) = \sum_{k=0}^{\infty} k(1-p)^k p.$$

Here is the trick. We know the sum of the geometric series: $\displaystyle\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$.

Differentiate both sides: $\displaystyle\sum_{k=0}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$.

Multiply by x: $\displaystyle\sum_{k=0}^{\infty} kx^k = \frac{x}{(1-x)^2}$.

Replace x by $1-p$: $\displaystyle\sum_{k=0}^{\infty} k(1-p)^k = \frac{1-p}{p^2}$.

Multiply by p: $\displaystyle\sum_{k=0}^{\infty} k(1-p)^k p = \frac{1-p}{p}$.

This last expression is the mean.

$$E(X) = \frac{1-p}{p}.$$

Example 10. Flip a fair coin until you get heads for the first time. What is the expected number of times you flipped tails?

answer: The number of tails before the first head is modeled by $X \sim \text{geo}(1/2)$. From the previous example $E(X) = \dfrac{1/2}{1/2} = 1$. This is a surprisingly small number.

Example 11. Michael Jordan, the greatest basketball player ever, made 80% of his free throws. In a game what is the expected number he would make before his first miss.

answer: Here is an example where we want the number of successes before the first failure. Using the neutral language of heads and tails: success is tails (probability $1-p$) and failure is heads (probability $= p$). Therefore $p = .2$ and the number of tails (made free throws) before the first heads (missed free throw) is modeled by a $X \sim \text{geo}(.2)$. We saw in Example 9 that this is

$$E(X) = \frac{1-p}{p} = \frac{.8}{.2} = 4.$$

1.4 Expected values of functions of a random variable

(The change of variables formula.)

If X is a discrete random variable taking values x_1, x_2, ... and h is a function the $h(X)$ is a new random variable. Its expected value is

$$E(h(X)) = \sum_j h(x_j)p(x_j).$$

We illustrate this with several examples.

Example 12. Let X be the value of a roll of one die and let $Y = X^2$. Find $E(Y)$.

answer: Since there are a small number of values we can make a table.

X	1	2	3	4	5	6
Y	1	4	9	16	25	36
prob	1/6	1/6	1/6	1/6	1/6	1/6

Notice the probability for each Y value is the same as that of the corresponding X value. So,

$$E(Y) = E(X^2) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \ldots + 6^2 \cdot \frac{1}{6} = 15.167.$$

Example 13. Roll two dice and let X be the sum. Suppose the payoff function is given by $Y = X^2 - 6X + 1$. Is this a good bet?

answer: We have $E(Y) = \sum_{j=2}^{12} (j^2 - 6j + 1)p(j)$, where $p(j) = P(X = j)$.

We show the table, but really we'll use R to do the calculation.

X	2	3	4	5	6	7	8	9	10	11	12
Y	-7	-8	-7	-4	1	8	17	28	41	56	73
prob	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Here's the R code I used to compute $E(Y) = 13.833$.

```
x = 2:12
y = x^2 - 6*x + 1
p = c(1 2 3 4 5 6 5 4 3 2 1)/36
ave = sum(p*y)
```

It gave ave = 13.833.

To answer the question above: since the expected payoff is positive it looks like a bet worth taking.

Quiz: If $Y = h(X)$ does $E(Y) = h(E(X))$? **answer: NO!!!** This is not true in general!

Think: Is it true in the previous example?

Quiz: If $Y = 3X + 77$ does $E(Y) = 3E(X) + 77$?

answer: Yes. By property (2), scaling and shifting does behave like this.