V PROBABILITY

In its simplest form, computing the probability reduces to counting, namely the lucky outcomes and all possible outcomes. The probability is then the ratio of the two, which is a real number between zero and one.

15 Inclusion-Exclusion

Today, we introduce basic concepts in probability theory and we learn about one of its fundamental principles.

Throwing dice. Consider a simple example of a probabilistic experiment: throwing two dice and counting the total number of dots. Each die has six sides with 1 to 6 dots. The result of a throw is thus a number between 2 and 12. There are 36 possible outcomes, 6 for each die, which we draw as the entries of a matrix; see Figure 15.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

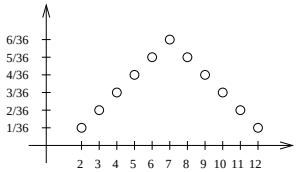

Figure 15: Left: the two dice give the row and the column index of the entry in the matrix. Right: the most likely sum is 7, with probability $\frac{1}{6}$, the length of the diagonal divided by the size of the matrix.

Basic concepts. The set of possible outcomes of an experiment is the *sample space*, denoted as Ω. A possible outcome is an *element*, $x \in \Omega$. A subset of outcomes is an *event*, $A \subseteq \Omega$. The *probability* or *weight* of an element x is $P(x)$, a real number between 0 and 1. For finite sample spaces, the *probability* of an event is $P(A) = \sum_{x \in A} P(x)$.

For example, in the two dice experiment, we set $\Omega = \{2, 3, \ldots, 12\}$. An event could be to throw an even number. The probabilities of the different outcomes are given in Figure 15 and we can compute

$$P(\text{even}) \quad = \quad \frac{1+3+5+5+3+1}{36} \quad = \quad \frac{1}{2}.$$

More formally, we call a function $P : \Omega \to \mathbb{R}$ a *probability distribution* or a *probability measure* if

(i) $P(x) \geq 0$ for every $x \in \Omega$;

(ii) $P(A \dot\cup B) = P(A) + P(B)$ for all disjoint events $A \cap B = \emptyset$;

(iii) $P(\Omega) = 1$.

A common example is the *uniform probability distribution* defined by $P(x) = P(y)$ for all $x, y \in \Omega$. Clearly, if Ω is finite then

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

for every event $A \subseteq \Omega$.

Union of non-disjoint events. Suppose we throw two dice and ask what is the probability that the outcome is even or larger than 7. Write A for the event of having an even number and B for the event that the number exceeds 7. Then $P(A) = \frac{1}{2}$, $P(B) = \frac{15}{36}$, and $P(A \cap B) = \frac{9}{36}$. The question asks for the probability of the union of A and B. We get this by adding the probabilities of A and B and then subtracting the probability of the intersection, because it has been added twice,

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

which gives $\frac{6}{12} + \frac{5}{12} - \frac{3}{12} = \frac{2}{3}$. If we had three events, then we would subtract all pairwise intersections and add back in the triplewise intersection, that is,

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

Principle of inclusion-exclusion. We can generalize the idea of compensating by subtracting to n events.

PIE THEOREM (FOR PROBABILITY). The probability of the union of n events is

$$P(\bigcup_{i=1}^{n} A_i) \quad = \quad \sum_{k=1}^{n} (-1)^{k+1} \sum P(A_{i_1} \cap \ldots \cap A_{i_k}),$$

where the second sum is over all subsets of k events.

PROOF. Let x be an element in $\bigcup_{i=1}^{n} A_i$ and H the subset of $\{1, 2, \ldots, n\}$ such that $x \in A_i$ iff $i \in H$. The contribution of x to the sum is $P(x)$ for each odd subset of H and $-P(x)$ for each even subset of H. If we include $\emptyset$ as an even subset, then the number of odd and even subsets is the same. We can prove this using the Binomial Theorem:

$$(1-1)^n \quad = \quad \sum_{i=0}^{n} (-1)^i \binom{n}{i}.$$

But in the claimed equation, we do not account for the empty set. Hence, there is a surplus of one odd subset and therefore a net contribution of $P(x)$. This is true for every element. The PIE Theorem for Probability follows. □

Checking hats. Suppose n people get their hats returned in random order. What is the chance that at least one gets the correct hat? Let A_i be the event that person i gets the correct hat. Then

$$P(A_i) \quad = \quad \frac{(n-1)!}{n!} \quad = \quad \frac{1}{n}.$$

Similarly,

$$P(A_{i_1} \cap \ldots \cap A_{i_k}) \quad = \quad \frac{(n-k)!}{n!}.$$

The event that at least one person gets the correct hat is the union of the A_i. Writing $P = P(\bigcup_{i=1}^n A_i)$ for its probability, we have

$$\begin{aligned} P &= \sum_{i=1}^{k} (-1)^{k+1} \sum P(A_{i_1} \cap \ldots \cap A_{i_k}) \\ &= \sum_{i=1}^{k} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \\ &= \sum_{i=1}^{k} (-1)^{k+1} \frac{1}{k!} \\ &= 1 - \frac{1}{2} + \frac{1}{3!} - \ldots \pm \frac{1}{n!}. \end{aligned}$$

Recall from Taylor expansion of real-valued functions that $e^x = 1 + x + x^2/2 + x^3/3! + \ldots$. Hence,

$$P \quad = \quad 1 - e^{-1} = 0.6\ldots$$

Inclusion-exclusion for counting. The principle of inclusion-exclusion generally applies to measuring things. Counting elements in finite sets is an example.

PIE THEOREM (FOR COUNTING). For a collection of n finite sets, we have

$$|\bigcup_{i=1}^{n} A_i| \quad = \quad \sum_{k=1}^{n} (-1)^{k+1} \sum |A_{i_1} \cap \ldots \cap A_{i_k}|,$$

where the second sum is over all subsets of k events.

The only difference to the PIE Theorem for probability is we count one for each element, x, instead of $P(x)$.

Counting surjective functions. Let M and N be finite sets, and $m = |M|$ and $n = |N|$ their cardinalities. Counting the functions of the form $f : M \to N$ is easy. Each $x \in M$ has n choices for its image, the choices are independent, and therefore the number of functions is n^m. How many of these functions are surjective? To answer this question, let $N = \{y_1, y_2, \ldots, y_n\}$ and let A_i be the set of functions in which y_i is not the image of any element in M. Writing A for the set of all functions and S for the set of all surjective functions, we have

$$S \quad = \quad A - \bigcup_{i=1}^{n} A_i.$$

We already know $|A|$. Similarly, $|A_i| = (n-1)^m$. Furthermore, the size of the intersection of k of the A_i is

$$|A_{i_1} \cap \ldots \cap A_{i_k}| \quad = \quad (n-k)^m.$$

We can now use inclusion-exclusion to get the number of functions in the union, namely,

$$|\bigcup_{i=1}^{n} A_i| \quad = \quad \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} (n-k)^m.$$

To get the number of surjective functions, we subtract the size of the union from the total number of functions,

$$|S| \quad = \quad \sum_{i=0}^{n} (-1)^k \binom{n}{k} (n-k)^m.$$

For $m < n$, this number should be 0, and for $m = n$, it should be $n!$. Check whether this is indeed the case for small values of m and n.

16 Conditional Probability

If we have partial information, this effectively shrinks the available sample space and changes the probabilities. We begin with an example.

Monty Hall show. The setting is a game show in which a prize is hidden behind one of three curtains. Call the curtains X, Y, and Z. You can win the prize by guessing the right curtain.

STEP 1. You choose a curtain.

This leaves two curtains you did not choose, and at least one of them does not hide the prize. Monty Hall opens this one curtain and this way demonstrates there is no prize hidden there. Then he asks whether you would like to reconsider. Would you?

STEP 2A. You stick to your initial choice.

STEP 2B. You change to the other available curtain.

Perhaps surprisingly, Step 2B is the better strategy. As shown in Figure 16, it doubles your chance to win the prize.

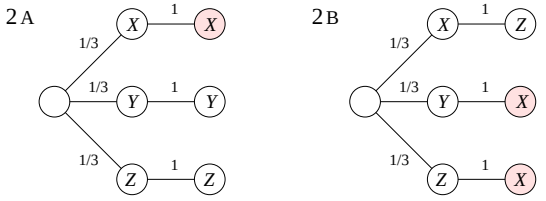

Figure 16: Suppose the prize is behind curtain X. The chance of winning improves from $\frac{1}{3}$ in 2A to $\frac{2}{3}$ in 2B.

Formalization. We are given a sample space, Ω, and consider two events, $A, B \subseteq \Omega$. The *conditional probability* of even A given event B is

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

We illustrate this definition in Figure 17. If we know that the outcome of the experiment is in B, the chance that it is also in A is the fraction of B occupied by $A \cap B$. We say A and B are *independent* if knowing B does not change the probability of A, that is,

$$P(A \mid B) \quad = \quad P(A).$$

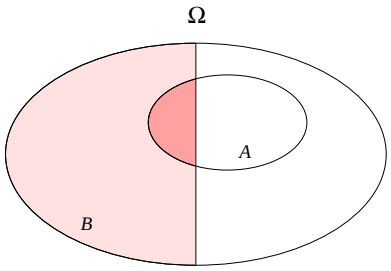

Figure 17: Assuming B, the probability of A is represented by the fraction of the shaded region, B, that is dark shaded, $A \cap B$.

Since $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = P(A)$, we have

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

We thus see that independence is symmetric. However, it fails to be an equivalence relation because it is neither reflexive not transitive. Combining the definition of conditional probability with the condition of independence, we get a formula for the probability of two events occurring at the same time.

PRODUCT PRINCIPLE FOR INDEPENDENT PROB. If A and B are independent then $P(A \cap B) = P(A) \cdot P(B)$.

Trial processes. In many situations, a probabilistic experiment is repeated, possibly many times. We call this a *trial process*. It is *independent* if the i-th trial is not influenced by the outcomes of the preceding $i-1$ trials, that is,

$$P(A_i \mid A_1 \cap \ldots \cap A_{i-1}) \quad = \quad P(A_i),$$

for each i.

An example is picking a coin from an bag that contains one nickel, two dimes, and two quarters. We have an independent trial process if we always return the coin before the next draw. The choice we get a quarter is therefore $\frac{2}{5}$ each time. The chance to pick the quarter three times in a row is therefore $(\frac{2}{5})^3 = \frac{8}{125} = 0.064$. More generally, we have the

INDEPENDENT TRIAL THEOREM. In an independent trial process, the probability of a sequence of outcomes, $a_1, a_2, \ldots, a_n$, is $P(a_1) \cdot P(a_2) \cdot \ldots \cdot P(a_n)$.

Trial processes that are not independent are generally more complicated and we need more elaborate tools to

compute the probabilities. A useful such tool is the tree diagram as shown in Figure 18 for the coin picking experiment in which we do not replace the picked coins.

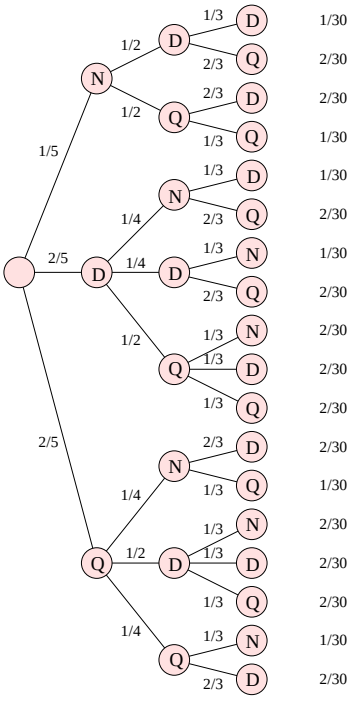

Figure 18: What is the probability of picking the nickel in three trials?

Medical test example. Probabilities can be counterintuitive, even in situations in which they are important. Consider a medical test for a disease, D. The test mostly gives the right answer, but not always. Say its false-negative rate is 1% and its false-positive rate is 2%, that is,

$$\begin{aligned} P(\text{y} \mid D) &= 0.99; \\ P(\text{n} \mid D) &= 0.01; \\ P(\text{y} \mid \neg D) &= 0.02; \\ P(\text{n} \mid \neg D) &= 0.98. \end{aligned}$$

Assume that the chance you have disease D is only one in a thousand, that is, $P(D) = 0.001$. Now you take the test and the outcome is positive. What is the chance that you have the disease? In other words, what is $P(D \mid \text{y})$? As illustrated in Figure 19,

$$P(D \mid \text{y}) = \frac{P(D \cap \text{y})}{P(\text{y})} = \frac{0.00099}{0.02097} = 0.047\ldots.$$

This is clearly a case in which you want to get a second opinion before starting a treatment.

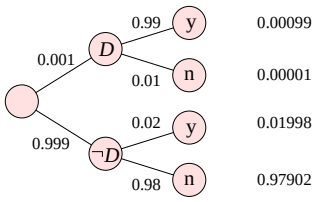

Figure 19: Tree diagram showing the conditional probabilities in the medical test question.

Summary. Today, we learned about conditional probabilities and what it means for two events to be independent. The Product Principle can be used to compute the probability of the intersection of two events, if they are independent. We also learned about trial processes and tree diagrams to analyze them.

17 Random Variables

A *random variable* is a real-value function on the sample space, $X : \Omega \to \mathbb{R}$. An example is the total number of dots at rolling two dice, or the number of heads in a sequence of ten coin flips.

Bernoulli trial process. Recall that an independent trial process is a sequence of identical experiments in which the outcome of each experiment is independent of the preceding outcomes. A particular example is the *Bernoulli trial process* in which the probability of success is the same at each trial:

$$\begin{aligned} P(\text{success}) &= p; \\ P(\text{failure}) &= 1-p. \end{aligned}$$

If we do a sequence of n trials, we may define X equal to the number of successes. Hence, Ω is the space of possible outcomes for a sequence of n trials or, equivalently, the set of binary strings of length n. What is the probability of getting exactly k successes? By the Independent Trial Theorem, the probability of having a sequence of k successes followed by $n-k$ failures is $p^k(1-p)^{n-k}$. Now we just have to multiply with the number of binary sequences that contain k successes.

BINOMIAL PROBABILITY LEMMA. The probability of having exactly k successes in a sequence of n trials is $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$.

As a sanity check, we make sure that the probabilities add up to one. Using the Binomial Theorem, get

$$\sum_{k=0}^{n} P(X=k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k},$$

which is equal to $(p+(1-p))^n = 1$. Because of this connection, the probabilities in the Bernoulli trial process are called the *binomial probabilities*.

Expectation. The function that assigns to each $x_i \in \mathbb{R}$ the probability that $X = x_i$ is the *distribution function* of X, denoted as $f : \mathbb{R} \to [0,1]$; see Figure 20. More formally, $f(x_i) = P(A)$, where $A = X^{-1}(x_i)$. The *expected value* of the random variable is $E(X) = \sum_i x_i P(X = x_i)$.

As an example, consider the Bernoulli trial process in which X counts the successes in a sequence of n trials,

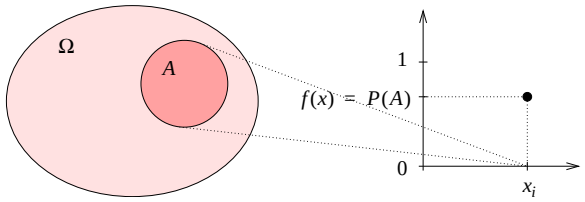

Figure 20: The distribution function of a random variable is constructed by mapping a real number, x_i, to the probability of the event that the random variable takes on the value x_i.

that is, $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$. The corresponding distribution function maps k to the probability of having k successes, that is, $f(k) = \binom{n}{k} p^k (1-p)^{n-k}$. We get the expected number of successes by summing over all k.

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} k f(k) \\ &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-k-1}. \end{aligned}$$

The sum in the last line is equal to $(p+(1-p))^{n-1} = 1$. Hence, the expected number of successes is $X = np$.

Linearity of expectation. Note that the expected value of X can also be obtained by summing over all possible outcomes, that is,

$$E(X) = \sum_{s \in \Omega} X(s) P(s).$$

This leads to an easier way of computing the expected value. To this end, we exploit the following important property of expectations.

LINEARITY OF EXPECTATION. Let $X, Y : \Omega \to \mathbb{R}$ be two random variables. Then

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

(ii) $E(cX) = cE(X)$, for every real number c.

The proof should be obvious. Is it? We use the property to recompute the expected number of successes in

a Bernoulli trial process. For i from 1 to n, let X_i be the expected number of successes in the i-th trial. Since there is only one i-th trial, this is the same as the probability of having a success, that is, $E(X_i) = p$. Furthermore, $X = X_1 + X_2 + \ldots + X_n$. Repeated application of property (i) of the Linearity of Expectation gives $E(X) = \sum_{i=1}^{n} E(X_i) = np$, same as before.

Indication. The Linearity of Expectation does not depend on the independence of the trials; it is also true if X and Y are dependent. We illustrate this property by going back to our hat checking experiment. First, we introduce a definition. Given an event, the corresponding *indicator random variable* is 1 if the event happens and 0 otherwise. Thus, $E(X) = P(X = 1)$.

In the hat checking experiment, we return n hats in a random order. Let X be the number of correctly returned hats. We proved that the probability of returning at least one hat correctly is $P(X \geq 1) = 1 - e^{-1} = 0.6\ldots$ To compute the expectation from the definition, we would have to determine the probability of returning exactly k hats corrects, for each $0 \leq k \leq n$. Alternatively, we can compute the expectation by decomposing the random variable, $X = X_1 + X_2 + \ldots + X_n$, where X_i is the expected value that the i-th hat is returned correctly. Now, X_i is an indicator variable with $E(X_i) = \frac{1}{n}$. Note that the X_i are not independent. For example, if the first $n-1$ hats are returned correctly then so is the n-th hat. In spite of the dependence, we have

$$E(X) = \sum_{i=1}^{n} E(X_i) = 1.$$

In words, the expected number of correctly returned hats is one.

Example: computing the minimum. Consider the following algorithm for computing the minimum among n items stored in a linear array.

```
min = A[1];
for i = 2 to n do
  if min > A[i] then min = A[i] endif
endif.
```

Suppose the items are distinct and the array stores them in a random sequence. By this we mean that each permutation of the n items is equally likely. Let X be the number of assignments to min. We have $X = X_1 + X_2 + \ldots + X_n$, where X_i is the expected number of assignments in the i-th step. We get $X_i = 1$ iff the i-th item, $A[i]$, is smaller than all preceding items. The chance for this to happen is one in i. Hence,

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} E(X_i) \\ &= \sum_{i=1}^{n} \frac{1}{i}. \end{aligned}$$

The result of this sum is referred to as the *n-th harmonic number*, $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$. We can use $\int_{x=1}^{n} = \ln n$ to show that the n-th harmonic number is approximately the natural logarithm of n. More precisely, $\ln(n+1) \leq H_n \leq 1 + \ln n$.

Waiting for success. Suppose we have again a Bernoulli trial process, but this time we end it the first time we hit a success. Defining X equal to the index of the first success, we are interested in the expected value, $E(X)$. We have $P(X = i) = (1-p)^{i-1}p$ for each i. As a sanity check, we make sure that the probabilities add up to one. Indeed,

$$\begin{aligned} \sum_{i=1}^{\infty} P(X = i) &= \sum_{i=1}^{\infty} (1-p)^{i-1} p \\ &= p \cdot \frac{1}{1-(1-p)}. \end{aligned}$$

Using the Linearity of Expectation, we get a similar sum for the expected number of trials. First, we note that $\sum_{j=0}^{\infty} jx^j = \frac{x}{(1-x)^2}$. There are many ways to derive this equation, for example, by index transformation. Hence,

$$\begin{aligned} E(X) &= \sum_{i=0}^{\infty} iP(X = i) \\ &= \frac{p}{1-p} \sum_{i=0}^{\infty} i(1-p)^i \\ &= \frac{p}{1-p} \cdot \frac{1-p}{(1-(1-p))^2}, \end{aligned}$$

which is equal to $\frac{1}{p}$.

Summary. Today, we have learned about random variable and their expected values. Very importantly, the expectation of a sum of random variables is equal to the sum of the expectations. We used this to analyze the Bernoulli trial process.

18 Probability in Hashing

A popular method for storing a collection of items to support fast look-up is hashing them into a table. Trouble starts when we attempt to store more than one item in the same slot. The efficiency of all hashing algorithms depends on how often this happens.

Birthday paradox. We begin with an instructive question about birthdays. Consider a group of n people. Each person claims one particular day of the year as her birthday. For simplicity, we assume that nobody claims February 29 and we talk about years consisting of $k = 365$ days only. Assume also that each day is equally likely for each person. In other words,

$$P(\text{person } i \text{ is born on day } j) \quad = \quad \frac{1}{k},$$

for all i and all j. Collecting the birthdays of the n people, we get a multiset of n days during the year. We are interested in the event, A, that at least two people have the same birthday. Its probability is one minus the probability that the n birthdays are distinct, that is,

$$\begin{aligned} P(A) &= 1 - P(\bar{A}) \\ &= 1 - \frac{k}{k} \cdot \frac{k-1}{k} \cdot \ldots \cdot \frac{k-n+1}{k} \\ &= 1 - \frac{k!}{(k-n)!k^n}. \end{aligned}$$

The probability of A surpasses one half when n exceeds 21, which is perhaps surprisingly early. See Figure 21 for a display how the probability grows with increasing n.

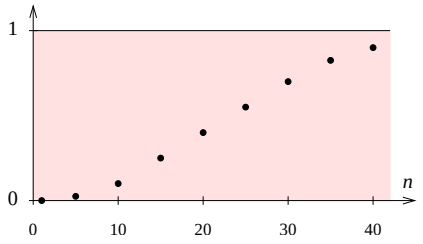

Figure 21: The probability that at least two people in a group of n share the same birthday.

Hashing. The basic mechanism in hashing is the same as in the assignment of birthdays. We have n items and map each to one of k slots. We assume the n choices of slots are independent. A *collision* is the event that an item is mapped to a slot that already stores an item. A possible resolution of a collision adds the item at the end of a linked list that belongs to the slot, but there are others. We are interested in the following quantities:

1. the expected number of items mapping to same slot;
2. the expected number of empty slots;
3. the expected number of collisions;
4. the expected number of items needed to fill all k slots.

Different hashing algorithms use different mechanisms for resolving collisions. The above quantities have a lot to say about the relative merits of these algorithms.

Items per slot. Since all slots are the same and none is more preferred than any other, we might as well determine the expected number of items that are mapped to slot 1. Consider the corresponding indicator random variable,

$$X_i \quad = \quad \begin{cases} 1 & \text{if item } i \text{ is mapped to slot 1;} \\ 0 & \text{otherwise.} \end{cases}$$

The number of items mapped to slot 1 is therefore $X = X_1 + X_2 + \ldots + X_n$. The expected value of X_i is $\frac{1}{k}$, for each i. Hence, the expected number of items mapped to slot 1 is

$$E(X) \quad = \quad \sum_{i=1}^{n} E(X_i) \quad = \quad \frac{n}{k}.$$

But this is obvious in any case. As mentioned earlier, the expected number of items is the same for every slot. Writing Y_j for the number of items mapped to slot j, we have $Y = \sum_{j=1}^{k} Y_j = n$. Similarly,

$$E(Y) \quad = \quad \sum_{j=1}^{k} E(Y_j) \quad = \quad n.$$

Since the expectations are the same for all slots, we therefore have $E(Y_j) = \frac{n}{k}$, for each j.

Empty slots. The probability that slot j remains empty after mapping all n items is $(1 - \frac{1}{k})^n$. Defining

$$X_j \quad = \quad \begin{cases} 1 & \text{if slot } j \text{ remains empty;} \\ 0 & \text{otherwise,} \end{cases}$$

we thus get $E(X_j) = (1 - \frac{1}{k})n$. The number of empty slots is $X = X_1 + X_2 + \ldots + X_k$. Hence, the expected

number of empty slots is

$$E(X) \;=\; \sum_{j=1}^{k} E(X_j) \;=\; k\left(1-\frac{1}{k}\right)^n.$$

For $k=n$, we have $\lim_{n\to\infty}(1-\frac{1}{n})^n = e^{-1} = 0.367\ldots$ In this case, we can expect about a third of the slots to remain empty.

Collisions. The number of collisions can be determined from the number of empty slots. Writing X for the number of empty slots, as before, we have $k-X$ items hashed without collision and therefore a total of $n-k+X$ collisions. Writing Z for the number of collisions, we thus get

$$\begin{aligned} E(Z) &= n-k+E(X) \\ &= n-k+k\left(1-\frac{1}{k}\right)^n. \end{aligned}$$

For $k=n$, we get $\lim_{n\to\infty} n(1-\frac{1}{n})^n = \frac{n}{e}$. In words, about a third of the items cause a collision.

Filling all slots. How many items do we need to map to the k slots until they store at least one item each? For obvious reasons, this question is sometimes referred to as the coupons collector problem. The crucial idea here is to define X_j equal to the number of items it takes to go from $j-1$ to j filled slots. Filling the j-th slot is an infinite Bernoulli process with success probability equal to $p = \frac{k-j+1}{k}$. Last lecture, we learned that the expected number of trials until the first success is $\frac{1}{p}$. Hence, $E(X_j) = \frac{k}{k-j+1}$. The number of items needed to fill all slots is $X = X_1 + X_2 + \ldots + X_k$. The expected number is therefore

$$\begin{aligned} E(X) &= \sum_{j=1}^{k} E(X_j) \\ &= \sum_{j=1}^{k} \frac{k}{k-j+1} \\ &= k\sum_{j=1}^{k} \frac{1}{j} \\ &= kH_k. \end{aligned}$$

As mentioned during last lecture, this is approximately k times the natural logarithm of k. More precisely, we have $k\ln(k+1) \le kH_k \le k(1+\ln k)$.

19 Probability Distributions

Although individual events based on probability are unpredictable, we can predict patterns when we repeat the experiment many times. Today, we will look at the pattern that emerges from independent random variables, such as flipping a coin.

Coin flipping. Suppose we have a fair coin, that is, the probability of getting head is precisely one half and the same is true for getting tail. Let X count the times we get head. If we flip the coin n times, the probability that we get k heads is

$$P(X = k) \quad = \quad \binom{n}{k} / 2^n.$$

Figure 22 visualizes this distribution in the form of a histogram for $n = 10$. Recall that the *distribution function* maps every possible outcome to its probability, $f(k) = P(X = k)$. This makes sense when we have a discrete domain. For a continuous domain, we consider the *cumulative distribution function* that gives the probability of the outcome to be within a particular range, that is, $\int_{x=a}^{b} f(x)\,\mathrm{d}x = P(a \leq X \leq b)$.

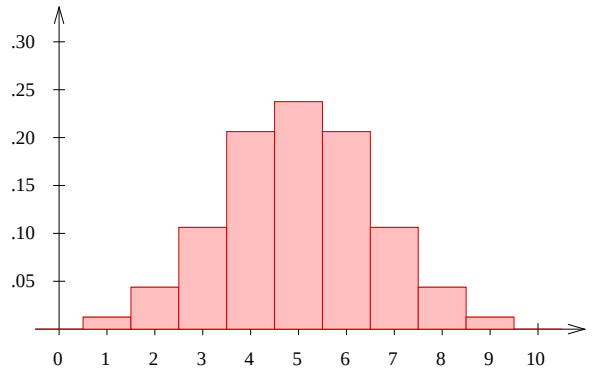

Figure 22: The histogram that the shows the probability of getting $0, 1, \ldots, 10$ heads when flipping a coin ten times.

Variance. Now that we have an idea of what a distribution function looks like, we wish to find succinct ways of describing it. First, we note that $\mu = E(X)$ is the expected value of our random variable. It is also referred to as the *mean* or the *average* of the distribution. In the example above, where X is the number of heads in ten coin flips, we have $\mu = 5$. However, we would not be surprised if we had four or six heads but we might be surprised if we had zero or ten heads when we flip a coin ten times. To express how surprised we should be we measure the spread of the distribution. Let us first determine how close we expect a random variable to be to its expectation, $E(X - E(X))$. By linearity of expectation, we have

$$E(X - \mu) \quad = \quad E(X) - E(\mu) \quad = \quad \mu - \mu \quad = \quad 0.$$

Hence, this measurement is not a good description of the distribution. Instead, we use the expectation of the square of the difference to the mean. Specifically, the *variance* of a random variable X, denoted as $V(X)$, is the expectation $E\left((X - \mu)^2\right)$. The *standard deviation* is the square root of the variance, that is, $\sigma(X) = V(X)^{1/2}$. If X_4 is the number of heads we see in four coin flips, then $\mu = 2$ and

$$V(X_4) \quad = \quad \frac{1}{16}\left[(-2)^2 + 4 \cdot (-1)^2 + 4 \cdot 1^2 + 2^2\right],$$

which is equal to 1. For comparison, let X_1 be the number of heads that we see in one coin flip. Then $\mu = \frac{1}{2}$ and

$$V(X_1) \quad = \quad \frac{1}{2}\left[(0 - \frac{1}{2})^2 + (1 - \frac{1}{2})^2\right],$$

which is equal to one quarter. Here, we notice that the variance of four flips is the sum of the variances for four individual flips. However, this property does not hold in general.

Variance for independent random variables. Let X and Y be independent random variables. Then, the property that we observed above is true.

ADDITIVITY OF VARIANCE. If X and Y are independent random variables then $V(X + Y) = V(X) + V(Y)$.

We first prove the following more technical result.

LEMMA. If X and Y are independent random variables then $E(XY) = E(X)E(Y)$.

PROOF. By definition of expectation, $E(X)E(Y)$ is the product of $\sum_i x_i P(X = x_i)$ and $\sum_j y_j P(Y = y_j)$. Pushing the summations to the right, we get

$$\begin{aligned} E(X)E(Y) \quad &= \quad \sum_i \sum_j x_i y_j P(X = x_i) P(Y = y_j) \\ &= \quad \sum_{i,j} z_{ij} P(X = x_i) P(Y = y_j), \end{aligned}$$

where $z_{ij} = x_i y_j$. Finally, we use the independence of the random variables X and Y to see that $P(X = x_i)P(Y = y_j) = P(XY = z_{ij})$. With this, we conclude that $E(X)E(Y) = E(XY)$. □

Now, we are ready to prove the Additivity of Variance, that is, $V(X+Y) = V(X) + V(Y)$ whenever X and Y are independent.

PROOF. By definition of variance, we have

$$V(X+Y) = E\left((X+Y-E(X+Y))^2\right).$$

The right hand side is the expectation of $(X-\mu_X)^2 + 2(X-\mu_X)(Y-\mu_Y) + (Y-\mu_Y)$, where μ_X and μ_Y are the expected values of the two random variables. With this, we get

$$\begin{aligned} V(X+Y) &= E\left((X-\mu_X)^2\right) + E\left((Y-\mu_Y)^2\right) \\ &= V(X) + V(Y), \end{aligned}$$

as claimed. □

Normal distribution. If we continue to increase the number of coin flips, then the distribution function approaches the *normal distribution*,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

This is the limit of the distribution as the number of coin flips approaches infinity. For a large number of trials, the

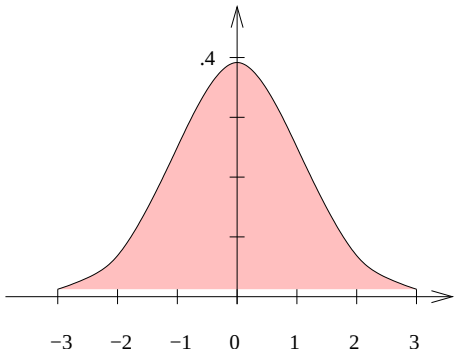

Figure 23: The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The probability that the random variable is between $-\sigma$ and σ is 0.68, between -2σ and 2σ is 0.955, and between -3σ and 3σ is 0.997.

normal distribution can be used to approximate the probability of the sum being between a and b standard deviations from the expected value.

STANDARD LIMIT THEOREM. The probability of the number of heads being between $a\sigma$ and $b\sigma$ from the mean goes to

$$\frac{1}{\sqrt{2\pi}} \int_{x=a}^{b} e^{-\frac{x^2}{2}} \, \mathrm{d}x$$

as the number of flips goes to infinity.

For example, if we have 100 coin flips, then $\mu = 50$, $V(X) = 25$, and $\sigma = 5$. It follows that the probability of having between 45 and 55 heads is about 0.68.

Summary. We used a histogram to visualize the probability that we will have k heads in n flips of a coin. We also used the mean, μ, the standard deviation, σ, and the variance, $V(X)$, to describe the distribution of outcomes. As n approaches infinity, we see that this distribution approaches the normal distribution.

Fifth Homework Assignment

Write the solution to each problem on a single page. The deadline for handing in solutions is 8 April 2009.

Question 1. (20 points). Use the Principle of Inclusion-Exclusion to count the surjective functions $f : M \to N$, where where both sets are finite with $m = |M|$ and $n = |N|$.

Question 2. ($20 = 6 + 7 + 7$ points). (Problems 5.3-1 to 3 in our textbook). Suppose you have a fair coin, one in which a flip gives head with probability one half and tail with probability one half. You do three flips with this coin.

(a) What is the probability that two flips in are row are heads, given that there is an even number of heads?

(b) Is the event that two flips in a row are heads independent of the event that there is an even number of heads?

(c) Is the event of getting at most one tail independent of the event that not all flips are identical?

Question 3. (20 points). (Problem 5.4-16 in our textbook). Suppose you have two nickels, two dimes, and two quarters in a bag. You draw three coins from the bag, without replacement. What is the expected amount of money you get?

Question 4. ($20 = 6 + 7 + 7$ points). (Problem 5.5-8 in our textbook). Suppose you hash n items into k locations.

(a) What is the probability that all n items has to different locations?

(b) What is the probability that the i-th item gives the first collision?

(c) What is the expected number of items you hash until the first collision?

Question 5. ($20 = 7 + 7 + 6$ points). In the programming language of your choice, write the following two functions:

(a) GETMEAN

(a) GETVARIANCE

These methods should take an array of values as input (the experimental results for each trial) and return a floating point number. Then, flip a coin 20 times (or simulate this on the computer) and use these methods to compute the mean and the variance of your trials. Are the results what you would have expected?

Question 6. ($20 = 10 + 10$ points). (Problems 5.7-8 and 14 in our textbook).

(a) Show that if X and Y are independent, and b and c are constant, then $X - b$ and $Y - c$ are independent.

(b) Given a random variable X, how does the variance of cX relate to that of X?