Lectures 1 and 2

jacques@ucsd.edu

1.1 Sets

To define probability rigorously, we first need to discuss sets and operations on sets. A *set* is an unordered collection of elements. We list the elements of a set between braces, for example, $\{1, 2, 3\}$ is the set whose elements are 1, 2 and 3. The elements are unordered, so this means that $\{1, 2, 3\}$ is the same set as $\{2, 3, 1\}$ and $\{1, 3, 2\}$. Two sets are *equal* if and only if they have the same elements; thus $\{1, 2, 3\} = \{1, 3, 2\}$. The number of elements in a set is called the *size* or *cardinality* of the set. For instance, the set $\{1, 2, 3\}$ has cardinality 3. The cardinality of the set $\mathbb{N}$ of natural numbers $\{1, 2, \dots\}$ is infinite, and so is the cardinality of the set $\mathbb{Z}$ of integers $\{\dots, -2, -1, 0, 1, \dots\}$ and the set $\mathbb{R}$ of real numbers. If S is a set, then $|S|$ denotes its cardinality. We do not spend time here on rigorously defining sets or constructing these well known sets of numbers. The set with no elements – the *empty set* – is denoted $\emptyset$.

Intersections, unions and complements. We can perform the following operations on sets. The *intersection* of two sets A and B is the set of elements which are in A and in B, written

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

The union of A and B is

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

In other words, $A \cup B$ is the set of elements which are in A or in B. We say that A is a *subset* of B if every element of A is also an element of B, in which case we write $A \subseteq B$. If A and B are sets, then the set of elements of A which are not in B is

$$A \backslash B = \{x \in A : x \notin B\}$$

The *symmetric difference* of A and B is

$$A \triangle B = (A \backslash B) \cup (B \backslash A).$$

If we are considering only subsets of a single set Ω, and A is a subset of a set Ω, then we write $\overline{A}$ instead of $\Omega \backslash A$. This is the *complement* of A.

Disjoint sets and partitions. Two sets A and B are *disjoint* if $A \cap B = \emptyset$ – in other words, they have no elements in common. A *partition* of a set Ω is a set of disjoint sets whose union is Ω. So if $A_1, A_2, \ldots, A_n$ are the sets in the partition, then every two different sets are disjoint – we say the sets are *pairwise disjoint* – and

$$A_1 \cup A_2 \cup \cdots \cup A_n = \Omega.$$

Sometimes to denote that the sets in this union are pairwise disjoint, we instead write

$$A_1 \sqcup A_2 \sqcup \cdots \sqcup A_n = \Omega.$$

Countable and uncountable sets. An infinite set is countable if its elements can be labeled with $\mathbb{N}$, and uncountable otherwise.

Venn Diagrams. A convenient way to represent sets is via Venn diagrams. In the picture below, we depict the Venn diagram of three sets A, B and C from which we can deduce a number of things, such as $A \cap B \cap C = \emptyset$ (no element is in common to A, B and C) and $A \subseteq B \cup C$ (the set A is contained in $B \cup C$ – every element of A is also an element of B or an element of C).

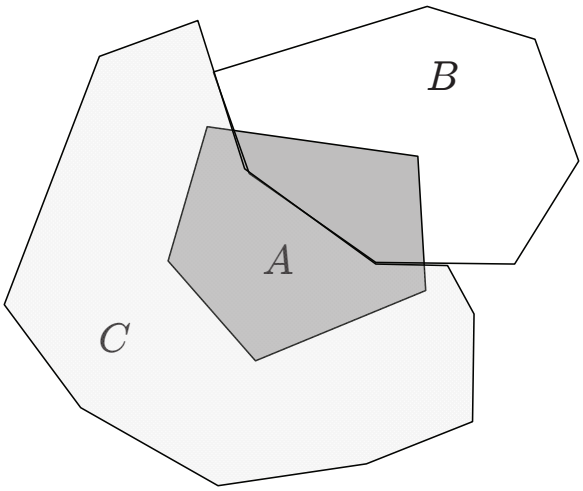


Figure 1 : Venn Diagram

To illustrate $\Omega = A \sqcup B \sqcup C \sqcup D$, in other words, that the sets A, B, C, D form a partition of the set Ω, we might draw the following diagram. From the diagram we easily see $\overline{A \cup B} = C \cup D = \overline{A} \cap \overline{B}$, for instance.

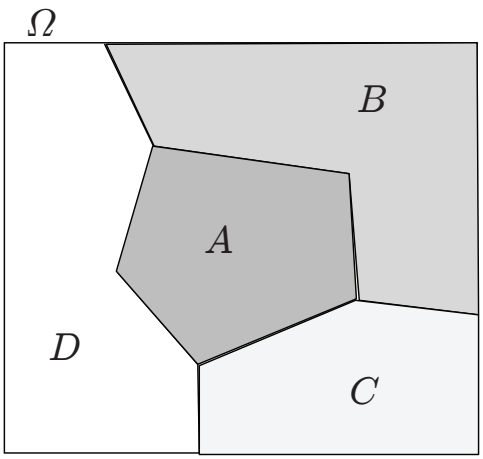


Figure 1 : Partition of Ω into four sets

Set identities. The following is sometimes called *deMorgan's Law.*

Proposition 1 *Let A, B be subsets of a set Ω. Then $\overline{\overline{A}} = A$ and*

$$\overline{A} \cap \overline{B} = \overline{A \cup B}$$
$$\overline{A} \cup \overline{B} = \overline{A \cap B}.$$

This can be proved by drawing the Venn diagrams of the sets, or by writing down logically what each side means: if a is an element of $\overline{A} \cap \overline{B}$, then a is not an element of A and not an element of B. Equivalently, this means that a is not an element of $A \cup B$. We observe the *distributivity laws* of union and intersection, which can again be checked with Venn diagrams.

Proposition 2 *Let A, B be sets. Then*

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C). \end{aligned}$$

1.2 Probability Spaces

To define probability in this course, we need three ingredients: a sample space, a set of events, and a probability measure. A *sample space* is just a set, which will will conventionally denote by Ω. The elements $\omega \in \Omega$ will be called *sample points*, and subsets of Ω are called *events*. The set of all events is denoted $\mathcal{F}$ and is called a *σ-field* or *σ-algebra*, and is required to satisfy the following requirements: $\Omega \in \mathcal{F}$, if $A \in \mathcal{F}$ then $\overline{A} \in \mathcal{F}$, and finally if $A_1, A_2, \cdots \in \mathcal{F}$ then $A_1 \cup A_2 \cup \cdots \in \mathcal{F}$. We imagine Ω to be the set of all possible outcomes of an experiment, and then the σ-field $\mathcal{F}$ of events are just sets of outcomes. An event is said to *occur* if the outcome of the experiment is contained in the event.

For example, we might flip a coin and record whether we get heads or tails. Then Ω could be $\{h, t\}$ and possible events are $\emptyset, \{h\}, \{t\}, \{h, t\}$. Then $\emptyset$ is the event that neither heads nor tails occurred, $\{h\}$ is the event of getting heads, $\{t\}$ is the event of getting tails, and $\{h, t\}$ is the event that heads occurred or tails occurred. Note that this is just Ω and $\mathcal{F} = \{\emptyset, \{h\}, \{t\}, \{h, t\}\}$ in this case.

A *probability measure* is a function $P : \mathcal{F} \to [0, 1]$ from the σ-field $\mathcal{F}$ of events in Ω into the unit interval $[0, 1]$ satisfying the following *axioms of probability*:

Axioms of Probability

1. $P(\Omega) = 1$
2. For any event $A \in \mathcal{F}$, $P(A) \geq 0$.
3. For disjoint events $A_1, A_2, \cdots \in \mathcal{F}$,

$$P(A_1 \sqcup A_2 \sqcup \dots) = \sum_{i=1}^{\infty} P(A_i).$$

Together, $\Omega, \mathcal{F}$ and P form what is called a *probability space*, denoted $(\Omega, \mathcal{F}, P)$. The axioms of probability are motivated by natural considerations of how probability should behave. The above framework allows us to set up probability as a mathematical concept, and thereby introduce the machinery of mathematics to make probability theory rigorous. We now give some basic examples of probability spaces.

1.3 Examples of probability spaces

First Example: Tossing a coin. Consider the sample space $\Omega = \{h, t\}$ and set of events $\mathcal{F} = \{\emptyset, \{h\}, \{t\}, \{h, t\}\}$. These arise from tossing a coin and recording the outcome, heads or tails. Let $P(\emptyset) = 0$, $P(\{h\}) = 1/2 = P(\{t\})$ and $P(\Omega) = 1$. This probability measure is a natural one: it is the case that the coin is a *fair coin*, since the events $\{h\}$ and $\{t\}$ are equally likely to occur – there is an equal probability of tossing heads and tossing tails. Now $(\Omega, \mathcal{F}, P)$ is a probability space, since it satisfies the three axioms of probability. In fact, if we want to define P so that the coin is a fair coin, i.e. so that $P(\{h\}) = P(\{t\})$, then the probability of heads and tails must both be $1/2$, since by Axioms 1 and 3,

$$1 = P(\Omega) = P(\{h\}) + P(\{t\}) = 2P(\{h\})$$

and so $P(\{h\}) = 1/2$.

In this first example, we did not have a choice in writing $P(\emptyset) = 0$. In fact, this is true in every probability space, since by Axioms 1 and 3,

$$1 = P(\Omega) = P(\Omega \sqcup \emptyset) = P(\Omega) + P(\emptyset) = 1 + P(\emptyset)$$

and so $P(\emptyset) = 0$.

Second Example: Biased coin. A coin is tossed and the outcome record. If the coin is twice as likely to land on heads as tails, what should the probability space be? The sample space is still $\Omega = \{h, t\}$ and the events are the same as in the first example. But now the probability measure satisfies

$$P(\{h\}) = 2P(\{t\})$$

since heads is twice as likely. So by Axioms 1 and 3,

$$1 = P(\Omega) = P(\{h\}) + P(\{t\}) = 3P(\{t\}).$$

So $P(\{t\}) = 1/3$ and then $P(\{h\}) = 2/3$, as expected. We still have $P(\emptyset) = 0$.

Third Example: Fair die. We could have repeated the first example for the case of tossing a six-sided die. There $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{F}$ consists of all the subsets of Ω. If the die is fair, then it would make sense to define

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{6\}) = \frac{1}{6}.$$

Now for a general event A, such as $A = \{1, 2, 5\}$, how do we define the probability measure? According to Axiom 3,

$$P(A) = P(\{1\} \sqcup \{2\} \sqcup \{5\}) = P(\{1\}) + P(\{2\}) + P(\{5\}) = \frac{3}{6} = \frac{1}{2}.$$

In general, for an event A, $P(A) = \frac{|A|}{6}$. We could ask: what is the probability of tossing a number larger than 2 on the die? The event we want is $A = \{3, 4, 5, 6\}$ and so

$$P(A) = \frac{|A|}{6} = \frac{4}{6} = \frac{2}{3}.$$

We will see soon that the first and third examples are special cases of something called a *uniform probability measure* – loosely speaking, this is a probability measure which assigns to every element of Ω the same probability.

Fourth Example: An infinite space. Imagine dropping a pin vertically onto the interval $[0, 1]$. Let Ω be the set of locations where the tip of the pin could land, namely $\Omega = [0, 1]$ – here we are under the assumption that the tip has zero width. We could define a probability measure P by saying $P([a, b]) = b - a$ – in other words, the chance that the pin lands in the interval $[a, b]$ is $b-a$. By taking unions and complements of those intervals, we create $\mathcal{F}$, which is rather complicated looking. We observe that the event $\{a\} = [a, a]$ that the tip lands at a particular point $a \in [0, 1]$ has probability $P([a, a]) = a - a = 0$. Even though the pin has no chance of landing at any particular point, $P(\Omega) = 1$. This does not violate Axiom 3, since we cannot write $\Omega = \{a_1\} \sqcup \{a_2\} \sqcup \ldots$ for any countable sequence of points $a_1, a_2, \ldots$ because $[0, 1]$ is *uncountable*. What is the probability of the event Q landing at a rational number[1] in $[0, 1]$? Now the rationals in $[0, 1]$ are known to be countable, so we can use Axiom 3 to obtain $P(Q) = 0$. By the complement rule, this means that $P(I) = 1$ where I is the set of irrational numbers in $[0, 1]$. We have not mentioned $\mathcal{F}$ explicitly here for a very good reason. The reason is that $\mathcal{F}$ is difficult to describe: it contains all countable unions and complements of intervals $[a, b]$, but does this give all subsets of Ω? It turns out that it does *not*: this means that there are subsets of Ω to which a probability cannot be assigned. These are called *non-measurable sets*, and are beyond the scope of this course. This example serves to illustrate some complications which can arise when dealing with infinite probability spaces, even though infinite probability spaces will arise frequently and very naturally in subsequent material.

[1]Recall, a rational number is a real number of the form m/n where m, n are integers and $n \neq 0$ – i.e. they are the fractions. The ones in $[0, 1]$ are those of the form m/n with $0 \leq m \leq n$ and there are countably many of them.

Lectures 3 and 4

jacques@ucsd.edu

2.1 Two consequences of the axioms

One of the main reasons for introducing sets at the beginning was to determine formulas for the probability of unions, intersections, and complements of sets, using the axioms of probability. These general rules apply in any probability space. Throughout this section, (Ω, F, P) is a probability space. The first consequence of the axioms is the rule of *complements*. Recall the complement of a set A, which is denoted $\overline{A}$[1] is

$$\Omega \backslash A = \{\omega \in \Omega : \omega \notin A\}.$$

This is the set of outcomes which is not in A. From the axioms, we obtain:

Proposition 1 *Let A be an event. Then*

$$P(\overline{A}) = 1 - P(A).$$

Proof ▷ The events $\overline{A}$ and A partition Ω, by definition, so from Axioms 1 and 3,

$$1 = P(\Omega) = P(\overline{A}) + P(A).$$

This gives the formula. ∎

The above formula is useful, as it is sometimes much easier to compute $P(A)$ than $P(\overline{A})$, or vice versa. The *inclusion-exclusion* formula is a general rule which helps to compute combinatorial probabilities, and follows as a useful generalization of Axiom 3 of probability.

Proposition 2 (Inclusion-Exclusion Formula) *Let A and B be events in a probability space. Then*

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

[1] Some books write A^c instead

Proof ▷ We may write $A \cup B$ as a disjoint union of the following three sets:

$$A \cup B = (A \backslash B) \sqcup (B \backslash A) \sqcup (A \cap B)$$

So by Axiom 3 of probability,

$$P(A \cup B) = P(A \backslash B) + P(B \backslash A) + P(A \cap B).$$

Now $A \backslash B$ and $A \cap B$ are two disjoint sets whose union is A. So using Axiom 3,

$$P(A) = P(A \backslash B) + P(A \cap B).$$

This gives $P(A \backslash B) = P(A) - P(A \cap B)$. Similarly, $P(B \backslash A) = P(B) - P(A \cap B)$. Putting these two equations into the equation for $P(A \cup B)$, we get

$$\begin{aligned} P(A \cup B) &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

This completes the proof. ∎

We could actually take unions of even more sets, for example, for three sets one obtains an inclusion-exclusion formula for three sets:

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

2.2 Computing probabilities.

Let's do examples of how to use the formulas for sets and probabilities to compute a probability.

First Example: Fair dice. Two dice are tossed, and the sum of the two numbers is recorded. The sample space Ω is the set of all ordered pairs of numbers up to six, i.e.

$$\Omega = \{(1,1), (1,2), (2,1), \ldots, (6,6)\}.$$

The events are all the subsets of Ω, and the probability measure we will use is

$$P(A) = \frac{|A|}{36}.$$

Note that we could imagine this as the probability measure which comes from tossing two fair dice independently of one another. It is also the *uniform measure* on Ω, since $P(\{(i,j)\}) = 1/36$ for every pair $(i,j) \in \Omega$ – this means that i was tossed on the first die, and j was tossed on the second. Consider now the event A that the sum of the two numbers on the dice is a single digit number, namely $1, 2, 3, 4, 5, 6, 7, 8$ or 9. We could directly write down all the outcomes in A, there are 30 of them, and then we could compute $P(A) = |A|/36 = 30/36 = 5/6$. But it is much easier to use the complement rule: $\overline{A}$ is the event that the sum is $10, 11$ or 12, so

$$\overline{A} = \{(4,6), (6,4), (5,5), (5,6), (6,5), (6,6)\}$$

and then

$$P(\overline{A}) = \frac{|\overline{A}|}{36} = \frac{6}{36} = \frac{1}{6}.$$

By the complement rule,

$$P(A) = 1 - P(\overline{A}) = \frac{5}{6}$$

which agrees with the preceding answer.

Second Example: Two fair dice. If two fair dice are thrown, what is the chance that at least one of the dice shows an odd number? Let A be the event that the first die shows an odd number, and let B be the event that the second shows an odd number. We are looking for $P(A \cup B)$ in the probability space (Ω, F, P) given in the first example. Now we know by the inclusion-exclusion formula that

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

We carefully write down A:

$$A = \{(1,1), (1,2), \ldots, (1,6), (3,1), (3,2), \ldots, (3,6), (5,1), (5,2), \ldots, (5,6)\}.$$

There are 18 outcomes in A, so

$$P(A) = \frac{|A|}{36} = \frac{1}{2}.$$

We could have argued this in a simpler way: half the numbers on the first die are odd, so there is a fifty percent chance that the first die shows an odd number, regardless of what the second die shows. Similarly, we get $P(B) = 1/2$. Now $A \cap B$ is the event that both numbers are odd, which is $1/4$, since

$$A \cap B = \{(1,1), (3,3), (5,5), (1,3), (3,1), (1,5), (5,1), (3,5), (5,3)\}$$

and so

$$P(A \cap B) = \frac{|A \cap B|}{36} = \frac{9}{36} = \frac{1}{4}.$$

Again we could have argued more simply: there are three odd numbers possible on the first die, and three on the second, so there are nine combinations which give two odd numbers. Finally, by inclusion-exclusion,

$$P(A \cup B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}.$$

We remark that whenever one sees in a probability problem the phrase *at least one*, this indicates the union of some events.

The purpose of the next example is to illustrate how one has sometimes to use careful logic to compute the required probability in an efficient way.

Third Example: Logic and sets. Two fair six-sided dice are thrown, with probability measure $P(A) = |A|/36$ in the same probability space as the last two examples. Let A be the event that the sum of the dice is odd, or at least one of the numbers is odd and both dice show different numbers. To compute $P(A)$, we could list all outcomes (as an exercise) in A and we would find that there are 24 pairs of numbers in A, so $P(A) = 24/36 = 2/3$. However,

$$A = B \cup (C \cap D)$$

where B is the event that the sum of the dice is odd, C is the event that at least one of the numbers is odd, and D is the event that both dice show different numbers. Therefore

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

using inclusion-exclusion. Observe that $B \cap C \cap D = B$, for if B occurs (sum is odd), then C must occur, and D must also occur. Therefore

$$P(A) = P(C \cap D).$$

Now $C \cap D$ is the event that one of the dice shows an odd number and the other shows a different number. It is easier to deal with $\overline{C \cap D} = \overline{C} \cup \overline{D}$ by deMorgan's law – this is the event that both dice show an even number or both dice show the same number. Then by inclusion-exclusion

$$P(\overline{C \cap D}) = P(\overline{C} \cup \overline{D}) = P(\overline{C}) + P(\overline{D}) - P(\overline{C} \cap \overline{D}).$$

Now $P(\overline{C}) = 1/4$ (probability both dice are even) and $P(\overline{D}) = 1/6$ (probability both dice show the same number) and $\overline{C} \cap \overline{D} = \{(2,2),(4,4),(6,6)\}$ (both dice are even and show the same number). Therefore

$$P(\overline{C \cap D}) = \frac{1}{4} + \frac{1}{6} - \frac{3}{36} = \frac{1}{3}.$$

By the complement rule,

$$P(A) = P(C \cap D) = 1 - P(\overline{C \cap D}) = 1 - \frac{2}{3} = \frac{2}{3}.$$

This answers the problem, but it is not the only way to do this problem. It is important to become familiar with manipulating sets to compute probabilities.

Fourth Example: Probability in the integers. Let $\Omega = \mathbb{Z}$ and let $\mathcal{F}$ consist of all subsets of Ω. Suppose we define

$$P(\{\omega\}) = \frac{1}{2^{\omega}}.$$

Then P defines a probability measure by letting for any $A \subset \Omega$

$$P(A) = \sum_{\omega \in A} 2^{-\omega}.$$

To see this, certainly $P(A) \geq 0$ for every A and

$$P(\Omega) = \sum_{\omega \in \mathbb{Z}} 2^{-\omega} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 1$$

since the sum is a geometric series. In general, if P is a probability measure defined on the integers, then the requirement

$$\sum_{\omega \in \mathbb{Z}} P(\{\omega\}) = 1$$

implies from elementary calculus that

$$\lim_{\omega \to \infty} P(\{\omega\}) = 0$$

in other words, larger and larger integers must have smaller and smaller probabilities.

Fifth Example: Combinatorial probability. We are going to spend some time later on looking at combinatorial probability, here we do a first example. Consider a standard deck of 52 cards (thirteen cards in each of four suits, diamonds, clubs, spades and hearts, and thirteen different card values $2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A$, four of each value –

for example the diamonds and spades are shown below). Suppose we consider drawing five cards from the deck with the uniform probability measure on all five card hands. Thus the sample space Ω is the set of all five card hands, $\mathcal{F}$ is the set of all subsets of Ω and, for any event A, $P(A) = |A|/|\Omega|$. Here Ω is very large:

$$|\Omega| = 52 \cdot 51 \cdot 50 \cdot 49 \cdot 48$$

is the total number of five card hands. Let's compute the probability that at least two of the cards in a five card hand are of the same value (in poker, this is called a pair). This is a counting problem: how many ways can we fill five slots with five cards such that two of the cards are of the same suit (this is $|A|$), and then we divide by $|\Omega| = 52 \cdot 51 \cdots 48$ to get the probability. It turns out to be easier to compute $P(\overline{A})$ – this is the probability that all cards have different values. There are 52 choices to fill the first slot, but now the second slot can't have a card of the same value, so there are only 48 choices for the second slot. Similarly there are 44 choices for the third, 40 for the fourth, and 36 for the last. So

$$P(\overline{A}) = \frac{52 \cdot 48 \cdot 44 \cdot 40 \cdot 36}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} = 0.50708\ldots$$

Therefore

$$P(A) = 0.49292\ldots$$

So it is in fact more likely than not that all the cards will be different (but perhaps surprisingly close to 1/2).

Figure 1 : Diamonds and spades

Lectures 5 and 6

jacques@ucsd.edu

3.1 Discrete Uniform Probability Measure

Combinatorial probability refers loosely to problems involving finite sample spaces. These problems are largely speaking more about counting outcomes than about probability, especially when the probability measure is the uniform measure. For this reason, we define the uniform probability measure on finite probability spaces as follows.

Discrete Uniform Probability Measure. *Let $(\Omega, \mathcal{F}, P)$ be a probability space such that Ω is a finite set. Then P is the discrete uniform probability measure if for every event A,*

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

If P is uniform, then P must assign to each outcome $\omega \in \Omega$ the same probability (this is why this is the uniform probability measure), and so we must have $P(\{\omega\}) = 1/|\Omega|$ for every $\omega \in \Omega$. So to work with the discrete uniform probability measure, if we want $P(A)$ for some event A, we just need to compute $|A|$ and $|\Omega|$, and thus we need to count the outcomes in A. Here is a first simple example:

First Example. A roulette wheel is spun, and the ball is equally likely to land on any number $0, 1, \ldots, 36$. Determine the probability that the ball lands on a red number (see Figure 1 below). In this case, the probability space is $(\Omega, \mathcal{F}, P)$ with P the uniform measure defined by $P(A) = |A|/37$. Now the event A that the outcome is red contains eighteen outcomes, namely

$$1 \quad 3 \quad 5 \quad 7 \quad 9 \quad 12 \quad 14 \quad 16 \quad 18 \quad 23 \quad 25 \quad 27 \quad 30 \quad 32 \quad 34 \quad 36.$$

Therefore $P(A) = 18/37$ is the probability of landing on red.

Figure 1 : Roulette Wheel

Second Example. The second and third examples of Lecture 2 involved the uniform measure on $\Omega = \{(1,1),(1,2),\ldots,(6,6)\}$, which corresponds to tossing two fair dice. The probability measure is $P(A) = |A|/36$. We might ask for the probability that the product of the two dice is at most 9. This event A has outcomes

$$(1,1) \quad (1,2) \quad (1,3) \quad (1,4) \quad (1,5) \quad (1,6) \quad (2,1)$$

$$(2,2) \quad (2,3) \quad (2,4) \quad (3,1) \quad (3,2) \quad (3,3).$$

Therefore $|A| = 13$ and $P(A) = 13/36$.

Third Example. We cannot define a uniform measure on a countable probability space $(\Omega, \mathcal{F}, P)$ if $\mathcal{F}$ is all subsets of Ω. For if we tried to define a uniform probability measure on $\mathbb{N}$ where $\mathcal{F}$ is the set of all subsets of $\mathbb{N}$, we require all $P(\{\omega\})$ to be equal for every $\omega \in \mathbb{N}$, but then by Axiom 3,

$$1 = \mathbb{P}(\Omega) = \sum_{\omega \in \mathbb{N}} P(\{\omega\})$$

and this is impossible – we cannot add together the same number infinitely many times to get 1.

3.2 Counting sequences

A unified way to consider many counting problems is the multiplication principle. To define this principle, we first define *sequences*, which are the ordered counterparts of sets (recall the elements of sets are unordered). A *sequence* over a set Σ is an ordered list of elements of Σ. For example, a binary sequence is an ordered list over $\{0, 1\}$ – it is an ordered list of zeroes and ones. The notation for sequences is to use round brackets and to separate the elements by commas, so for example $(1, 0, 1)$ is a binary sequence, and it is different to the binary sequences $(1, 1, 0)$ and $(0, 1, 1)$. The multiplication principle is a method for counting sequences.

Multiplication Principle. *Let N be the number of sequences $(x_1, x_2, \ldots, x_k)$ such that given $x_1, x_2, \ldots, x_{i-1}$, there are N_i choices for x_i, for $i = 1, 2, \ldots, k$. Then*

$$N = N_1 N_2 \ldots N_k.$$

First Example. We have mentioned the example of tossing two dice frequently, where the sample space Ω of all pairs has $|\Omega| = 36$. The multiplication principle justifies that $|\Omega| = 36$, since Ω consists of all sequences (x_1, x_2) and there are six choices for x_1 and six choices for x_2, so

$$|\Omega| = N_1 N_2 = 6 \cdot 6 = 36.$$

Second Example. Determine the number of binary sequences of length ten which do not have two consecutive 1s or two consecutive 0s. First, observe that the total number of binary sequences of length ten is 2^{10} by the multiplication principle: there are $N_1 = 2$ choices for the first digit (0 or 1), then $N_2 = 2$ choices for the next, and so on. Now with the restriction that there are no two consecutive 0s and no two consecutive 1s, $N_1 = 2$ (the first digit can be 0 or 1), but once we have chosen the first digit, the next digit must be different, so there are $N_2 = 1$ choice for the next digit. Then the next digit is again different to the second one, so $N_3 = 1$, and so on. In other words, once we have chosen the first digit of the sequence, all the other digits are uniquely determined. So the number of binary sequences required is $N_1 N_2 \ldots N_{10} = 2$ by the multiplication principle. The reader should check that the problem of counting binary sequences of length 10 with no two consecutive 1s is more difficult.

Third Example. Five people's birthdays are recorded, and the sequence of their birthdays is recorded as a sequence of five numbers in $\{1, 2, \ldots, 365\}$. How many sequences

of birthdays arise in this way? There are $N_1 = 365$ choices for the first birthday, and then $N_2 = 365$ choices for the second, and so on up to $N_5 = 365$. By the multiplication principle, there are $N_1 N_2 N_3 N_4 N_5 = (365)^5$ possible sequences of five birthdays. Now suppose all the birthdays are on different days. Then how many sequences are possible? This is trickier, since the first birthday could be anything, so there are still $N_1 = 365$ choices, but the second birthday must then be different from the first, and so $N_2 = 364$. Similarly, $N_3 = 363$ and $N_4 = 362$ and $N_5 = 361$, so the number of sequences of five different birthdays is (by the multiplication principle)

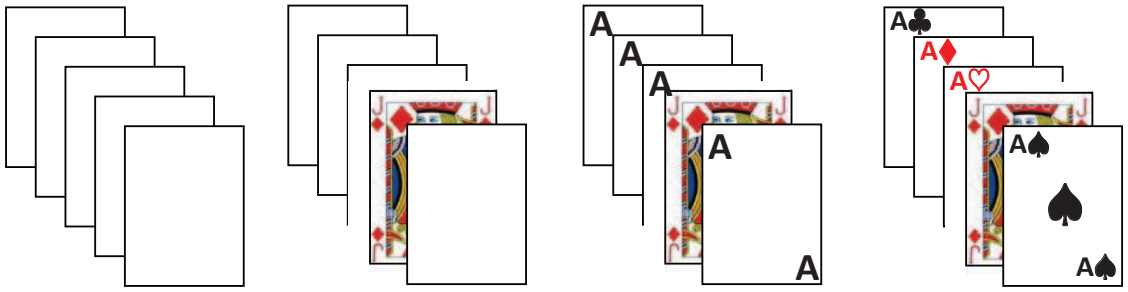
$$N_1 N_2 N_3 N_4 N_5 = 365 \cdot 364 \cdot 363 \cdot 362 \cdot 361.$$

We will come to this interesting example again when we discuss the *birthday paradox*.

Third Example. A poker hand (five cards) is drawn from a standard 52 card deck. What is the chance of getting four cards of the same value? We imagine this as a sequence of five choices of cards, where four of the cards must have the same value. One of the cards will be of a different value than the four others (call this the odd card). There are $N_1 = 5$ choices for the position of this card in the sequence and $N_2 = 52$ ways to decide what that card is. Having chosen where that card is, we must fill the remaining four places in the sequence with four cards of the same value (but different value than the odd card). There are $N_3 = 12$ possible card values. Once we know the card value, we have to assign the four cards (spade, diamond, heart, club) of that value to the four open slots in the sequence. There are $N_4 = 4$ ways to assign where the spade goes, then $N_5 = 3$ ways to assign where the diamond goes, $N_6 = 2$ ways to assign the heart and $N_7 = 1$ way to fill the remaining slot with the club suit. By the multiplication principle, the number of such poker hands is

$$N_1 N_2 N_3 N_4 N_5 N_6 N_7 = 5 \cdot 52 \cdot 12 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 74880.$$

An illustration of this argument is given in the picture below.



No cards chosen

Position of odd card chosen and odd card selected and placed

$5 \cdot 52$ ways

Value of four cards chosen

12 ways

Four cards placed

$4{\cdot}3{\cdot}2{\cdot}1$ ways

Figure 2 : Choosing four of a kind

3.3 Counting subsets

In order to help us with counting, we want to count subsets of sets. For example, how many subsets of $\{1, 2, 3, 4, 5\}$ have size 3? We could list them all: they are

$$\{1,2,3\} \quad \{1,2,4\} \quad \{1,2,5\} \quad \{1,3,4\} \quad \{1,3,5\}$$
$$\{1,4,5\} \quad \{2,3,4\} \quad \{2,3,5\} \quad \{2,4,5\} \quad \{3,4,5\}$$

It becomes impractical to list all the sets in this way, for example, there are a very large number of subsets of $\{1, 2, \ldots, 100\}$ of size 50. The following gives a general formula for counting these subsets. Recall that $n!$ – n *factorial* – is the product of all positive integers from 1 to n. For instance, $2! = 2 \cdot 1$, $3! = 3 \cdot 2 \cdot 1 = 6$, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$, and so on. We define $0! = 1$. Factorials are important in counting subsets:

Proposition 1 *The number of subsets of size $k \geq 0$ in a set of size $n \geq k$ is*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{k!}.$$

The notation $\binom{n}{k}$ is read "n choose k" – we are counting how many ways to choose k elements from an n element set. For example, the number of subsets of $\{1, 2, 3, 4, 5\}$ of

size three is 10 from the example above, and from the proposition with $n = 5$ and $k = 3$ we get

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = \frac{120}{6 \cdot 2} = 10$$

as we expected. The number of subsets of $\{1, 2, \ldots, 100\}$ of size 50 is

$$\binom{100}{50} = \frac{100(99)(98)\ldots(51)}{50(49)(48)\ldots 1} = 100891344545564193334812497256.$$

The quantities $\binom{n}{k}$ are called *binomial coefficients*, and we will encounter them frequently in the material to follow.

3.4 Combinatorial Probability

We now apply our methods for counting subsets and counting sequences to combinatorial probability problems. In most cases, we are dealing with a probability space $(\Omega, \mathcal{F}, P)$ where Ω is a finite set and P is the uniform measure, namely, for every $A \in \mathcal{F}$,

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

This probability measure is often indicated by saying that outcomes are "uniformly chosen" or occur "uniformly".

First Example. Determine the probability of the event F of drawing four of a kind in a uniformly chosen five card poker hand. The sample space Ω is the set of sequences of five cards (i.e. poker hands). By the multiplication principle,

$$|\Omega| = 52 \cdot 51 \cdot 50 \cdot 49 \cdot 48.$$

According to the third example of Section 3.2,

$$|F| = 74880.$$

Therefore, since P is the uniform measure,

$$P(F) = \frac{|F|}{|\Omega|} = \frac{74880}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} = \frac{13}{49980} = 0.000240096\ldots.$$

So there is roughly a 1 in 4000 chance of drawing four of a kind in a poker hand.

Second Example. Determine the probability of the event C of drawing at least two cards of the same value in a uniformly chosen five card poker hand. We considered this example before in Lecture 2. We wrote

$$|\Omega| = 52 \cdot 51 \cdot 50 \cdot 49 \cdot 48$$

and this is now justified by the multiplication principle. It turned out to be easier to look at the probability that all the cards have different values, which is the complement of C. By the multiplication principle, this event $\overline{C}$ has probability

$$\frac{52 \cdot 48 \cdot 44 \cdot 40 \cdot 36}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} = \frac{2112}{4165} = 0.50708\ldots$$

since after choosing the first card, there are only 48 choices for the next one because it has to be of a different value, and so on. Now $C = \overline{D}$, so by the complement rule,

$$P(C) = 1 - P(\overline{C}) = 0.49292\ldots.$$

So there is almost a fifty percent chance of drawing at least two cards of the same value (this is called a *pair* in poker).

Third Example. Determine the probability of the event B of drawing exactly three cards of the same value in a uniformly chosen five card poker hand from a standard 52 card deck. Now we must determine $|B|$. We are asking: how many ways can we make a five card hand with exactly three cards having the same value? First, let's decide where the three cards of the same value will be in the sequence of five cards. We have to choose a set of three positions for these cards out of five positions. Therefore there are

$$\binom{5}{3}$$

ways to decide where the three cards of the same value will be. There are 13 choices for what their value is. Once we have chosen their value, we have to choose what they actually are (since there are four cards of each value). By the multiplication principle, there are $4 \cdot 3 \cdot 2 = 24$ ways to choose how to put three of the four cards of the same value into the three chosen positions. Finally, we have to choose the remaining two cards in the poker hand to fill the remaining two positions. Those cards must have a different value to the three cards of the same value (otherwise we would have more than three cards of the same value). So there are $48 \cdot 47$ ways to choose those two cards, by the multiplication principle. Finally, by the multiplication principle,

$$|B| = \binom{5}{3} \cdot 13 \cdot 24 \cdot 48 \cdot 47.$$

The probability is

$$P(B) = \frac{|A|}{|\Omega|} = \frac{94}{4165} = 0.022569\ldots.$$

It is important to recognize that the probability of the event E of drawing at least three cards of the same value is a different problem: in fact we cannot choose the three positions for three cards of the same value, since there might be four cards of the same value, in which case we cannot tell which three positions to count. A better approach is to note $E = B \sqcup F$, where F is the event of four of a kind, and therefore by Axiom 3 of probability,

$$P(E) = P(B) + P(F) = \frac{13}{49980} + \frac{94}{4165} = \frac{163}{7140} = 0.022829\ldots$$

We conclude there is roughly a two percent chance of getting at least three cards of the same value.

The Birthday Paradox

jacques@ucsd.edu

Remarks. *These notes should be considered as part of the lectures. For proper treatment of the birthday paradox, the details are written here in full. These notes should be read in conjunction with Lectures 5 and 6, and after the multiplication principle.*

1. The bet. In class I bet that out of the 42 people in attendance, two would have the same birth day and birth month. This may seem a bit of a foolish bet, since there are 365 possibly days in the year, and only 42 people on the class. If there were more than 365 people in the class, then a win would be assured simply because there must be two people with the same birthday. However, with a bit of combinatorics, we can see that in fact there is a more than 90 percent chance that I bet correctly. This is a counter-intuitive fact which is known more generally as the birthday paradox. I put the results of the (correct as it turns out) bet on the course website.

2. Birthdays on any planet. Let us generalize the problem a little bit and write it in terms of probability measure. Suppose we are on a planet with N days in the year, and we want to know the probability that in a sequence of n uniformly chosen people on that planet, at least two have the same birthday. So the sample space Ω is the set of all sequences of n birthdays (there are N^n such sequences so $|\Omega| = N^n$). The probability measure here is the uniform measure on Ω:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{N^n}.$$

for every event $A \subseteq \Omega$. The event we want is the set A of sequences of n different birthdays, because then $P(\overline{A}) = 1 - P(A)$ is the probability that at least two birthdays are the same, by the complement rule.

3. The exact probability. By the multiplication principle,

$$|A| = N \cdot (N-1) \cdot (N-2) \cdot \cdots \cdot (N-n+1).$$

A short way of writing this is using product Π notation:

$$|A| = \prod_{i=1}^{n} (N - i + 1).$$

This means multiply out $N-i+1$ for values of i from 1 to n, and it is very similar to the sum Σ notation which you will have seen before in calculus and earlier in the course. Then

$$P(A) = \frac{|A|}{N^n} = \frac{1}{N^n}\prod_{i=1}^{n}(N-i+1) = \prod_{i=1}^{n}\Big(1-\frac{i-1}{N}\Big).$$

This is the probability that the bet fails (all birthdays different). Note that we divided each of the n terms in the first product by N to get the second product. We want to let $n = 42$ and $N = 365$. However, the product is hard to compute for such large n, so instead we find an upper estimate for the product: that is we show that the product is in fact less than 0.1.

4. A fact from calculus. To find an upper estimate for $P(A)$, we use the following fact from calculus:

Fact 1. *If $x_1, x_2, \ldots, x_n$ are any real numbers, then*

$$\prod_{i=1}^{n}(1-x_i) \le e^{-\sum_{i=1}^{n} x_i}.$$

Proof of Fact 1. Well the right hand side of the inequality is a product

$$e^{-x_1}\cdot e^{-x_2}\cdots e^{-x_n}.$$

The line $y = 1-x$ is tangent to $y = e^{-x}$ at $x = 0$, and otherwise lies below the curve $y = e^{-x}$, and therefore $1-x \le e^{-x}$ for any real number x. Applying this to each e^{-x_i}, we get

$$e^{-x_1}\cdot e^{-x_2}\cdots e^{-x_n} \ge (1-x_1)(1-x_2)\ldots(1-x_n) = \prod_{i=1}^{n}(1-x_i)$$

and this proves Fact 1.

5. Sum of first n integers. The next fact we need is the well-known formula for the sum of the first n integers:

Fact 2. *For a natural number n,*

$$\sum_{i=1}^{n} i = 1+2+\cdots+n = \frac{1}{2}n(n+1).$$

Proof of Fact 1. Gauss had a beautiful way of doing this sum: it is half the area of a rectangle with side lengths n and $n+1$. To see this, multiply the sum by two. Then we can imagine computing the sum in pairs: $1+n$ and then $2+(n-1)$ all the way up to $n+1$. But the pairs all add up to $n+1$, and there are n of them, so twice the given sum is $n(n+1)$, the area of the rectangle shown below. Therefore the sum itself, which is the total area of the grey blocks in the picture, is $n(n+1)/2$, as required.

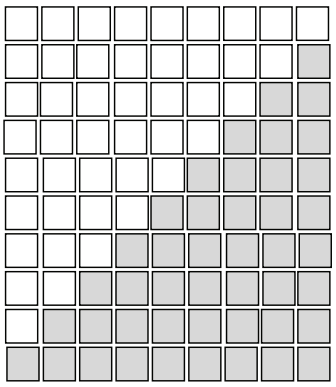


Figure : Gauss' counting

6. The final upper estimate. Using Facts 1 and 2, we find an upper estimate for $P(A)$ as follows (we will call it Fact 3). Let $x_i = (i-1)/N$ for $i = 1, 2, \ldots, n$. Then by Fact 1,

$$P(A) \leq e^{-\sum_{i=1}^{n} x_i} \leq e^{-\frac{1}{N}\sum_{i=1}^{n}(i-1)}$$

Using Fact 2 in the exponent,

$$\sum_{i=1}^{n}(i-1) = \sum_{i=1}^{n-1} i = \frac{1}{2}(n-1)n.$$

Therefore our general upper estimate for $P(A)$ is as follows:

Fact 3. *Let N and n be natural numbers. Then*

$$P(A) \leq e^{-\frac{1}{N}\sum_{i=1}^{n}(i-1)} = e^{-\frac{n(n-1)}{2N}}.$$

7. Back to earth. Consider finally the case of $n = 42$ people and $N = 365$ possible birthdays. By Fact 3,

$$P(A) \leq e^{-\frac{1722}{730}} \leq 0.095\cdots < 0.1.$$

This confirms that there is more than a 90 percent chance that at least 2 out of 42 people have the same birthday, as

$$P(\overline{A}) = 1 - P(A) > 0.9$$

so the original bet had more than a 90 percent chance of being correct.

8. Concluding Remarks. The estimate in Fact 3 says that in general, if n is much more than $\sqrt{N}$ then it is very likely that at least two people will have the same birthday, since $n(n-1)/N$ becomes large when n is much larger than $\sqrt{N}$. In the course, we will be making this kind of asymptotic statement more precise by giving limit theorems. Exercises related to the birthday paradox can be found in Exercise Sheet 2. You may be asked in homework or exams to write down the exact probability for a birthday-type problem (Step 3 above), but you would not be asked to derive (Steps 4–6) the upper estimate. You may be given the upper estimate and then asked to estimate a probability (as in Step 7).

Lectures 7-8

jacques@ucsd.edu

4.1 Conditional Probability

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Suppose that we have prior information which leads us to conclude that an event $A \in \mathcal{F}$ occurs. Based on this information, how does the probability space $(\Omega, \mathcal{F}, P)$ change to reflect this information? This is the notion of conditional probability. Intuitively, we should consider for any event $B \in \mathcal{F}$, given that A occurs, the probability that B occurs is the probability of that part of B which is in A divided by the probability of A. This leads to the definition:

Definition of conditional probability. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A \in \mathcal{F}$ such that $P(A) \neq 0$. Then the conditional probability of an event $B \in \mathcal{F}$ given A is defined by*

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

Note that $(A, \mathcal{G}, Q)$ defines a new probability space, with

$$\mathcal{G} = \{A \cap B : B \in \mathcal{F}\}$$

and $Q(B) = P(B|A)$. To see this, it is not hard to check that $\mathcal{G}$ is again a σ-field, and to see that Q satisfies the axioms of probability, we observe that $Q(A) = P(A \cap A)/P(A) = 1$ (Axiom 1 is satisfied), $Q(B) \geq 0$ for every $B \in \mathcal{G}$ (Axiom 2 is satisfied). To check Axiom 3, suppose we have disjoint events $B, C \in \mathcal{G}$. Then

$$\begin{aligned} Q(B \sqcup C) &= P(B \sqcup C|A) \\ &= \frac{P((B \sqcup C) \cap A)}{P(A)} \\ &= \frac{P(B \cap A) + P(C \cap A)}{P(A)} \\ &= P(B|A) + P(C|A) \quad = \quad Q(A) + Q(B). \end{aligned}$$

The same argument extends to countably many disjoint events $B_1, B_2, \cdots \in \mathcal{G}$. Having verified that $Q(B) = P(B|A)$ is actually a probability measure, we do some computational examples.

First Example. A fair six-sided die is tossed once and the outcome is recorded. Given that the outcome is an odd number, what is the probability that the outcome is 5? Let A be the event that the outcome is odd, and let B be the event that the outcome is 5. Then we are asking for $P(B|A)$. According to the definition,

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

Now $P(A) = 1/2$, and $P(A \cap B) = P(B) = P(\{5\}) = 1/6$. Therefore

$$P(B|A) = \frac{1/6}{1/2} = \frac{1}{3}$$

as expected.

Second Example. A fair six-sided die is tossed twice and both outcomes are recorded. Given that at least one of the numbers is a 6, what is the probability that the total on the dice is at least 10? Let A be the event that at least one of the numbers is a 6, and let B be the event that the total is at least 10. Then we are asking for $P(B|A)$. Now $P(A) = 1 - P(\overline{A})$, where $\overline{A}$ is the event that neither of the dice shows 6. The probability of that event, by the multiplication principle is

$$\frac{|\overline{A}|}{36} = \frac{5 \cdot 5}{36} = \frac{25}{36}.$$

We conclude $P(A) = 11/36$. Next, we compute $P(A \cap B)$. This is the event that at least one of the dice shows a 6 and the total on the dice is at least 10. We can list the outcomes in $A \cap B$:

$$A \cap B = \{(6,4), (6,5), (6,6), (5,6), (4,6)\}$$

and therefore

$$P(A \cap B) = \frac{|A \cap B|}{36} = \frac{5}{36}.$$

So finally,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{5/36}{11/36} = \frac{5}{11}.$$

4.2 Law of total probability

The third axiom of probability gives a law called the law of total probability, which is useful in computing conditional probabilities. It essentially says that if a given event B can be written as

$$B = (B \cap A_1) \sqcup (B \cap A_2) \sqcup \ldots$$

where $A_1, A_2, \ldots$ are events, then we can compute $P(B)$ in terms of the probabilities $P(B|A_i)$ and $P(A_i)$. This allows us to "break down" probability computations into simple pieces.

Law of total probability. *Let $B, A_1, A_2, \ldots$ be events in a probability space and suppose*

$$B = (B \cap A_1) \sqcup (B \cap A_2) \sqcup \ldots.$$

Then

$$P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).$$

Proof ▷ By Axiom 3 of probability,

$$P(B) = \sum_{i=1}^{\infty} P(B \cap A_i).$$

By definition of conditional probability,

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

and so

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

Putting this into the sum,

$$P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).$$

This verifies the law of total probability. ∎

First Example. A red urn and a blue urn each contain black and white balls. The red urn contains as many white balls as black balls, whereas the blue urn contains three times as many white balls as black balls. A player randomly and uniformly picks one of the urns, and then randomly and uniformly picks a ball from that urn. Determine the probability that the ball picked is black. Let A_r be the event that we pick the red urn, and let A_b be the event that we pick the blue urn. Then $P(A_r) = P(A_b) = 1/2$. Let B be the event that the ball picked is black. To determine $P(B)$, note that $B = (A_r \cap B) \sqcup (A_b \cap B)$ and therefore

$$P(B) = P(A_r \cap B) + P(A_b \cap B)$$

by Axiom 3 of probability. We would expect

$$P(A_r \cap B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

since there is half a chance of picking the red urn, and then half a chance that we pick a white ball from the red urn. Similarly,

$$P(A_b \cap B) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}.$$

To make these statements precise, we use the law of total probability:

$$P(B) = P(B|A_r)P(A_r) + P(B|A_b)P(A_b) = \frac{1}{2} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{1}{2} = \frac{7}{12}.$$

The reader should find the problem with the following argument: $P(A_r \cap B) = 1/2$ since $A_r \cap B$ is the event that a black ball is picked from the red urn, and half the balls in the red urn are black; this ignores that A_r is an event and $P(A_r) < 1$. For this reason we apply $P(A_r \cap B) = P(B|A_r)P(A_r)$.

Second Example. Continuing from the last example, suppose that we know that the ball picked is black. What is the probability that the ball was picked from the blue urn? Using the notation of the last example, this is $P(A_b|B)$. By definition of conditional probability,

$$P(A_b|B) = \frac{P(A_b \cap B)}{P(B)} = \frac{1/3}{7/12} = \frac{4}{7}.$$

Third Example. Two decks of cards lie on a table. One of the decks is a standard deck, but the other is not, since every card in it is the ace of spades. A dealer randomly and uniformly picks one of the two decks, and asks you to pick a card from it. The card you pick is the ace of spades. Determine the probability that the deck offered to you is the non-standard deck. This problem is basically the same as the last one: let B be the event that the ace of spades is picked, and let A_s and A_n be the events that the chosen deck is standard and non-standard respectively. Then we want $P(A_n|B)$. We know that

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

Let us compute $P(B)$ using the law of total probability:

$$P(B) = P(B|A_n)P(A_n) + P(B|A_s)P(A_s).$$

Then $P(A_s) = P(A_n) = 1/2$ and $P(B|A_n) = 1$ and $P(B|A_s) = 1/52$. Therefore

$$P(B) = \frac{1}{2} + \frac{1}{104} = \frac{53}{104}.$$

We also have $P(A_n \cap B) = P(B|A_n)P(A_n) = 1/2$, so

$$P(A_n|B) = \frac{P(A_n \cap B)}{P(B)} = \frac{1/2}{53/104} = \frac{52}{53}.$$

So there is a $52/53$ chance that the deck chosen is non-standard.

4.3 Bayes' Rule

Bayes' rule is an exceedingly important but simple to state tool in probability, and appears in numerous applications.

Bayes' Rule. *Let A and B be events in a probability space, such that $P(A) \neq 0$ and $P(B) \neq 0$. Then*

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

Proof ▷ By definition of conditional probability,

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

Also

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

Solving for $P(A \cap B)$, we get

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

and dividing both sides by $P(A)$ gives Bayes' Rule. ∎

The probability $P(B)$ is called the prior probability, and $P(B|A)$ is called the posterior probability. We imagine that an experiment is performed, from which we want to obtain some information about B. The prior probability $P(B)$ is the probability that B occurs, without any further information. The posterior probability is the probability that B occurs, given the information obtained from the experiment, which is the event A. In practice, to compute $P(A)$, we may use the law of total probability $P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$.

First Example. A patient X at a medical exam reveals a symptom x of Type II diabetes. The percentage of all people in the United States with Type II diabetes is 10 percent (the prior probability is 10/100). The percentage of Type II diabetics which are asymptomatic in a medical exam is 0.1 percent and the percentage of Type II diabetics which have symptoms is 80 percent. Determine the probability that patient X has Type II diabetes (the posterior probability). We set up the problem in the framework of conditional probability. Let B be the event that a patient has Type II diabetes. Let A be the event that a patient has symptom x at a medical exam. Then we have from Bayes' Rule that the posterior probability is $P(B|A) = P(A|B)P(B)/P(A)$. Now $P(B|A)$ is the probability we want: it is the probability that given that symptom x is detected, the patient has Type II diabetes. The prior probability is $P(B) = 10/100$, and we have $P(A|B) = 80/100$, and $P(A|\overline{B}) = 1/1000$, since these quantities are given. Therefore

$$P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B}) = \frac{80}{100} \cdot \frac{10}{100} + \frac{1}{1000} \cdot \frac{90}{100} = \frac{809}{10000}.$$

Then

$$P(B|A) = \frac{(80/100) \cdot (10/100)}{(809/10000)} = \frac{800}{809} = 0.98887\ldots.$$

So the patient has roughly a 98.9 percent chance of having Type II diabetes.

Second Example. A student S is suspected of cheating on an exam, due to evidence E of cheating being present. Suppose that in the case of a cheating student, evidence E is present with 60 percent probability, and that in the case of a student that does not cheat, evidence E is present with a 0.01 percent probability. Suppose also that the proportion of students that cheat is 1 percent. Determine the probability that S cheated. Let B be the event S cheated, and let A be the event that evidence E is present. We seek

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

This time we do not know $P(A)$, the probability that evidence E is present. However, by the law of total probability,

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

We know $P(A|B) = 0.6$ and $P(A|\overline{B}) = 0.0001$, and $P(B) = 0.01$, since these are given. Therefore

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.6 \cdot 0.01}{0.6 \cdot 0.01 + 0.0001 \cdot 0.99} = 0.983\ldots.$$

So there is roughly a 98 percent chance the student S cheated.

4.4 Monty-Hall Problem.

The Monty-Hall problem is a good illustration of conditional probability. A game show consists of three closed doors, behind one of which there is a prize. A game show host, who knows which door hides the prize, asks a contestant to pick one of the three doors, and then randomly opens a different door with no prize behind it. it. Then the host asks: would you like to open the door you originally picked (call this *stay* strategy), or would you like to pick the other closed door (i.e. *switch* strategy)?

The first remark is that it does not matter which door you first selected, so we might as well assume you picked door number 1. The question is: do you stick with door number 1 or do you switch to another door? It seems intuitive that it is equally likely that the prize is behind the door 1 and the door that the game show host did not pick. However, this is not the case, since by revealing that there is nothing behind door 2, the game show host increases the likelihood that there is something behind door 3. Thus it seems plausible that to maximize the probability of revealing the prize, you should switch doors. The key is to make this rigorous in terms of probability. Let A_i be the event that the prize lies behind door i. Let B_i be the event that the game show host picks door i. Let B be the event that switching doors leads to a win. Then we have $P(A_i) = 1/3$ for $i = 1, 2, 3$ and also

$$B = (B_3 \cap A_2) \sqcup (B_2 \cap A_3).$$

By the law of total probability,

$$P(B) = P(B_3|A_2)P(A_2) + P(B_2|A_3)P(A_3) = \frac{2}{3}$$

since $P(B_3|A_2) = P(B_2|A_3) = 1$. Let C be the event that keeping the door (staying) wins. Then

$$C = (B_2 \cap A_1) \sqcup (B_3 \cap A_1)$$

and by the law of total probability,

$$P(C) = P(B_2|A_1)P(A_1) + P(B_3|A_1)P(A_1) = \frac{1}{3}$$

since $P(B_2|A_1) = 1/2 = P(B_3|A_1)$. Therefore a win is twice as likely if you switch doors than if you stay.

4.5 Independence

Two events in a probability space are independent if the occurrence of either of them does not affect the probability of the occurrence of the other. In mathematical terms, this means events A and B are independent if

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

By definition of conditional probability, this is the same as $P(A \cap B) = P(A)P(B)$. More generally, we can define independence of a set of events.

Definition of independence. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A_1, A_2, \ldots, A_n$ be events in $\mathcal{F}$. Then $A_1, A_2, \ldots, A_n$ are independent if for any non-empty set $I \subseteq \{1, 2, \ldots, n\}$,*

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i).$$

The definition says that if events are independent, the the probability of their intersection is just the product of all their probabilities. So for instance, $P(A \cap B) = P(A)P(B)$, and $P(A \cap B \cap C) = P(A)P(B)P(C)$. We have been implicitly using the notion of independence. When we say two fair die are tossed independently, we now have the formal definition of what that means: if A is any set of outcomes on the first die and B is any set of outcomes on the second, then

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

In words, the outcome of the first die does not affect the outcome of the second.

First Example. An event A cannot be independent of itself, unless $P(A) = 0$ or $P(A) = 1$. To see this, if A is independent of A, then

$$P(A \cap A) = P(A)P(A) = P(A)^2$$

by definition. But $P(A \cap A) = P(A)$, so we get

$$P(A) = P(A)^2$$

which is only possible if $P(A) = 0$ or $P(A) = 1$. Similarly, one can show that if A and B are disjoint and independent, then $P(A)$ or $P(B)$ is 0, and one can also show that

if A and B are independent, then so are $\overline{A}$ and $\overline{B}$, using inclusion-exclusion and the complement rule.

Second Example. A coin is tossed independently n times. Suppose the probability of heads on each toss is p. What is the probability that no heads occur? Let A_i be the event of tails on the ith toss. Then we want

$$P(A_1 \cap A_2 \cap \cdots \cap A_n).$$

By independendence, this is $P(A_1)P(A_2)\dots P(A_n)$. Now by the complement rule, $P(A_i) = 1 - p$, so

$$P(\text{no heads}) = (1-p)^n.$$

We will come back to this example when we compute the probability of exactly k heads in n coin tosses; this is the famous binomial distribution.

Third Example. A dart is thrown uniformly and randomly at a circular simplified dartboard consisting of eight sectors with equal area, labelled clockwise $1, 2, 3, 4$. Let A be the event that the dart lands in sectors $1, 2$, let B be the event that the dart lands in sectors $2, 3$, and let C be the event that the dart lands in sectors $1, 3$. Are the events A, B, C independent? To check independence, we have to check $P(A)P(B) = P(A \cap B)$, $P(B)P(C) = P(B \cap C)$, $P(A)P(C) = P(A \cap C)$ and $P(A)P(B)P(C) = P(A \cap B \cap C)$. Now $P(A) = P(B) = P(C) = 1/2$, since they each have 2 sectors, and since each pairwise intersection $A \cap B, B \cap C, A \cap C$ covers 1 sector, $P(A \cap B) = P(B \cap C) = P(A \cap C) = 1/4$. So any two of the events A, B, C are independent. However, they are not independent together, since $A \cap B \cap C = \emptyset$ and so

$$P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C) = \frac{1}{8}.$$

The last example illustrates that to check independence of a set of events, independence of all combinations of at least two of those sets must also be checked. Finally, we give an example which tells us that if we repeat an experiment independently many times, then provided there is a positive chance the experiment succeeds, the experiment is very likely to succeed after a few tries.

Fourth Example. An experiment is performed, and it is known that the probability of success of the experiment is 0.1. Suppose the experiment is repeated n times. How large must n be in order to guarantee that the experiment succeeds at least once in n

times with more than 90 percent certainty? Let A_i be the event that the experiment succeeds on the ith attempt. Since the A_i are independent, the probability that in n trials the experiment fails every time is

$$P(\overline{A}_1)P(\overline{A}_2)\cdots P(\overline{A}_n) = (1-0.1)^n = 0.9^n.$$

So we want to ensure $0.9^n < 0.001$, by the complement rule. A calculator shows this happens already when $n = 22$. So in $n = 22$ trials, the experiment has more than a 90 percent chance of succeeding on at least one of the trials. More generally, if an experiment has probability p of succeeding in one trial, then the probability that it succeeds at least once in n trials is

$$1 - P(\overline{A}_1)P(\overline{A}_2)\cdots P(\overline{A}_n) = 1 - (1-p)^n.$$

How large should n be to guarantee at least a fifty percent chance of success? For this we solve

$$1 - (1-p)^n \geq \frac{1}{2}.$$

Taking logs and solving for n, we get

$$n \geq \frac{-\ln 2}{\ln(1-p)}.$$

So in this many trials we have at least a fifty percent chance of success.

Lectures 9-10

jacques@ucsd.edu

5.1 Random Variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space. The *Borel sets* in $\mathbb{R}$ are the sets in the smallest σ-field on $\mathbb{R}$ that contains all countable unions and complements of intervals. This σ-field is called the *Borel σ-field* on $\mathbb{R}$. If $X : \Omega \to \mathbb{R}$ is a function, then for a set $S \subseteq \mathbb{R}$, let $X^{\dashv}(S) = \{\omega \in \Omega : X(\omega) \in S\}$. In other words, this is the *inverse image* of S – the set of all things in Ω which got mapped by X into S.

Definition of random variables. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable is a function $X : \Omega \to \mathbb{R}$ such that for any Borel set $B \subset \mathbb{R}$, $X^{\dashv}(B) \in \mathcal{F}$.*

We do not spend time here on the subtleties of Borel sets, except to remark that there are subsets of $\mathbb{R}$ that are not Borel sets. Loosely speaking, a random variable assigns to each outcome $\omega \in \Omega$ a number, such that sets such as $\{\omega \in \Omega : a \leq X(\omega) \leq b\}$ are events. Throughout these notes, for a Borel set $B \subset \mathbb{R}$, the event $\{\omega \in \Omega : X(\omega) \in B\}$ is abbreviated to $[X \in B]$, and we write $P(X \in B)$ instead of $P([X \in B])$. So for instance $P(X \leq x)$ really means $P(\{\omega \in \Omega : X(\omega) \leq x\})$. Some examples of random variables are given below.

Example. Two fair dice are tossed and the two outcomes recorded. As is familiar, we have

$$\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$$

consisting of 36 possible outcomes. Now in this case, any function $X : \Omega \to \mathbb{R}$ could be a random variable since we take $\mathcal{F}$ in this case is all subsets of Ω. Let's consider the two random variables

$$X((i,j)) = i + j \quad Y((i,j)) = ij.$$

The first is the sum of the two dice, which we have considered before. For instance

$$[X = 3] = \{\omega \in \Omega : X(\omega) = 3\} = \{(i,j) \in \Omega : i + j = 3\} = \{(1,2), (2,1)\}.$$

It follows that

$$P(X = 3) = \frac{2}{36} = \frac{1}{18}.$$

The reader can check for example that

$$P(X = 12) = \frac{1}{36} \quad P(X = 11) = \frac{1}{18}.$$

Now consider $[Y = 12]$. This is the same as

$$\{\omega \in \Omega : Y(\omega) = 12\} = \{(i, j) \in \Omega : ij = 12\} = \{(3, 4), (4, 3), (2, 6), (6, 2)\}$$

and so

$$P(Y = 12) = \frac{4}{36} = \frac{1}{9}.$$

As another example, we observe $[Y = 7] = \emptyset$, since $ij = 7$ has no solution with $(i, j) \in \Omega$. Therefore $P(Y = 7) = 0$.

5.2 Cumulative Distribution Functions

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let X be a random variable on Ω. Then the probabilities $P(X \leq x)$ for $x \in \mathbb{R}$ can be considered as a function $F(x)$ where $F : \mathbb{R} \to [0, 1]$. If we can find an explicit formula for $F(x)$, then we can work with all the probabilities at once, and introduce the tools of calculus to analyse $F(x)$. In many practical examples, one can actually determine $F(x)$.

Definition of cdf. *Let $(\Omega, \mathcal{F}, P)$ be any probability space, and let X be a random variable on Ω. Then the cumulative distribution function of X or* cdf *is the function $F : \mathbb{R} \to [0, 1]$ defined by $F(x) = P(X \leq x)$ for $x \in \mathbb{R}$.*

Sometimes the cdf is also called the *distribution function*. A random variable whose cdf is continuous will be referred to as a *continuous random variable*. In most of our work on uncountable probability spaces, we deal with continuous random variables. The cdf has some simple properties, from the definition and the axioms of probability:

Proposition 1 *Let F be the cdf of a random variable X. Then F is a non-decreasing function and $\lim_{x\to\infty} F(x) = 1$.*

Axiom 1 of probability gives $P(\Omega) = 1$. So intuitively,

$$\lim_{x\to\infty} F(x) = \lim_{x\to\infty} P(X \leq x) = P(\Omega) = 1.$$

For the first statement, that F is non-decreasing, it follows from the fact that $[X \leq x] \subseteq [X \leq y]$ for $x \leq y$, and therefore $P(X \leq x) \leq P(X \leq y)$ and so $F(x) \leq F(y)$ for $x \leq y$. A useful point to remember is that

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

In the next example we give a specific instance of *deriving* a cdf from the definition above for a given random variable. In practice, however, we are going to be given the cdf of a random variable and then work with it to compute various statistics.

First Example. The *uniform measure* on $\Omega = [0, 1]$ is the probability measure defined by taking, for an interval A of length $b - a$, for instance $A = [a, b)$, $P(A) = b - a$, and then using the axioms of probability to define $P(A)$ for all events in $[0, 1]$. This measure is sometimes called the *Lebesgue measure*. We imagine this as picking uniformly a point $y \in [0, 1]$, and then an event $A \subseteq [0, 1]$ is just the event that the picked point is in A. Let $X : [0, 1] \to \mathbb{R}$ be the numerical value of the picked point y. Then the cdf of X is, by definition,

$$F(x) = P(X \leq x) = P([0, x]) = x$$

for $x \in [0, 1]$. If $x \geq 1$ then $F(x) = 1$ and if $x \leq 0$ then $F(x) = 0$. This cdf is graphed below, and we conclude that X is a continuous random variable since the cdf is continuous.

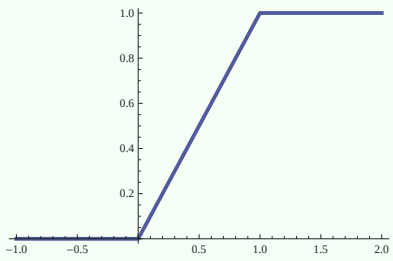


Figure 1 : The cdf of X

This example illustrates that we have to be very careful to define the value of $F(x)$ for all $x \in \mathbb{R}$; we had to take care separately of the facts that when $x \leq 0$, $F(x) = 0$, and that when $x \geq 1$, $F(x) = 1$.

Second Example. Consider the same probability space as in the first example, but now consider the random variable $Y = X^2$ – the square of the number picked. Then for $0 \leq x \leq 1$,

$$F(x) = P(Y \leq x) = P([0, \sqrt{x}]) = \sqrt{x}.$$

For $x \geq 1$, $F(x) = 1$ and for $x \leq 0$, $F(x) = 0$. So Y is also continuous, and the cdf of Y is graphed below:

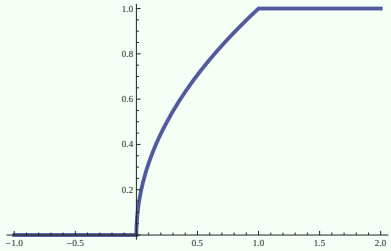


Figure 2 : The cdf of $Y = X^2$

5.2 Probability Density Functions

We might ask why we did not rather consider the function $P(X = x)$ instead of $P(X \leq x)$. One reason is (as in the last two examples) that for uncountable probability spaces in general, $P(X = x) = 0$, so this does not give the same valuable information as the cdf $F(x)$. However, it would still be informative to know how the mass of the probability measure is distributed relative to a given random variable. Informally, the *probability density function* indicates how the probability measure relative to the random variable X is spread out, and is defined as follows:

Definition of pdf. *Let $(\Omega, \mathcal{F}, P)$ be any probability space, let X be a random variable on Ω, and let $f : \mathbb{R} \to \mathbb{R}$ be a function. Then f is a probability density function or pdf of X if and only if for every $a, b \in \mathbb{R}$,*

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

The pdf is sometimes referred to as the *density function*. So the pdf of a random variable is a function whose integral gives the various probabilities $P(a \leq X \leq b)$. Some very important properties of pdfs are listed below:

Proposition 2 *Let f be a pdf of random variable X, and let F be the cdf of X. Then*

1. $\int_{-\infty}^{\infty} f(x)dx = 1.$
2. *If f is differentiable at x, then $f(x) = F'(x)$.*
3. $F(x) = \int_{-\infty}^{x} f(x)dx.$

The second statement gives an explicit way to find the cdf from the pdf. We should remark that if F is smooth enough – technically F must be absolutely continuous – then f is unique up to a subset of $\mathbb{R}$ of Lebesgue measure 0.

Example. Let X be a randomly and uniformly chosen real number in $[0, 1]$. In a preceding example, we saw

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Now F is differentiable at every $x \notin \{0, 1\}$, and it is *not* differentiable at $x = 0$ and $x = 1$. So for $x \notin \{0, 1\}$, the pdf $f(x)$ is defined, and equals $F'(x)$ by the proposition above. For $0 < x < 1$, $F'(x) = 1$, and for $x < 0$ or $x > 1$, $F'(x) = 0$. Therefore

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \text{ or } x > 1 \\ 1 & \text{if } 0 < x < 1 \end{cases}$$

This pdf is shown in the figure below. It reflects very well the uniform probability measure since all the mass in on $[0, 1]$ and it is constant there.

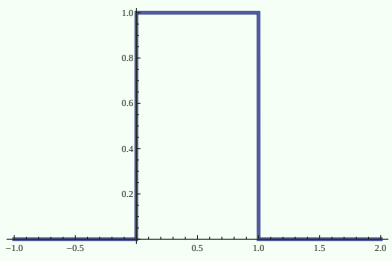


Figure 3 : Pdf of a uniform random variable

5.3 Probability Mass Functions

A *discrete random variable* is a random variable X such that for some countable set S,

$$P(X \in S) = \sum_{x \in S} P(X = x) = 1.$$

In other words, all the mass of X is concentrated on countable many points. Then X cannot have a pdf, since if $P(X = x) > 0$ where $x \in S$, then also

$$P(X = x) = P(x \leq X \leq x) = \int_x^x f(x)dx = 0$$

which is a contradiction. So discrete random variables *cannot have a pdf.*

We have mostly encountered discrete random variables so far, for example where X is the sum of the values on two dice, or X counts the number of heads in coin tosses, or X is an element of $\mathbb{N}$, and so on. Since no pdf exists for these, it makes sense rather to consider the *probability mass function* or *pmf*, namely

$$f(x) = P(X = x) \quad \text{for } x \in \mathbb{R}.$$

Given the probability mass function f, we have

$$F(x) = \sum_{y=-\infty}^{x} f(y)$$

and also

$$\sum_{x=-\infty}^{\infty} f(x) = 1.$$

In the case that X is discrete, the cdf $F(x)$ is not a continuous function: it jumps by $P(X = x)$ at each $x \in S$. We will soon give classical examples of discrete random variables.

5.3 Expectation

If X is a random variable, we can ask for the average or mean value of X. For example, if we toss a fair coin and score 1 for heads and 0 for tails, and let X be the score, then the average value of X is clearly $\frac{1}{2}$. In this section, we define this average value for any probability space and for both discrete and continuous random variables.

Definition of expectation. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. If X is a continuous random variable with pdf f, then the* expectation *of X is*

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

If X is a discrete random variable with pmf f, then the expectation of X is

$$E(X) = \sum_{x=-\infty}^{\infty} x f(x) dx.$$

First Example. Let X denote the numerical outcome when a fair die is tossed. To determine $E(X)$, we note that X is discrete, and has pmf $f(x) = P(X = x) = 1/6$ for $x \in \{1, 2, 3, 4, 5, 6\}$ and $f(x) = 0$ otherwise. So

$$E(X) = \sum_{x=-\infty}^{\infty} x f(x) = \sum_{x=1}^{6} x f(x) = \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}.$$

So we conclude the average value when tossing a fair die is 3.5.

Second Example. Let X denote the value of a uniformly chosen number in $[0, 1]$. Then the cdf $F(x)$ was determined in a preceding example, and it is continuous, so X is a *continuous random variable.* We also determined

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \text{ or } x > 1 \\ 1 & \text{if } 0 < x < 1 \end{cases}$$

This is the derivative of F. By definition,

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{1} x dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}.$$

So the average value of a uniformly chosen number in $[0, 1]$ is $1/2$, as expected.

5.4 Classical Discrete Distributions.

A. The Bernouilli Distribution. Let X be a random variable such that $P(X = 1) = p$ and $P(X = 0) = 1 - p$ (so X is a discrete random variable). Then X is said to have

the *Bernouilli distribution* with success probability p. We can imagine this as tossing a coin with heads probability p, and scoring 1 if the coin lands heads, and 0 if it lands tails. Thus X records the score. The probability mass function defining the Bernouilli distribution is

$$f(x) = p^x(1-p)^{1-x} \quad \text{for } x \in \{0,1\}.$$

Note that this is a probability mass function since

$$\sum_{x=-\infty}^{\infty} f(x) = p^0(1-p)^1 + p^1(1-p)^0 = p + 1 - p = 1.$$

The mean of a Bernouilli random variable is

$$E(X) = \sum_{x=-\infty}^{\infty} xf(x) = 0 \cdot (1-p) + 1 \cdot p = p.$$

This makes sense: the average number of heads is p if the coin has heads probability p in one toss. We write $X \sim \text{BER}(p)$ to represent that X has the Bernouilli distribution with probability p.

B. The Binomial Distribution. We now look at what happens if the coin above is tossed n times independently. Surely the average number of heads in n coin tosses is pn, and we shall now check this mathematically. For this, we need the *binomial theorem*. This gives a general way of expanding binomials like $(a+b)^2 = a^2 + 2ab + b^2$, $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$, $(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$, and so on.

Proposition 3 (Binomial Theorem). *Let $n \in \mathbb{N}$ and let $a, b \in \mathbb{R}$. Then*

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

Proof ▷ If $a = b = 0$, this is trivial. Suppose say $b \neq 0$. If we divide both sides by b^n, and let $y = a/b$, we get

$$(1+y)^n = \sum_{x=0}^{n} \binom{n}{x} y^x.$$

If we multiply out the n brackets on the left, what is the coefficient of y^x? Well this means that from x of the brackets we picked y (call these special brackets) and from the others we picked 1. How many ways can we choose the x special brackets out of the n brackets? From our work on counting subsets, it is exactly $\binom{n}{x}$. In other words, y^x must appear $\binom{n}{x}$ times for each x. This matches the y^x term on the right, so we are done. ∎

Now we define the binomial distribution. Let X be a random variable with pmf

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $x \in \{0, 1, 2, \ldots, n\}$. Then X is a discrete random variable that is said to have the *binomial distribution* with success probability p in n trials. A plot of f is shown below:

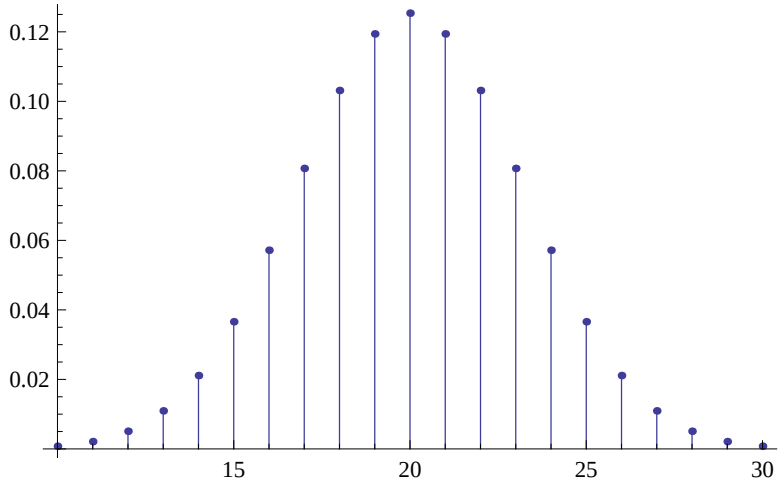


Figure 4 : The binomial distribution $X \sim \text{BIN}(40, 1/2)$

To check f is indeed a pmf, we use the binomial theorem with $a = p$ and $b = 1 - p$.

$$\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1.$$

We can imagine this as tossing a coin n times independently, with heads probability p on each trial, and then X is the number of heads obtained: there are $\binom{n}{x}$ ways to choose on which turns the x heads occur, then the chance they occurred there is p^x, and the chance that everything else is tails is $(1-p)^{n-x}$. There is no particularly nice formula for the cdf, which is

$$F(x) = \sum_{y \le x} \binom{n}{y} p^y (1-p)^{n-y}.$$

Finally, we compute the expectation carefully. There is a clever way to do it using derivatives. Consider the sum

$$g(z) = \sum_{x=0}^{n} \binom{n}{x} z^x (1-p)^{n-x}.$$

By the binomial theorem, $g(z) = (z + 1 - p)^n$ and so

$$g'(z) = n(z + 1 - p)^{n-1}.$$

On the other hand, as a function of z, g is a polynomial (infinite radius of convergence), so the derivative trick applies and

$$g'(z) = \sum_{x=0}^{n} \binom{n}{x} x z^{x-1} (1-p)^{n-x}$$

for all $z \in \mathbb{R}$. Now let $z = p$. Then the two formulas we just found for $g'(z)$ with $z = p$ give

$$\sum_{x=0}^{n} \binom{n}{x} x p^{x-1} (1-p)^{n-x} = n.$$

Now the reason we did this is that

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x} \\ &= p \sum_{x=1}^{n} \binom{n}{x} x p^{x-1} (1-p)^{n-x} \\ &= pn. \end{aligned}$$

So indeed the expected number of heads in n coin tosses is pn. We write $X \sim \textsc{Bin}(n, p)$ to represent that X has the Binomial distribution with probability p in n trials.

Two important technical things were done here: first, using the binomial theorem, and second using a derivative trick to determine the expectation. This is not an isolated phenomenon: we will see that often the expectation can be computed using this derivative trick. So let's state it as a proposition:

Proposition 4 (Derivative of power series). *Let $g(z)$ denote the power series*

$$\sum_{n=0}^{\infty} a_n z^n$$

and suppose that g has radius of convergence $\rho > 0$. Then for $|z| < \rho$,

$$g'(z) = \sum_{n=0}^{\infty} n a_n z^{n-1}.$$

Here is another example where it works, on a geometric series. Let's recall the formula for a geometric series:

Proposition 5 (Geometric series). *Let $m \in \mathbb{N} \cup \{0\}$ and let $|z| < 1$. Then*

$$\sum_{x=m}^{\infty} z^x = \frac{z^m}{1-z}.$$

C. The Geometric Distribution. Let X be a random variable with pmf

$$f(x) = p(1-p)^{x-1}$$

where $x \in \mathbb{N}$ and $p \in [0,1]$, and $f(x) = 0$ otherwise. To check that f is a valid pmf, we use the formula for a geometric series with $m = 0$ and $z = 1 - p$:

$$\begin{aligned} \sum_{x=1}^{\infty} f(x) &= \sum_{x=1}^{\infty} p(1-p)^{x-1} \\ &= p\sum_{x=0}^{\infty} (1-p)^x \\ &= \frac{p}{1-(1-p)} = 1. \end{aligned}$$

We can imagine X to be the turn at which a coin first shows heads when it is tossed repeatedly and independently, and the probability of heads on each turn is p. For then if $X = x$, we have to get $x-1$ tails (probability is $(1-p)^{x-1}$) followed by heads (probability p). Finally, we can compute the expectation:

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

Let for $|z| < 1$,

$$g(z) = \sum_{x=1}^{\infty} z^x.$$

Then $g(z) = z/(1-z)$ since it is a geometric series. Using the derivative trick,

$$g'(z) = \frac{1}{(1-z)^2} = \sum_{x=1}^{\infty} xz^{x-1}.$$

Note that this is valid for $|z| < 1$, since the radius of convergence of the power series is 1. Now put $z = (1-p)$, so that

$$E(X) = pg'(1-p) = p \cdot \frac{1}{(1-(1-p))^2} = \frac{p}{p^2} = \frac{1}{p}.$$

So the expected number of coin tosses up to and including the first heads is $1/p$. We write $X \sim \textsc{Geo}(p)$ to represent that X has the geometric distribution with probability p.