Introduction to Probability

Nader H. Bshouty
Department of Computer Science
Technion 32000
Israel
e-mail: bshouty@cs.technion.ac.il

1 Combinatorics

1.1 Simple Rules In Combinatorics

The **rule of sum** says that the number of ways to choose an element from one of two disjoints sets is the sum of the cardinalities of the sets. That is, if A and B are two finite sets with no members in common, then $|A \cup B| = |A| + |B|$.

The **rule of product** says that the number of ways to choose an ordered pair is the number of ways to choose the first element from A and the second element from B. That is, if A and B are two finite sets, then $|A \times B| = |A| \cdot |B|$.

1.2 Simple Counting

Strings

A **string** over a set S is a sequence of elements of S. For example, there are 8 binary strings of length 3: 000,001,010,011,100,101,110,111. A k-string over a set S can be viewed as an element of the Cartesian product S^k of k-tuples; thus, there are $|S|^k$ strings of length k.

Permutations

A **permutation** of a finite set S is an ordered sequence of all the elements of S, with each element appearing exactly once. For example, if $S = \{a, b, c\}$, there are 6 permutations of S: abc,acb,bac,bca,cab,cba. There are $n!$ permutations of a set of n elements. A **k-permutation** of S is an ordered sequence of k of S, with no element appearing more than once in the sequence. The number of k-permutations of an n-set is:

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

Combinations

A k-**combination** of an n-set S is simply a k-subset of S. There are six 2-combinations of the 4-set $\{a, b, c, d\}$: ab, ac, ad, bc, bd, cd. The number of k-combinations of an n-set.

$$\frac{n!}{k!(n-k)!}$$

1.3 Equalities

Binomial coefficients

We use the notation $\binom{n}{k}$ (read "n choose k") to denote the number of k-combinations of an n-set. We have

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

This formula is symmetric in k and $n-k$:

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

These numbers are also known as **binomial coefficients**, due to their appearance in the **binomial expansion**:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

A special case:

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

We also have

$$\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}$$

$$\binom{n}{k} = \frac{n}{n-k}\binom{n-1}{k}$$

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

1.4 Inequalities

Binomial bounds

For $1 \leq k \leq n$, we have the lower bound:

$$\binom{n}{k} \geq \left(\frac{n}{k}\right)^k.$$

The upper bound:

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k.$$

For all $0 \leq k \leq n$ (see 2),

$$\binom{n}{k} \leq \frac{n^n}{k^k(n-k)^{n-k}},$$

where for convenience we assume that $0^0 = 1$. For $k = \lambda n$, where $0 \leq \lambda \leq 1$, this bound can be rewritten as (see R3):

$$\binom{n}{\lambda n} \leq 2^{nH(\lambda)},$$

where $H(\lambda) = -\lambda lg\lambda - (1-\lambda)lg(1-\lambda)$ is the **entropy function** and where, for convenience, we assume that $0 \lg 0 = 0$, so that $H(0) = H(1) = 0$. For $0 \leq \lambda \leq \frac{1}{2}$ we have (see R4)

$$\sum_0^{\lambda n} \binom{n}{i} \leq 2^{nH(\lambda)}.$$

For any $n \geq 0, j \geq 0, k \geq 0$, and $j + k \leq n$,

$$\binom{n}{j+k} \leq \binom{n}{j}\binom{n-j}{k}.$$

The Stirling's approximation is

$$\log n! = \left(n + \frac{1}{2}\right)\log n - n - \frac{1}{2}\log(2\pi) + o(1),$$

$$\binom{2n}{n} = \frac{2^{2n}}{\sqrt{\pi n}}(1 + O(1/n)).$$

2 Probability

2.1 Introduction

We define probability in terms of a **sample space** S, which is a set whose elements are called **elementary events**. For flipping two distinguishable coins, we can view the sample space $S = HH, HT, TH, TT$. An **event** is a subset of the sample space S. The event S is called the **certain event**, and the event $\emptyset$ is called the **null event**. We say that two events A and B are **mutually** exclusive if $A \cap B = \emptyset$.

Axioms of probability

A **probability distribution** $\Pr\{\}$ on a sample space S is a mapping from events of S to real numbers such that the following probability axioms are satisfied:

1. $\Pr[A] \geq 0$ for any event A.
2. $\Pr[S] = 1$.
3. $\Pr[A \cup B] = \Pr[A] + \Pr[B]$ for any two mutually exclusive events A and B.

We call $\Pr[A]$ the **probability** of the event A.

Results:

$$\Pr[\emptyset] = 0,$$

if $A \subseteq B$, then $\Pr[A] \leq \Pr[B]$. For $\overline{A}$ (the **complement** of A),

$$\Pr[\overline{A}] = 1 - \Pr[A].$$

For any two events A and B,

$$\begin{aligned} \Pr[A \cup B] &= \Pr[A] + \Pr[B] - \Pr[A \cap B] \\ &\leq \Pr[A] + \Pr[B] \end{aligned}$$

Discrete probability distributions

A probability distribution is **discrete** if it is defined over a finite or or countably infinite sample space.Then for any event A,

$$\Pr[A] = \sum_{s \in A} \Pr[s].$$

If S is finite and every elementary event $s \in S$ has probability $\Pr[s] = 1/|S|$ then we have the **uniform probability distribution** on S. A **fair coin** is one which the probability of obtaining a head is the same as the probability of obtaining a tail, that is, 1/2.

2.2 Conditional probability and Independence

Conditional probability formalizes the notion of having prior partial knowledge of the outcome of an experiment. The conditional probability of an event A given that another event B occurs is defined to be:

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

Whenever $\Pr[B] \neq 0$, (we read "$\Pr[A|B]$" as "the probability of A given B.") Two events are **independent** if, $\Pr[A \cap B] = \Pr[A] \Pr[B]$, which is equivalent, if $\Pr[B] \neq 0$, to the condition $\Pr[A|B] = \Pr[A]$. A collection $A_1, A_2, ..., A_n$ of events is said to be **pairwise independent** if $\Pr[A_i \cap A_j] = \Pr[A_i] \Pr[A_j]$. For all $1 \leq i < j \leq n$. We say that they are (**mutually**) **independent** if every k-subset $A_{i_1}, A_{i_2}, ..., A_{i_k}$ of collection, where $2 \leq k \leq n$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$, satisfies:

$$\Pr[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}] = \Pr[A_{i_1}] \Pr[A_{i_2}] \cdots \Pr[A_{i_k}]$$

Bayes's Theorem

$$\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[B]}.$$

From Bayes's theorem:

$$\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[A]\Pr[B|A] + \Pr[\overline{A}]\Pr[B|\overline{A}]}$$

Boole's inequality: For any finite or countably infinite sequence of events $A_1, A_2, ...,$

$$\Pr[A_1 \cup A_2 \cup \cdots] \leq \Pr[A_1] + \Pr[A_2] + \cdots.$$

For collection of events $A_1, A_2, ..., A_n$,

$$\begin{aligned}\Pr[A_1 \cap A_2 \cap \cdots \cap A_n] &= \Pr[A_1] \cdot \Pr[A_2|A_1] \cdot \Pr[A_3|A_1 \cap A_2] \cdots \\ &\quad \Pr[A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}]\end{aligned}$$

2.3 Discrete Random Variables and Expectation

A (**discrete**) **random variable** X is a function from a finite or countably infinite sample space S to the real numbers. For random variable X and a real number x, we define the event $X = x$ to be $[s \in S : X(s) = x]$; thus, $\Pr[X = x] = \sum_{[s \in S: X(s)=x]} \Pr[s]$. The function: $f(x) = \Pr[X = x]$, is the **probability density function** of the random variable X.

It is common for several random variables to be defined on the same sample space . If X and Y are random variables, the function: $f(x, y) = \Pr[X = x \text{ and } Y = y]$ Is the **joint probability density function** of X and Y. For a fixed value y, $\Pr[Y = y] = \sum_x \Pr[X = x$ and $Y = y]$. We define two random variables X and Y to be **independent** if for all x and y, the events $X = x$ and $Y = y$ are independent or, equivalently, if for all x and y, we have $\Pr[X = x \text{ and } Y = y] = \Pr[X = x]\Pr[Y = y]$.

Expected value of a random variable

The **expected value** (or, synonymously, **expectation** or **mean**) of a discrete random variable X is

$$E[X] = \sum_x x \Pr[X = x].$$

Sometimes the expectation of X is denoted by μ_x or , when the random is apparent from context, simply by μ. The expectations of the sum of two random variables is the sum of their expectations, that is,

$$E[X + Y] = E[X] + [Y].$$

If X is any random variable, any function $g(x)$ defines a new random variable $g(X)$. If the expectation of $g(X)$ is defined, then

$$E[g(X)] = \sum_x g(x) \Pr[X = x].$$

Letting $g(x) = ax$, we have for any constant a,

$$E[aX] = aE[X].$$

$$E[aX + Y] = aE[X] + E[Y].$$

When two random variables X and Y are independent and each has a defined expectation, $E[XY] = E[X]E[Y]$. In general, when n random variable $X_1, X_2, ..., X_n$ are mutually independent, $E[X_1 X_2 \cdots X_n] = E[X_1]E[X_2] \cdots E[X_n]$. When a random variable X takes on values from the natural numbers $N = \{0, 1, 2, ...\}$, there is a nice formula for its expectation:

$$E[X] = \sum_{i=1}^{\infty} \Pr[X \geq i].$$

Variance and standard deviation

The **variance** of a random variable X with mean $E[X]$ is:

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[X^2] - E^2[X] \end{aligned}$$

We have

$$E[X^2] = Var[X] + E^2[X]$$

$$Var[aX] = a^2 Var[X]$$

When X and Y are independent random variables, $Var[X+Y] = Var[X]+Var[Y]$. In general, if n random variables $X_1, X_2, ..., X_n$ are pairwise independent, then

$$Var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} Var[X_i].$$

The **standard deviation** of a random variable X is the positive square root of the variance of X.

Markov's inequality: (see Rule 1)

For a nonnegative random variable X

$$\Pr[X \geq t] \leq \frac{E[X]}{t}$$

for all $t > 0$ and

$$\Pr[X \geq \lambda E[X]] \leq \frac{1}{\lambda}.$$

See more results in [G] Appendix A.

Chebychev's inequality:

$$\Pr[|X - E[X]| \geq k] \leq \frac{Var[X]}{k^2}.$$

Bienayme-Chebyschev Let $X_1, \ldots, X_m$ be pairwise independent random variables such that $\mathbf{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\Pr\left[\left|\frac{\sum_{i=1}^{m} X_i}{m} - \mu\right| \geq \lambda\right] \leq \frac{\sigma^2}{\lambda^2 m}.$$

See the proof in [G] Appendix A.

2.4 Geometric and Binomial distributions

A coin flip is an instance of a **Bernoulli trail**, which is defined as an experiment with only two possible outcomes: **success**, which occurs with probability p, and **failure**, which occurs with probability $q = 1 - p$. The trails are mutually independent ok. each has the same probability p for success.

The geometric distribution

Suppose we have a sequence of Bernoulli trails, each with a probability p of success and a probability $q = 1 - p$ of failure. Let the random variable X be the number of trails needed to obtain a success.

$$\Pr[X = k] = q^{k-1}p$$

A probability distribution satisfying this is said to be a **geometric distribution**.

$$E[X] = 1/p$$

$$Var[X] = q/p^2$$

The binomial distribution

Define the random variable X to be the number of success in n trails.

$$\Pr[X = x] = \binom{n}{k} p^k q^{n-k}$$

A probability distribution satisfying this is said to be a **binomial distribution** . Define:

$$b(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

We have: $E[X] = np$ and $Var[X] = npq$. The binomial distribution $b(k; n, p)$ increases as k runs from 0 to n until it reaches the mean np, and then it decreases.

Let $n \geq 0$, let $0 < p < 1$, let $q = 1 - p$, and let $0 \leq k \leq n$. Then,

$$b(k; n, p) \leq \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}.$$

For $0 \leq k \leq n$,

$$b(k; n, 1/2) \leq 2^{nH(k/n)-n}.$$

The tails of the binomial distribution

Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in \{0, 1\}$ and $E[X_i] = p$. Then this is equivalent to a sequence of n Bernoulli trails , where success occurs with probability p. Let X be the random variable denoting the total number of successes. Then for $0 \leq k \leq n$, the probability of at least k success is:

$$\Pr\left[\sum_{i=1}^{n} X_i \geq k\right] = \sum_{i=k}^{n} b(i; n, p) \leq \binom{n}{k} p^k.$$

For $np < k < n$, the probability of more than k successes is:

$$\Pr\left[\sum_{i=1}^{n} X_i > k\right] = \sum_{i=k+1}^{n} b(i; n, p) < \frac{(n-k)p}{k - np} b(k; n, p),$$

and

$$\Pr\left[\frac{\sum_{i=1}^{n} X_i}{n} - p \geq \lambda\right] \leq \left(\frac{pe}{\lambda}\right)^{\lambda n}.$$

$$\Pr\left[\sum_{i=1}^{n} X_i \leq k\right] = \sum_{i=1}^{k} b(i; n, p) \leq \binom{n}{k} (1-p)^{n-k}.$$

For $0 < k < np$, the probability of fewer than k success is:

$$\Pr\left[\sum_{i=1}^{n} X_i < k\right] = \sum_{i=0}^{k-1} b(i; n, p) < \frac{kq}{np - k} b(k; n, p).$$

Chernoff Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in \{0, 1\}$ and $\mathbf{E}[X_i] = p$ Then

Multiplicative Form

$$\Pr\left[\frac{\sum_{i=1}^{n} X_i}{n} - p \geq \lambda p\right] \leq e^{-\lambda^2 np/3}.$$

$$\Pr\left[p - \frac{\sum_{i=1}^{n} X_i}{n} \geq \lambda p\right] \leq e^{-\lambda^2 np/2}.$$

Additive Form

$$\Pr\left[\frac{\sum_{i=1}^{n} X_i}{n} - p \geq \lambda\right] \leq e^{-2\lambda^2 n}.$$

$$\Pr\left[p - \frac{\sum_{i=1}^{n} X_i}{n} \geq \lambda\right] \leq e^{-2\lambda^2 n}.$$

and

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} X_i}{n} - p\right| \geq \lambda\right] \leq 2e^{-2\lambda^2 n}.$$

See another bound in [G] Appendix A.

Hoeffding Let $X_1, \ldots, X_m$ be independent random variables such that $X_i \in [a, b]$ and $\mathbf{E}[X_i] = p$. Then

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} X_i}{n} - p\right| \geq \lambda\right] \leq 2e^{-2\lambda^2 n/(b-a)^2}.$$

Other Results. Consider independent random variables $X_1, \ldots, X_n$ (a sequence of n Bernoulli trails) where $E[X_i] = p_i$ (where in the ith trail, for $i = 1, 2, ..., n$, success occurs with probability p_i and failure occurs with probability $q_i = 1 - p_i$). Let $\mu = E\left[\sum_{i=1}^{n} X_i\right]$. Then for $r > \mu$,

$$\Pr\left[\sum_{i=1}^{n} X_i - \mu \geq r\right] \leq \left(\frac{\mu e}{r}\right)^r$$

$$\Pr[X - \mu \geq r] \leq e^{-r^2/2n}$$

3 Examples

Example 1 Random Halving*: A game where you randomly uniformly choose an integer* $1 \leq i_1 \leq n$ *and then* $1 \leq i_2 \leq i_1$ *until you get* 1*. Show that the expected number of steps is* $1 + \sum_{i=1}^{n-1} 1/i$*.*

Solution: Let X_n be the expected number of steps then $X_1 = 1$ and

$$E[X_n] = 1 + E_{1 \leq i \leq n}[E[X_i]].$$

Now

$$\frac{n+1}{n} E[X_{n+1}] - E[X_n] = \frac{1}{n} + \frac{1}{n} E[X_{n+1}]$$

and therefore $E[X_{n+1}] = E[X_n] + (1/n)$.

Acknowledgement: To Vivian Bshouty for helping me with this survey.

References

[AS] N. Alon and J. H. Spencer. The Probabilistic Method.

[CLR] T. Cormen, C. E. Leiserson and R. L. Rivest. Introduction to Algorithms.

[G] O. Goldreich. Modern Cryptography, Probabilistic Proofs and Pseudo-randomness.

[MR] R. Motwani and P. Raghavan. Randomized Algorithms.

4 Proofs

Rule 1 Markov's inequality:

$$\Pr[X \geq t] \leq \frac{E[X]}{t}$$

for all $t > 0$. **Proof.** *We have*

$$E[X] = \sum_x \Pr[X = x] \cdot x \geq \sum_{x \geq t} \Pr[X = x] x \geq \Pr[X \geq t] t. \square$$

Rule 2 *We have*

$$\binom{n}{k} \leq \frac{n^n}{k^k (n-k)^{n-k}}.$$

Proof. We have

$$n^n = (k + (n-k))^n \geq \binom{n}{k} k^k (n-k)^{n-k}. \square$$

Rule 3 *For* $0 \leq \lambda \leq 1$ *we have*

$$\binom{n}{\lambda n} \leq 2^{nH(\lambda)},$$

where $H(\lambda) = -\lambda lg\lambda - (1-\lambda) lg(1-\lambda)$.

Proof. Use Rule 2.$\square$

Rule 4 *For* $0 \leq \lambda \leq \frac{1}{2}$ *we have*

$$\sum_0^{\lambda n} \binom{n}{i} \leq 2^{nH(\lambda)}.$$

Proof. We have

$$\begin{aligned} 1 &= (\lambda + (1-\lambda))^n \geq \sum_{i=1}^{\lambda n} \binom{n}{i} \lambda^i (1-\lambda)^{n-i} = \sum_{i=1}^{\lambda n} \binom{n}{i} (1-\lambda)^n \left(\frac{\lambda}{1-\lambda}\right)^{\lambda i} \\ &\leq \sum_{i=1}^{\lambda n} \binom{n}{i} (1-\lambda)^n \left(\frac{\lambda}{1-\lambda}\right)^{\lambda n} = 2^{-nH(\lambda)} \sum_0^{\lambda n} \binom{n}{i}. \square \end{aligned}$$