CSCE-658 Randomized Algorithms

Lecture #5, February 9, 2016

Lecturer: Professor Jianer Chen

5 How to count: sampling and allocation

In this section, we study some combinatorial techniques that help counting the number of possible cases in various circumstances. These techniques are useful when we compute probability and develop randomized algorithms.

Let us start with the *fundamental counting principle*: suppose that you have a sequence of k choices, and that there are m_1 ways for the first choice, m_2 ways for the second choice, ..., and m_k ways for the k-th choice. then the total number of different ways for your choices is $m_1 m_2 \cdots m_k$.

Counting for sampling

Sampling problems play an important role in the study of probability theory. The problems can be modeled by drawing balls from an urn, with various specified conditions. Assume that there are n distinguished balls, marked with 1, 2, ..., n, respectively, in the urn, from which k balls will be drawn. We want to count the number of all possible outcomes of the sampling. This number depends on how you draw the balls and how you define an outcome. We say that a sampling is *with replacement* if after drawing each ball, we put it back to the urn (but record the result of the drawing). We say that a sampling is *with ordering* if the result of the sampling is recorded as an *ordered* sequence of numbers in $\{1, 2, \ldots, n\}$ where the h-th number in the sequence records the result of the h-th drawing. In particular, two sequences that have the same numbers but different orders are treated as two different outcomes of the sample. On the other hand, in a sampling *without ordering*, two drawings are regarded to have the same outcome if they get the same set of balls, even in different orderings.

Model 1. Sampling with replacement and with ordering.

We have k drawings. Since the drawn balls are put back to the urn, each drawing can pick any of the n balls in the urn. By the fundamental counting principle, the number of different outcomes is

$$\underbrace{n \cdot n \cdot \cdots \cdot n}_{k} = n^k.$$

Model 2. Sampling without replacement but with ordering.

First note that under this condition, we must have $n \geq k$. The first drawing has n possible results. Since the drawn ball is not put back to the urn, the second drawing has only $n-1$ possible results. In general, the h-th drawing, with $h \leq k$, has only $n-h+1$ possible results. Therefore, the total number of different outcomes is

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

In particular, if $n = k$, then each outcome of the sampling gives us a *permutation* of the numbers 1, 2, ..., n. The total number of permutations of the n numbers is $n!$ (by definition, $0! = 1$).

This may be a proper place to have some discussions on the *factorial* function $n!$. In particular, we would like to have an estimation on the value $n!$. The value $n!$ is the product of the n numbers 1, 2, ..., n. If we replace each of these numbers by n, we certainly get a larger number. Thus,

$$n! \leq n^n. \tag{11}$$

On the other hand, if we ignore the first $n/2$ numbers 1, 2, ..., $n/2$, and replace each of the last $n/2$ numbers by a smaller number $n/2$, then we certainly get a smaller number. Thus,

$$n! \geq (n/2)^{n/2}. \tag{12}$$

From (11) and (12), we conclude that the value $n!$ is of the order $\Theta(n)^{\Theta(n)}$ (recall that $\Theta(n)$ means a function that is of the same order as that of n).

If a more accurate estimation is needed, we can use *Stirling's formula*, which says that $n!$ is approximately $\sqrt{2\pi n}(n/e)^n$, where $e = 2.718\cdots$ is the base of the natural logarithm. In fact, the following upper and lower bounds for $n!$ are known [17], which hold true for all positive integers n:

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/(12n+1)} < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/(12n)}.$$

Since $e^{1/(12n+1)} > e^0 = 1$ and $e^{1/(12n)} \le e^{1/12} < 1.09$ for all integers $n > 0$, we have

Theorem 5.1 (Stirling's formula) *For all integers $n > 0$.*

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! < 1.09\sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

Stirling's formula will be used repeatedly in our analysis.

Model 3. Sampling without replacement and without ordering.

This case can be derived from that of Model 2. Recall that in Model 2, when we draw k balls, each ordering of the drawings of these k balls counts as a different outcome. Fixed k balls, then these k balls have $k!$ different orderings, which constitute $k!$ outcomes in Model 2. If we do not care the drawing ordering, then for any k balls, the $k!$ different outcomes in Model 2 should be treated as a single outcome in Model 3. Since the total number of outcomes in Model 2 is $n!/(n-k)!$, we conclude that the total number of different outcomes under the condition of Model 3 is

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

Indeed, sampling under Model 3 actually asks the number of ways to pick k balls from the urn of n balls, which by definition is equal to the *binomial coefficient* $\binom{n}{k}$ (i.e., "n choose k").

Model 4. Sampling with replacement but without ordering.

This model is a bit tricky. Since the sampling is with replacement, each ball may be drawn multiple times. On the other hand, the sampling is without ordering. Thus, we are only interested in the number of times each ball is drawn. Therefore, each outcome of the sampling can be stated as "ball 1 was drawn k_1 times, ball 2 was drawn k_2 times, ..., and ball n was drawn k_n times," where the numbers k_1, k_2, ..., k_n must satisfy $k_1 + k_2 + \cdots + k_n = k$.

Fix an outcome of the sampling, which consists of a length-k sequence of the numbers from 1, 2, ..., n. Note that a number in $\{1, 2, \ldots, n\}$ may appear 0 time, one time, or more than one times in the sequence. Sort the sequence in non-decreasing order. Now insert $n-1$ vertical bars in the sorted sequence to break the sequence into n subsequences so that each subsequence consists of the same number. Therefore, all numbers in the subsequence before the first vertical bar are all 1, all numbers in the subsequence between the first and the second vertical bars are all 2, and so on. Note that some of these subsequences can be empty so the two bounding vertical bars are adjacent to each other. The left sequence given in (13) below is an illustration of this representation, which corresponds to the sampling that made $k = 7$ drawings (with replacement) in an urn of $n = 9$ balls, in which ball 1 was drawn 3 times, ball 5 was drawn twice, each of the balls 4 and 7 was drawn once, and the other 5 balls were not drawn.

$$1\,1\,1\,|\,|\,|\,4\,|\,5\,5\,|\,|\,7\,|\,| \qquad \longrightarrow \qquad \star\,\star\star\,|\,|\,|\,\star\,|\,\star\star\,|\,|\,\star\,|\,| \tag{13}$$

It is easy to see that the actual values of the numbers in the above representation are not necessary, which can be uniquely determined by the slot in which they are in: for all h, all the numbers between the $(h-1)$-th vertical bar and the h-th vertical bar are the number h. Therefore, we can simply replace all the number values in the sequence with an anonymous symbol "$\star$", as shown in the right sequence in (13), and the resulting sequence, which consists of k $\star$'s and $(n-1)$ |'s still uniquely represents the outcome of the sampling. It is not hard to observe and verify that there is a one-to-one correspondence

between the set of such sequences and the set of all outcomes of the sampling under this model. Since each such sequence is uniquely determined by the positions of the $n-1$ |'s in the sequence, we conclude that the total number of such sequences, i.e., the total number of outcomes of the sampling under this model, is equal to

$$\binom{k+n-1}{n-1} = \binom{k+n-1}{k},$$

where $\binom{k+n-1}{n-1}$ is the number of ways to pick $n-1$ positions from the $k+n-1$ positions in the sequence to place the symbol |, while $\binom{k+n-1}{k}$ is the number of ways to pick k positions in the sequence to place the symbol $\star$. By mathematics, we know that these two binomial coefficients are equal.

Model 5. Permuting n balls with k different colors.

Finally, we consider a special model of sampling. In this model, we have n balls, and each ball is colored with a color from $\{c_1, c_2, \ldots, c_k\}$, where balls of the same color are indistinguishable. To be more specific, suppose that there are n_1 balls colored with c_1, n_2 balls colored with c_2, ..., and n_k balls colored with c_k. The sampling draws n balls and is without replacement. On the other hand, the sampling *is* with ordering, though it cannot distinguish balls of the same color. How many different outcomes are there under this model? Note that this question is the same as asking the number of different sequences of length n that consists of the k symbols c_1, c_2, ..., c_k, with n_1 c_1's, n_2 c_2's, ..., and n_k c_k's, where $n_1 + n_2 + \cdots + n_k = n$.

This model can be studied based on Model 2. Suppose that we also assign each ball a distinct number from $\{1, 2, \ldots, n\}$. By the discussion on Model 2, there are $n!$ different permutations in terms of the n numbers we assigned to the balls. Now fix such a permutation S, and fix a color c_i. Since two balls with the same color c_i are not distinguishable in terms of their color, any permutation of the n_i balls of color c_i (without changing the positions that hold the balls of color c_i) in the sequence S will correspond to the same outcome under Model 5 where we only consider the colors of the balls. Therefore, because of color c_i, there are $n_i!$ permutations in terms of the n numbers we assigned that correspond to the same outcome under Model 5. This is true for all colors. Thus, we conclude that the total number of outcomes of the sampling under Model 5 is

$$\frac{n!}{n_1! n_2! \cdots n_k!},$$

where again we take the convention $0! = 1$ (thus there can be unused colors).

Counting for allocation

In the model of allocation, we have n boxes and k tokens, where the boxed are labeled with b_1, b_2, ..., b_n, and the tokens are labeled with t_1, t_2, ..., t_k. We consider the number of ways we can put the tokens in the boxes, where each outcome is a different distribution of the k tokens in the n boxes. Again, we need to specify various conditions for the allocation, which will significantly affect the number of possible outcomes. Some natural conditions include:

1. Do you allow each box to hold more than one token? and
2. Do you ignore the token labels in an allocation?

This model and the conditions can be directly translated into the model of sampling, as follows. Take the box b_i in the allocation model as the ball i in the sampling model, for all $i = 1, 2, \ldots, n$, and take the token t_j in the allocation model as the "j-th drawing" in the sampling model, for all $j = 1, 2, \ldots, k$. Thus, "putting the token t_j into the box b_i" in the allocation model corresponds to "the j-th drawing gets the ball i" in the sampling model. With this correspondence, in many cases, conclusions on the allocation model can be derived directly from the sampling model.

Model 1'. Each box can hold any number of tokens and the token labels are observed.

Putting the token t_j in the box b_i corresponds to the j-th drawing that gets the ball i in sampling. Since a box is allowed to hold more than one token, which corresponds to a ball that can be drawn multiple times in sampling, this is the model of sampling with replacement. Moreover, since the token

labels are observed, the ordering of the drawing in the sampling model matters. Thus, this corresponds to the sampling model with ordering. Therefore, Model 1' corresponds exactly to Model 1 in sampling, and we conclude that there are totally n^k different outcomes under Model 1'.

This can be verified purely based on the allocation model. An outcome of this model has each token in a particular box. If we interpret the token t_j in the box b_i as "assigning value i to token t_j," then each token can be assigned n different values. By the fundamental counting principle, the total number of ways to assign values to the k tokens, i.e., the total number of outcomes of Model 1', is equal to n^k.

Model 2'. Each box can hold at most one token and the token labels are observed.

Similar to the analysis given for Model 1', we derive that this model corresponds to Model 2 in sampling, which is without replacement but with ordering. Therefore, the total number of outcomes of Model 2' is $n!/(n-k)!$. In particular, in the case of $k = n$, this model asks the number of ways to distribute the n tokens into the n boxes, with each box holding exactly one token, which is just the number of permutations of n symbols, and is equal to $n!$.

Model 3'. Each box can hold at most one token and the token labels are ignored.

Again, this corresponds to Model 3 in sampling, thus has $\binom{n}{k}$ different outcomes. In fact, this corresponds to the task of selecting k boxes among the n boxes and putting the k (indistinguishable) tokens into these k selected boxes, one token per box.

Model 4'. Each box can hold any number of tokens and the token labels are ignored.

This corresponds to Model 4 in sampling, thus has $\binom{k+n-1}{n-1} = \binom{k+n-1}{k}$ different outcomes. The allocation model has a more intuitive interpretation based on the right sequence in (13): the $\star$'s between the $(h-1)$-th vertical bar and the h-th vertical bar correspond to the tokens in the box b_h for $h = 2, \ldots, n-1$, the $\star$'s before the first vertical bar correspond to the tokens in the box b_1, and the $\star$'s after the last vertical bar correspond to the tokens in the box b_n (recall that the tokens are indistinguishable). Thus, each such a sequence corresponds uniquely to an outcome of Model 4'.

Both the model of sampling and the model of allocation are useful when we count. Sometimes one model is probably more natural and intuitive than the other. The allocation model may be easier in some cases for imposing further conditions. For example, we may require that no box be left empty when $k \geq n$, or specify loads for some specific boxes. These conditions can also be translated into the sampling model but may probably look less natural.