5 Introduction to Probability Theory

In this section we review the basic concepts in the area of probability theory.

5.1 Set Theory

Let M be an arbitrary set. Given any subset $A \subseteq M$, $|A|$ denotes its *cardinality*, that is, the number of elements in A, and $\bar{A}$ denotes its *complement* $M \setminus A$. 2^M is called the *power set* of M and consists of all subsets $A \subseteq M$. We summarize a few standard sets.

- $\emptyset$ denotes the empty set.
- $\mathbb{N}$ means the set of integers $\{1, 2, 3, 4, \ldots\}$, and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.
- For any $c \in \mathbb{N}$, $[c]$ represents the set $\{1, \ldots, c\}$.

Given two sets A and B,

- $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$ is called the *intersection* of A and B,
- $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$ is called the *union* of A and B, and
- $A \setminus B = \{x \mid x \in A \text{ and } x \notin B\}$ is called the *difference* of A and B.

Two sets A and B are called *disjoint* if $A \cap B = \emptyset$.

A very important principle in set theory is the *inclusion-exclusion principle*: Consider any collection of n sets $A_1, \ldots, A_n$. Then,

$$\left| \bigcup_{i=1}^{n} A_i \right| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < i_2 < \ldots < i_k} \left| \bigcap_{j=1}^{k} A_{i_j} \right| .$$

5.2 Combinatorics

We start with some basic definitions:

- $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-1) \cdot n$ is called "n factorial". As an example, there are $n!$ ways of arranging n different numbers in a sequence. In order to prove this, let $P(n)$ denote the number of ways of arranging (or *permuting*) n different numbers. Then it is easy to see that

$$\begin{aligned} P(1) &= 1 \quad \text{and} \\ P(n) &= n \cdot P(n-1) \quad \text{for all } n > 1 \end{aligned}$$

  This allows to prove via induction that $P(n) = n!$ and therefore our statement above is indeed correct.

- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is called "n choose k" or, in general, "binomial coefficient". As an example, there are $\binom{n}{k}$ different ways of picking a set of k numbers out of a set of n different numbers.

Many combinatorial problems can be viewed as drawing balls. Consider the problem of determining the number of ways of drawing k balls one after the other out of n balls, numbered from 1 to n. The most important cases are summarized in Figure 1.

To give an example that the expressions are true: for the special case in which we draw n balls out of n balls in an ordered way without replacement, we get the number of all possible ways of permuting n numbers, which as we know is $n!$.

	ordered	unordered
with replacement	n^k	$\binom{n+k-1}{k}$
without replacement	$n^{(k)} = \frac{n!}{(n-k)!}$	$\binom{n}{k}$

Figure 1: Number of outcomes of drawing k out of n balls

5.3 Basic concepts in probability theory

Next we introduce some basic concepts in probability theory.

Let us consider a random experiment of which all possible results are included in a non-empty set Ω, usually called the *sample space*. An element $\omega \in \Omega$ is called a *sample point* or *outcome* of the experiment. An *event* of a random experiment is specified as a subset of Ω. Event A is called *true* if an outcome $\omega \in \Omega$ has been chosen with $\omega \in A$. Otherwise A is called *false*. A system $\mathcal{A}$ of events (or, in general, subsets of Ω) is called an *algebra* if

- $\Omega \in \mathcal{A}$,
- if $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$ and $A \cap B \in \mathcal{A}$, and
- if $A \in \mathcal{A}$, then $\bar{A} \in \mathcal{A}$.

Given an algebra $\mathcal{A}$, a function $\mu : \mathcal{A} \to \mathbb{R}_+$ is called a *measure* on $\mathcal{A}$ if for every pair of disjoint sets $A, B \in \mathcal{A}$ we have

$$\mu(A \cup B) = \mu(A) + \mu(B) \ .$$

This definition clearly implies that $\mu(\emptyset) = 0$ and that for any set of pairwise disjoint events $A_1, \ldots, A_k \in \mathcal{A}$ we have

$$\mu\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{n} \mu(A_i) \ .$$

Furthermore, it implies that for any pair of sets $A, B \in \mathcal{A}$ we have

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

We say that a function $p : \mathcal{A} \to [0,1]$ is a *probability measure* if

- p is a measure on $\mathcal{A}$ and
- $p(\Omega) = 1$.

Given a probability measure p, the *probability* of an event A to be true is defined as

$$\Pr[A] = p(A) \ .$$

We say that a triple $(\Omega, \mathcal{A}, p)$ is a *probability space* if $\mathcal{A}$ is an algebra over Ω and p is a probability measure on $\mathcal{A}$.

5.4 Events

Starting from a given collection of sets that represent events, we can form new events by means of statements containing the logical connectives "or," "and," and "not," which correspond in the language of set theory to the operations "union," "intersection," and "complement."

If A and B are events, their *union*, denoted by $A \cup B$, is the event consisting of all outcomes realizing either A or B. The *intersection* of A and B, denoted by $A \cap B$, consists of all outcomes realizing both A and B. The *difference* of B and A, denoted by $B \setminus A$, consists of all outcomes that belong to B but not to A. If A is a subset of Ω, its *complement*, denoted by $\bar{A}$, is the set of outcomes in Ω that do not belong to A. That is, $\bar{A} = \Omega \setminus A$.

Two events A and B are called *disjoint* if $A \cap B$ is empty. In probability theory, $\emptyset$ is called the *impossible* event. The set Ω is naturally called the *certain* event.

If Ω is a countable sample space (i.e., its elements can be arranged in a sequence so that the rth element is identifiable for any $r \in \mathbb{N}$), we define the *size* of an event A, denoted by $|A|$, to be the number of outcomes it contains.

5.5 The Inclusion-Exclusion Principle

Let $A_1, \ldots, A_n$ be any collection of events. The *inclusion-exclusion principle* stated earlier implies that

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < i_2 < \ldots < i_k} \Pr\left[\bigcap_{j=1}^{k} A_{i_j}\right] . \tag{1}$$

For the special case of $n = 2$ we obtain

$$\Pr[A_1 \cup A_2] = \Pr[A_1] + \Pr[A_2] - \Pr[A_1 \cap A_2] .$$

In cases where it is too difficult to evaluate (1) exactly, *Bonferroni's inequalities* may be used to find suitable approximations:

- For every odd m,
$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{k=1}^{m} (-1)^{k+1} \sum_{i_1 < i_2 < \ldots < i_k} \Pr\left[\bigcap_{j=1}^{k} A_{i_j}\right] .$$
- For every even m,
$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \geq \sum_{k=1}^{m} (-1)^{k+1} \sum_{i_1 < i_2 < \ldots < i_k} \Pr\left[\bigcap_{j=1}^{k} A_{i_j}\right] .$$

Special cases of these inequalities are *Boole's inequalities*:

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \Pr[A_i]$$

and

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \geq \sum_{i=1}^{n} \Pr[A_i] - \sum_{1 \leq i < j \leq n} \Pr[A_i \cap A_j] .$$

We will give two examples to illustrate the use of these inequalities.

Example 1: Given a set system $M = \{S_1, \ldots, S_m\}$ over some set of elements V, consider the problem of coloring each element in V with one out of two colors in such a way that no set S_i is *monochromatic*, i.e. contains only nodes of a single color. We would like to identify a class of sets S for which this is always possible. Such a class is given in the following claim.

Claim 5.1 *For every set system M of size m in which every set $S \in M$ is of size at least $\log m + 2$, there is a 2-coloring of the elements such that no set in M is monochromatic.*

Proof. Consider the random experiment of choosing for each element independently and uniformly at random one of the two possible colors. In this case, the probability that a set of size k is monochromatic is equal to $2 \cdot 2^{-k} = 2^{-k+1}$. For every $i \in \{1, \ldots, m\}$, let A_i be the event that set S_i is monochromatic. Then, by the inclusion-exclusion principle,

$$\Pr[A_1 \cup \ldots \cup A_m] \le \sum_{i=1}^{m} \Pr[A_i] \le \sum_{i=1}^{m} 2^{-(\log m + 1)} = \frac{1}{2} .$$

Hence,

$$\Pr[\bar{A}_1 \cap \ldots \cap \bar{A}_n] = 1 - \Pr[A_1 \cup \ldots \cup A_n] \ge \frac{1}{2} ,$$

and therefore there must exist a 2-coloring such that no set is monochromatic. □

Example 2: Consider the situation that we have n balls and n bins, and each ball is placed in a bin chosen independently and uniformly at random.

Claim 5.2 *The probability that bin 1 has at least one ball is at least $1/2$.*

Proof. For any $i \in [n]$, let A_i be the event that ball i is placed in bin 1. Then, by Boole's inequality,

$$\begin{aligned}
\Pr[\text{bin 1 has at least one ball}] &= \Pr\left[\bigcup_{i \in [n]} A_i\right] \\
&\ge \sum_{1 \le i \le n} \Pr[A_i] - \sum_{1 \le i < j \le n} \Pr[A_i \cap A_j] \\
&= \sum_{1 \le i \le n} \frac{1}{n} - \sum_{1 \le i < j \le n} \frac{1}{n^2} \\
&= 1 - \binom{n}{2} \frac{1}{n^2} \ge 1 - \frac{1}{2} = \frac{1}{2} .
\end{aligned}$$

□

Observe that the probability bound in Claim 5.2 is not far away from the exact bound:

$$\Pr[\text{bin 1 has at least one ball}] = 1 - \left(1 - \frac{1}{n}\right)^n \stackrel{n \to \infty}{=} 1 - \frac{1}{\mathrm{e}} .$$

5.6 Conditional probability

The *conditional probability* of event B assuming an event A with $\Pr[A] > 0$ is denoted by $\Pr[B \mid A]$. It satisfies

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

or equivalently,

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B \mid A] \ . \tag{2}$$

Expression 5.6 holds, since the space of outcomes that is left for $\Pr[B \mid A]$ is the set of all outcomes in A, and therefore we have to normalize the probabilities of the outcomes in A in a way that they sum up to 1. This is achieved by dividing $\Pr[A \cap B]$ by $\Pr[A]$.

We can generalize Expression 2 as follows: if $A_1, \ldots, A_n$ are events with $\Pr[A_1 \cap \ldots \cap A_{n-1}] > 0$, then

$$\Pr[A_1 \cap \ldots \cap A_n] = \prod_{i=1}^{n} \Pr[A_i \mid A_1 \cap \ldots \cap A_{i-1}] \ .$$

Suppose that A and B are events with $\Pr[A] > 0$ and $\Pr[B] > 0$. Then, in addition to the equality (2), we have

$$\Pr[A \cap B] = \Pr[B] \cdot \Pr[A \mid B] \ . \tag{3}$$

From (2) and (3) we obtain *Bayes's formula*

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

Two events A and B are called *independent* if and only if

$$\Pr[B \mid A] = \Pr[B] \ .$$

Note that, due to Bayes's formula, in this case also $\Pr[A \mid B] = \Pr[A]$, that is, the independence property is *symmetric*. Furthermore, it holds that

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B] \ .$$

If $\Pr[B \mid A] \neq \Pr[B]$, then A and B are said to be *correlated*. A and B are called

- *negatively correlated* if $\Pr[B \mid A] < \Pr[B]$ and
- *positively correlated* if $\Pr[B \mid A] > \Pr[B]$.

By Bayes's formula, all of these correlation properties are also symmetric.

As an example, any two disjoint events A and B with positive probabilities cannot be independent, since $\Pr[B \mid A] = 0$. However, they are always negatively correlated. Furthermore, they have the property that

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] \ .$$

5.7 Random Variables

Any numerical function $X = X(\omega)$ defined on a sample space Ω may be called a *random variable.* In this lecture we will only consider integer-valued random variables, i.e., functions of the form $X : \Omega \to \mathbb{Z}$. A random variable X is called *non-negative* if $X(\omega) \geq 0$ for all $\omega \in \Omega$. For the special case that X maps elements in Ω to $\{0, 1\}$, X is called a *binary* or *Bernoulli* random variable. A binary random variable X is called an *indicator* of event A (denoted by I_A) if $X(\omega) = 1$ if and only if $\omega \in A$ for all $\omega \in \Omega$.

For any random variable X and any number $x \in \mathbb{Z}$, we define $[X = x] = \{\omega \in \Omega : \; X(\omega) = x\}$. Instead of using set operations to express combinations of events associated with random variables, we will use logical expressions in the following, that is,

- instead of $\Pr[[X = x] \cap [Y = y]]$ we write $\Pr[X = x \wedge Y = y]$, and
- instead of $\Pr[[X = x] \cup [Y = y]]$ we write $\Pr[X = x \vee Y = y]$.

Furthermore, we define

$$\Pr[X \leq k] = \sum_{\ell \leq k} \Pr[X = \ell] \quad \text{and} \quad \Pr[X \geq k] = \sum_{\ell \geq k} \Pr[X = \ell] \; .$$

The function $p_X(k) = \Pr[X = k]$ is called the *probability distribution* of X, and the function $F_X(k) = \Pr[X \leq k]$ is called the *(cumulative) distribution function* of X.

Two random variables X and Y are called *independent* if, for all $x, y \in \mathbb{Z}$,

$$\Pr[X = x \mid Y = y] = \Pr[X = x] \; .$$

5.8 Expectation

The most important measure used in combination with random variables is the expectation.

Definition 5.3 *Let $(\Omega, \mathcal{A}, p)$ denote an arbitrary probability space and $X : \Omega \to \mathbb{Z}$ be an arbitrary function with integer values. Then the* expectation *of X is defined as*

$$\mathrm{E}[X] = \sum_{x \in \mathbb{Z}} x \cdot \Pr[X = x] \; . \tag{4}$$

The following fact lists some basic properties of the expectation.

Fact 5.4 *Let X and Y be arbitrary random variables and c be an arbitrary constant.*

- $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.
- $\mathrm{E}[c \cdot X] = c \cdot \mathrm{E}[X]$.
- *If X is a binary random variable, then* $\mathrm{E}[X] = \Pr[X = 1]$.
- *If X and Y are independent, then* $\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$.

The expectation enables us to prove some simple tail estimates.

Fact 5.5 *Let X be an arbitrary random variable. Then*

$$\Pr[X < \mathrm{E}[X]] < 1 \quad \textit{and} \quad \Pr[X > \mathrm{E}[X]] < 1 \; .$$

The next result provides a first, simple probability bound that depends on the deviation from the expected value. It has apparently first been used by Chebychev, but it is commonly called *Markov inequality.*

Theorem 5.6 (Markov Inequality) *Let X be an arbitrary non-negative random variable. Then, for any $k > 0$,*

$$\Pr[X \geq k] \leq \frac{\mathrm{E}[X]}{k} \ .$$

Proof. Obviously,

$$\mathrm{E}[X] = \sum_{x \geq 0} x \cdot \Pr[X = x] \geq k \cdot \Pr[X \geq k] \ .$$

□

Next we present a simple example to illustrate the use of these inequalities.

Example: In the first theorem we show that every graph has a bipartite subgraph with at least half of its nodes.

Theorem 5.7 *Let $G = (V, E)$ be a graph with n vertices and m edges. Then G contains a bipartite subgraph with at least $m/2$ edges.*

Proof. Let $T \subseteq V$ be a random subset given by a random experiment with independent probabilities $\Pr[v \in T] = 1/2$ for every $v \in V$. We call an edge $\{v, w\}$ *crossing* if exactly one of v, w are in T. Let X be the number of crossing edges. We define

$$X = \sum_{\{v,w\} \in E} X_{v,w}$$

where $X_{v,w}$ is the indicator random variable for $\{v, w\}$ being crossing. Then

$$\mathrm{E}[X_{v,w}] = 1/2$$

as two fair coin flips have probability $1/2$ of being different. Then

$$\mathrm{E}[X] = \sum_{\{v,w\} \in E} \mathrm{E}[X_{v,w}] = \frac{m}{2} \ .$$

Thus, according to Fact 5.5, there must be some choice of T with $X \geq m/2$ and the set of those crossing edges forms a bipartite graph. □

The theorem goes not automatically provide an efficient algorithm for finding such a subgraph. However, the next theorem shows that for a slightly smaller value than $m/2$ this can easily be done.

Theorem 5.8 *There is a randomized algorithm that only needs an expected linear amount of time steps to find a bipartite subgraph in a graph $G = (V, E)$, $|E| = m$, with at least $m/4$ edges.*

Proof. Let the random variable Y be defined as $Y = m - X$, i.e. Y counts the number of non-crossing edges. We would like to have Y as small as possible. From the previous proof and the linearity of expectation we know that

$$\mathrm{E}[Y] = m - \mathrm{E}[X] = m - m/2 = m/2 \ .$$

Hence, it follows from the Markov inequality that

$$\Pr[Y \geq 3m/4] \leq \frac{\mathrm{E}[Y]}{3m/4} = \frac{m/2}{3m/4} = \frac{2}{3} .$$

Thus, $\Pr[Y < 3m/4] \geq 1/3$ and therefore $\Pr[X > m/4] \geq 1/3$.

Now, consider the following algorithm:

```
repeat
   perform the random experiement in Theorem 5.7
until at least m/4 edges are crossing
```

Let the random variable T denote the number of rounds the algorithm needs to produce a bipartite graph with at least $m/4$ edges. Since each round has a probability of success of some $p \geq 1/3$, we obtain that

$$\Pr[T = t] = (1-p)^{t-1} \cdot p .$$

Hence,

$$\begin{aligned} \mathrm{E}[T] &= \sum_{t \in \mathbb{N}} t \cdot \Pr[T = t] \\ &= \sum_{t \in \mathbb{N}_0} (t+1) \cdot (1-p)^t p \\ &= p \left(\sum_{t \in \mathbb{N}_0} (1-p)^t \right)^2 = p \cdot \left(\frac{1}{1-(1-p)} \right)^2 = \frac{1}{p} . \end{aligned}$$

Since $p \geq 1/3$, it follows that $\mathrm{E}[T] \leq 3$. □

5.9 Variance

Definition 5.9 *The* variance *(or* dispersion*) of a random variable X, denoted by $\mathrm{V}[X]$, is defined as*

$$\mathrm{V}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] .$$

The number $\sigma = \sqrt{\mathrm{V}[X]}$ is called the standard deviation *of X.*

The following fact lists some basic properties of the variance.

Fact 5.10 *Let X and Y be arbitrary random variables and a, b be arbitrary constants.*

- $\mathrm{V}[a + b \cdot X] = b^2 \cdot \mathrm{V}[X]$.
- *If X is a binary random variable, then $\mathrm{V}[X] = \mathrm{E}[X] \cdot (1 - \mathrm{E}[X])$.*
- *If X and Y are independent, then $\mathrm{V}[X + Y] = \mathrm{V}[X] + \mathrm{V}[Y]$.*
- *If X and Y are independent, then*

$$\frac{\mathrm{V}[X \cdot Y]}{\mathrm{E}[X \cdot Y]^2} = -1 + \left(1 + \frac{\mathrm{V}[X]}{\mathrm{E}[X]^2}\right)\left(1 + \frac{\mathrm{V}[Y]}{\mathrm{E}[Y]^2}\right) .$$

The most well-known tail estimate that uses the variance of a random variable is due to Chebychev.

Theorem 5.11 (Chebychev Inequality) *Let X be an arbitrary random variable. Then, for every $k > 0$,*

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

Proof. From the Markov inequality it follows for all random variables X that

$$\Pr[|X| \geq k] = \Pr[X^2 \geq k^2] \leq \mathrm{E}[X^2]/k^2 .$$

Replacing $|X|$ by $|X - \mathrm{E}[X]|$ yields the theorem. □

5.10 Higher Order Moments

Expectations of powers of random variables are called *moments* and play an essential role in the investigations of probability theory. For us they will be of particular interest in the study of sums of random variables. Given a random variable X, the following expected values of functions of X are often studied:

- $\mathrm{E}[X^k]$, $k \in \mathbb{N}$: kth *moment* of X
- $\mathrm{E}[|X|^k]$, $k \in \mathbb{N}$: kth *absolute moment* of X
- $\mathrm{E}[|X - \mathrm{E}[X]|^k]$, $k \in \mathbb{N}$: kth *absolute central moment* of X
- $\mathrm{E}[\mathrm{e}^{h \cdot X}]$, $h > 0$: *exponential moment* of X

Loéve stated the following result in [Loé77], which easily follows from the Markov inequality.

Theorem 5.12 (Loéve Inequality) *Let X be an arbitrary non-negative random variable and g be an arbitrary function that is non-negative and non-decreasing on $\mathbb{N}_0$. Then, for any $k \geq 0$ with $g(k) > 0$,*

$$\Pr[X \geq k] \leq \frac{\mathrm{E}[g(X)]}{g(k)} .$$

Proof. Since g is non-negative and non-decreasing on $\mathbb{N}_0$, it holds for every $k \geq 0$ that

$$\mathrm{E}[g(X)] = \sum_{x \geq 0} g(x) \cdot \Pr[X = x] \geq g(k) \Pr[X \geq k] .$$

□

5.11 Basic Probability Distributions

In this section we introduce the most important probability distributions used in combination with discrete random variables.

5.11.1 Uniform distribution

A random variable X is called *uniformly distributed* over a finite set M of values in $\mathbb{R}$ if

$$\Pr[X = x] = \frac{1}{|M|}$$

for all $x \in M$. If $M = \{a, a+1, \ldots, b\}$ for some integers a, b with $a < b$, then

$$\mathrm{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \mathrm{V}[X] = \frac{(b-a)^2}{12} + \frac{b-a}{6}$$

5.11.2 Binomial distribution

A random variable X is called *binomially distributed* if there are parameters $p \in [0,1]$ and $n \in \mathbb{N}$ such that

$$\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

for all $k \in \{0, \ldots, n\}$. In this case,

$$\mathrm{E}[X] = n \cdot p \quad \text{and} \quad \mathrm{V}[X] = n \cdot p(1-p) \ .$$

As an example, consider the situation that we have m balls in an urn, of which m_1 are red and m_2 are blue. We draw n balls at random out of this urn, one after the other, replacing each drawn ball. Let the random variable X denote the number of red balls drawn. Then X is binomially distributed with $p = m_1/m$.

5.11.3 Poisson distribution

A random variable X is called *Poisson distributed* if there is a parameter $\lambda > 0$ such that

$$\Pr[X = k] = \frac{\lambda^k}{k!} \cdot \mathrm{e}^{-\lambda}$$

for all $k \in \mathbb{N}_0$. In this case,

$$\mathrm{E}[X] = \lambda \quad \text{and} \quad \mathrm{V}[X] = \lambda \ .$$

The binomial distribution converges towards the Poisson distribution if, in the example above, the number of balls in the urn, m, and the number of draws, n, is increased while keeping the number of red balls constant.

5.11.4 Geometric distribution

A random variable X is called *geometrically distributed* if there is a parameter $p \in [0,1]$ such that

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

for all $k \in \mathbb{N}$. In this case,

$$\mathrm{E}[X] = \frac{1}{p} \quad \text{and} \quad \mathrm{V}[X] = \frac{1-p}{p^2} \ .$$

As an example, suppose that we again have an urn with m balls, of which m_1 balls are red and m_2 balls are blue. We draw a ball at random from this urn and replace it until we draw a red ball. Let the random variable X denote the number of balls drawn. Then X is geometrically distributed with $p = m_1/m$.

5.11.5 Hypergeometric distribution

A random variable X is called *hypergeometrically distributed* if there are parameters N, M, and n such that

$$\Pr[X = k] = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

for all $k \in \{0, \ldots, n\}$. In this case, we obtain for $p = M/N$

$$\mathrm{E}[X] = n \cdot p \quad \text{and} \quad \mathrm{V}[X] = \frac{N-n}{N-1} \cdot n \cdot p(1-p) .$$

As an example, consider the situation that we have N balls in an urn, of which M are red and $N-M$ are blue. We draw n balls at random out of this urn, one after the other, *without* replacing each drawn ball. Let the random variable X denote the number of red balls drawn. Then X is hypergeometrically distributed.

5.12 Chernoff bounds

In the following we study sums of independent binary random variables. Tail estimates for these sums have among many others been investigated by Chernoff [Che52] and are often called *Chernoff bounds*. It has been found convenient to use exponential moments of random variables to derive these tail estimates. The idea of using exponential moments was apparently first used by S. N. Bernstein. His method had a significant influence on deriving tail estimates also for many other problems. We present here a proof given by Hagerup and Rüb [HR90]. Bound (6) was taken from [McD98].

Theorem 5.13 (Chernoff Bound) *If $X_1, \ldots, X_n$ are independent binary random variables, then it holds for all $\delta \geq 0$ that*

$$\begin{aligned} \Pr[S_n \geq (1+\delta)\mu] &\leq \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} && (5)\\ &\leq \mathrm{e}^{-\frac{\delta^2\mu}{2(1+\delta/3)}} && (6)\\ &\leq \mathrm{e}^{-\min[\delta^2,\,\delta]\cdot\mu/3} . \end{aligned}$$

Furthermore, it holds for all $0 \leq \delta \leq 1$ that

$$\begin{aligned} \Pr[S_n \leq (1-\delta)\mu] &\leq \left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} && (7)\\ &\leq \mathrm{e}^{-\delta^2\mu/2} . \end{aligned}$$

Proof. We will only show (5) and (7). First we present a proof for inequality (5). For all $i \in [n]$, let $p_i = \mathrm{E}[X_i]$. According to the Loéve inequality (see Theorem 5.12) we obtain that, for any function $g(x) = e^{h \cdot x}$ with $h > 0$ and any $\delta \geq 0$,

$$\Pr[S_n \geq (1+\delta)\mu] \leq \mathrm{e}^{-h(1+\delta)\mu} \cdot \mathrm{E}[\mathrm{e}^{h \cdot S_n}] . \quad (8)$$

Since $X_1, \ldots, X_n$ are independent, it follows that

$$\begin{aligned} \mathrm{E}[\mathrm{e}^{h\cdot S_n}] &= \mathrm{E}[\mathrm{e}^{h(X_1+\ldots+X_n)}] = \mathrm{E}[\mathrm{e}^{h\cdot X_1}\cdots\mathrm{e}^{h\cdot X_n}] = \textstyle\prod_{i=1}^n \mathrm{E}[\mathrm{e}^{h\cdot X_i}]\\ &= \prod_{i=1}^n (p_i\mathrm{e}^h + (1-p_i)) = \textstyle\prod_{i=1}^n (1 + p_i(\mathrm{e}^h - 1))\\ &\leq \prod_{i=1}^n \mathrm{e}^{p_i(\mathrm{e}^h-1)} \quad \text{since } 1+x \leq \mathrm{e}^x \text{ for all } x\\ &= \mathrm{e}^{\mu(\mathrm{e}^h-1)} . \end{aligned}$$

This yields together with inequality (8) that

$$\Pr[S_n \geq (1+\delta)\mu] \leq \mathrm{e}^{-h(1+\delta)\mu} \cdot \mathrm{e}^{\mu(\mathrm{e}^h-1)} = \mathrm{e}^{-(1+h(1+\delta)-\mathrm{e}^h)\mu} \ . \tag{9}$$

The right-hand side of (9) attains its minimum at $h = h_0$, where $h_0 = \ln(1+\delta)$. Inserting this in (9) yields

$$\Pr[S_n \geq (1+\delta)\mu] \leq (1+\delta)^{-(1+\delta)\mu} \cdot \mathrm{e}^{\delta \cdot \mu} = \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \ .$$

This completes the proof for inequality (5). The proof of (7) can be done in a similar way. □

References

[Che52] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.

[HR90] T. Hagerup and C. Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33:305–308, 1989/90.

[Loé77] M. Loéve. *Probability theory.* Springer Verlag, New York, 1977.

[McD98] McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–247. Springer Verlag, Berlin, 1998.