Summary of what has been covered in math302

Week 1.

Essentially we cover Sec. 1.2

Probability theory: early motivations given by gambling problems (cf textbook p. 30)

**Def (book p.18).** Sample space \( \Omega = \) set of all possible outcomes of an experiment, for now we will only consider \( \Omega \) finite (e.g. for tossing a die, \( \Omega = \{1, 2, 3, 4, 5, 6\} \)), or infinite countable, e.g. the lifetime of a bulb in number of days (\( \Omega = \mathbb{N} \)).

Any subset \( E \) of \( \Omega \) is called an event, e.g., the event “outcome of die is even” is \( E = \{2, 4, 6\} \). Recall that \( E \subset F \) means that all outcomes of \( E \) are in \( F \) (inclusion).

Operations on events (book p.21):

- **Union:** \( E \cup F \), \( E_1 \cup \cdots \cup E_n \) (union of \( n \) sets, also denoted \( \bigcup_{i=1}^{n} E_i \)), \( E_1 \cup E_2 \cup E_3 \cup \cdots \) (union of infinitely many sets, also denoted \( \bigcup_{i=1}^{\infty} E_i \)),

- **Intersection:** \( E \cap F \), \( E_1 \cap \cdots \cap E_n \) (intersection of \( n \) sets, also denoted \( \bigcap_{i=1}^{n} E_i \)), \( E_1 \cap E_2 \cap E_3 \cap \cdots \) (intersection of infinitely many sets, also denoted \( \bigcap_{i=1}^{\infty} E_i \)),

- \( E - F \) denotes the set of outcomes in \( E \) but not in \( F \),

- \( \tilde{E} \) is the complement of \( E \), that is, \( \tilde{E} = \Omega - E \).

Arithmetic rules on events (not all explicitly stated in book):

- **Associativity:** \((E \cup F) \cup G = E \cup (F \cup G) = E \cup F \cup G\), \((E \cap F) \cap G = E \cap (F \cap G) = E \cap F \cap G\),

- **Distributivity:** \((E \cup F) \cap G = (E \cap G) \cup (F \cap G)\), \((E \cap F) \cup G = (E \cap G) \cup (F \cup G)\),

- if \( F = \tilde{E} \) then \( F = E \).

- **De Morgan’s laws:** \( \bigcup_{i=1}^{n} E_i = \bigcap_{i=1}^{n} \tilde{E}_i \), \( \bigcap_{i=1}^{n} E_i = \bigcup_{i=1}^{n} \tilde{E}_i \).

We talk of \( \Omega \) as a probability space if to each event \( E \) is assigned a quantity \( P(E) \), called the probability of \( E \), so as to satisfy the following axioms (cf Kolgomorov’s axiomatic):

1. \( 0 \leq P(E) \leq 1 \),
2. \( P(\Omega) = 1 \),
3. for \( E, F \) disjoint (i.e., \( E \cap F = \emptyset \)),
   \[ P(E \cup F) = P(E) + P(F) \]

   More generally, for \( E_1, E_2, \ldots \) disjoint (i.e., \( E_i \cap E_j = \emptyset \) for \( i \neq j \)),
   \[ P\left( \bigcup_{i=1}^{\infty} E_i \right) = \sum_{i=1}^{\infty} P(E_i) \]

If \( \Omega \) is finite or infinite countable (discrete setting), a probability space is typically obtained from a distribution function (book p.19), i.e., a function \( m \) from \( \Omega \) to \([0,1]\) such that \( \sum_{\omega \in \Omega} m(\omega) = 1 \). Then the induced probability assignment is \( P(E) = \sum_{\omega \in E} m(\omega) \), and clearly satisfies the 3 axioms. In the case where \( \Omega \) is finite, we will very often consider the uniform distribution function, where \( m(\omega) = \frac{1}{|\Omega|} \) for
each $\omega \in \Omega$ (with the notation $|S|$ for the cardinality of a finite set $S$). Then we have $P(E) = \frac{|E|}{|\Omega|}$.

**Rk:** In the book (Theo.1.1 p. 22) they define a probability space as induced by a distribution function, and show that the axioms hold. We prefer the more usual (and more general, it also covers the continuous case to be seen later) definition in terms of axioms.

Properties that follow from the axioms (in all generality, not necessarily induced by a distribution function):

- $P(\tilde{E}) = 1 - P(E)$, in particular $P(\emptyset) = 1 - P(\Omega) = 0$.
- If $E \subset F$, then $P(E) \leq P(F)$,
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

**Week 2.**

We now move to Chapter 3 (on combinatorics) of the textbook. We focus here on the case where $\Omega$ is finite, and consider in that case the uniform distribution, assigning probability $\frac{|E|}{|\Omega|}$ to any event. Since the probability is in terms of cardinality quantities, we can compute probabilities by counting (an often difficult task).

We start discussing permutations (book p.79). For a finite set $A$ of objects (numbers, letters, etc.), a permutation on $A$ is a mapping $\sigma$ from $A$ to $A$ that is 1-to-1, where 1-to-1 means that for every $y \in A$ there is a unique $x \in A$ such that $\sigma(x) = y$ (otherly stated, every $y \in A$ has a unique preimage by $\sigma$).

Examples. For $A = \{a, b\}$, the mapping sending $a$ to $b$ and $b$ to $b$ is not a permutation ($b$ has two preimages), but the mapping sending $a$ to $b$ and $b$ to $a$ is a permutation. For $A = \{1, 2, 3, 4, 5, 6\}$, the mapping sending 1 to 3, 2 to 5, 3 to 6, 4 to 4, 5 to 2, and 6 to 1 is a permutation.

A permutation has two classical representations:

- as an (unordered) collection of directed cycles, on the last example above,
- as a word giving a linear arrangement of the elements of $A$; for this we need to consider $A$ as endowed with an order (e.g. if $A$ is a set of letters, we order the letters in alphabetic order, if $A$ is a set of integers, we order them in increasing order), on the last example above the word is $w = 3\, 5\, 6\, 4\, 2\, 1$, e.g. the 5th element of the word being 2 means that the permutation maps 5 to 2. Similarly the word $b\, c\, a$ represents the permutation mapping $a$ to $b$, $b$ to $c$, and $c$ to $a$.

The second point of view (word representation) lets us easily see that there are $n!$ permutations on an $n$-element set. Indeed, there are $n$ ways to choose the first symbol, $n - 1$ ways to choose the second symbol, ..., 2 ways to choose the $(n - 1)$th symbol, 1 way to choose the $n$th symbol, hence a total of $n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1 = n!$ possibilities.
For σ a permutation on an n-element set A, we call a fixed point of σ an element x ∈ A such that σ(x) = x. For n ≥ 1, we denote by \( \mathfrak{S}_n \) the set of permutations on the set \{1, \ldots, n\}.

We can see \( \mathfrak{S}_n \) as a sample space and look at events. For σ a random permutation in \( \mathfrak{S}_n \) (under the uniform distribution),

- what is the probability that 1 is a fixed point? More generally, for 1 ≤ i ≤ n what is the probability that i is a fixed point?
- what is the probability that σ has no fixed point (cf hat check problem, book p.85-86)?

For the first problem, it is easy to see that the number of permutations fixing 1 is \((n-1)!\) (indeed we have 1 possibility for the first symbol, \(n-1\) possibilities for the second symbol, etc.), so that the probability is \(\frac{(n-1)!}{n!} = \frac{1}{n}\) (similarly for all \(i \in \{1, \ldots, n\}\) there are \((n-1)!\) permutations fixing \(i\), hence the probability of fixing \(i\) is also \(\frac{1}{n}\)).

For the second problem, we do it by counting how many permutations in \( \mathfrak{S}_n \) have no fixed points. Such permutations are called derangements. Note that, in the decomposition of a permutation σ into cycles, a fixed point is a cycle of length 1. Hence σ is a derangement if and only if all its cycles have size larger than 1. For \(n \geq 1\) we call \(n\)-derangement a derangement in \( \mathfrak{S}_n \). We denote by \(w_n\) the number of \(n\)-derangements, so that the searched probability \(p_n\) is given by

\[
p_n = \frac{w_n}{n!}.
\]

We have the first values \(w_1 = 0, w_2 = 1, w_3 = 2, w_4 = 9, \ldots\).

We first note that for \(i \geq 2\) there are as many \(n\)-derangements mapping 1 to 2 as there are mapping 1 to \(i\), indeed we can turn a derangement mapping 1 to 2 into a derangement mapping 1 to \(i\) by just swapping the labels 2 and \(i\) in the cyclic representation of the derangement, as in the next figure where \(n = 5\) and \(i = 4\):

\begin{itemize}
  \item maps 1 to 2
  \item swap labels \{2, 4\}
  \item label becomes 4
  \item label becomes 2
\end{itemize}

If we denote by \(a_n\) the number of \(n\)-derangements mapping 1 to 2, and for \(i \in \{2, \ldots, n\}\) denote by \(w_n^{(i)}\) the number of \(n\)-derangements mapping 1 to \(i\), the previous argument gives \(w_n^{(i)} = a_n\) for any \(i \in \{2, \ldots, n\}\), hence ¹

\[
w_n = \sum_{i=2}^{n} w_n^{(i)} = \sum_{i=2}^{n} a_n = (n-1)a_n.
\]

We now show that \(a_n = w_{n-1} + w_{n-2}\). Let σ be an \(n\)-derangement mapping 1 to 2. Let \(e\) be the cycle containing 1 (and 2). If \(e\) has size 2, we can just see σ as the cycle (1, 2) together with a derangement (formed by the other cycles) on the \((n-2)\)-element set \{3, 4, \ldots, n\}. Hence there are \(w_{n-2}\) \(n\)-derangements mapping

¹We use the notation \(\sum_{i=0}^{n} d_i\) for \(d_0 + d_1 + \cdots + d_n\), and similarly the notation \(\prod_{i=0}^{n} d_i\) for the product \(d_0 \cdot d_1 \cdot d_2 \cdots d_n\).
1 to 2 and where $c$ has size 2. If $c$ has size larger than 2, we obtain from $\sigma$ a derangement $\sigma'$ on the $(n - 1)$-element set $\{2, 3, \ldots, n\}$ by “erasing” 1 on $c$, as in the example below:

![Diagram of derangement on \{2, 3, 4, 5, 6, 7\}]

In doing this there is no loss of information (to recover $\sigma$ from $\sigma'$ we just insert 1 on the cycle of $\sigma'$ containing 2, just before 2). Hence there are $w_{n-1}$ n-derangements mapping 1 to 2 and where $c$ has size larger than 2. Overall we obtain $a_n = w_{n-1} + w_{n-2}$, giving the recurrence (due to Euler),

$$w_n = (n - 1) \cdot (w_{n-2} + w_{n-1}) \quad \text{for each } n \geq 2,$$

with initial conditions $w_0 = 1, w_1 = 0$. The recurrence makes it possible to quickly compute the successive values of $w_n$: $w_2 = 1, w_3 = 2, w_4 = 9, w_5 = 44, w_6 = 265, w_7 = 1854, w_8 = 14833, w_9 = 133496, w_{10} = 1334961, \ldots$. We observe that the probability $p_n = w_n/n!$ seems to converge fast to a certain value $\approx 0.368$. A first step to understand this converging behaviour is to deduce from the recurrence (1) (see below) the following expression of $p_n$:

$$p_n = \frac{(-1)^2}{2!} + \frac{(-1)^3}{3!} + \cdots + \frac{(-1)^n}{n!}.$$

We then recognize the $n$th expansion of the exponential function $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$ at $x = -1$. Hence $p_n$ converges as $n \to \infty$ to the constant $e^{-1} = 1/c \approx 0.368$.

Proof that (1) implies (2). Note that if we subtract $n \cdot w_{n-1}$ from both sides of (1), we obtain

$$w_n - n \cdot w_{n-1} = (n - 1) \cdot (w_{n-2} + w_{n-1}) - n \cdot w_{n-1} = (n - 1)w_{n-2} - w_{n-1},$$

hence the sequence $u_n$ defined as $u_n = w_n - n \cdot w_{n-1}$ satisfies $u_n = -u_{n-1}$, and it has initial value $u_1 = w_1 - 0 = 1$, so for $n = 1, 2, 3, \ldots$, the sequence $u_n$ alternates between -1 and +1, and we have $u_n = (-1)^n$. Hence $\frac{u_n}{n!} = \frac{(-1)^n}{n!}$. Note also that

$$\frac{u_n}{n!} = \frac{w_n - n \cdot w_{n-1}}{n!} = \frac{w_n}{n!} - \frac{w_{n-1}}{(n-1)!} = p_n - p_{n-1}.$$

We thus have

$$p_n = p_1 + (p_2 - p_1) + (p_3 - p_2) + \ldots + (p_n - p_{n-1}) = 0 + \frac{(-1)^2}{2!} + \frac{(-1)^3}{3!} + \cdots + \frac{(-1)^n}{n!}.$$

We now move to counting samples (ordered or not) from a finite set $A$ of $n$ elements (Chap 3, p.80 and Section 3.2).

For $k \geq 1$ an ordered $k$-sample from $A$ (also called a $k$-permutation in book p.80) is an ordered listing of $k$ different elements from $A$. For example for $A = \{a, b, c\}$ the ordered 2-samples are $\{ab, ac, ba, bc, ca, cb\}$. In general, for an $n$-element set $A$, the number of ordered $k$-samples is $n \cdot (n - 1) \cdots (n - k + 1)$, since there are $n$ possibilities for choosing the first element, $n - 1$ possibilities for the 2nd elements, $n - 2$ possibilities for the 3rd element, \ldots, $n - k + 1$ possibilities for the $k$th element. This product also writes as $n!/(n-k)!$, and we denote it by $(n)_k$. 

Related to it is the birthday paradox (book p.77-79). We are interested in the probability \( P(E) \) that \( k \) people in a room have distinct birthdays (ignore Feb. 29th). There are \( 365^k \) ways to assign birthdays to the \( k \) people (365 possibilities for person number 1, 365 possibilities for person number 2, ..., 365 possibilities for person number \( k \)), but there are only \( 365 \cdot 364 \cdot 363 \cdots (365 - k + 1) \) ways to assign distinct birthdays to the \( k \) people (this can be seen as choosing an ordered \( k \)-sample of birthdays, person number 1 has the birthday given by the first element from the sample, person number 2 has the birthday given by the second element from the sample, etc.). Hence we have

\[
P(E) = \frac{(365)_k}{365^k} = \prod_{j=0}^{k-1} 365 - j = \prod_{j=0}^{k-1} 1 - \frac{j}{365}.
\]

Using this formula we can compute \( P(E) \) and observe that it becomes smaller than \( 1/2 \) at \( k = 23 \) already, and is only about 0.03 for \( k = 50 \) (book table p.79). To have a better understanding of the transition at \( k = 23 \), let us approximate \( P(E) \). Using \( e^x \approx 1 + x \) for small \( x \) (say \( |x| \leq 1/10 \)), we have

\[
P(E) \approx \prod_{j=0}^{k-1} e^{-j/365} = e^{-\sum_{j=0}^{k-1} j/365}.
\]

Now we use the formula \( \sum_{j=0}^{k-1} j = k(k-1)/2 \) to conclude that

\[
P(E) \approx e^{-k(k-1)/(2 \cdot 365)},
\]

and thus if we call \( y \) the value such that \( e^{-y} = 1/2 \) (i.e., \( y = \log(2) \)), the transition occurs for \( k \) such that \( k(k-1)/(2 \cdot 365) \approx y \), i.e., \( k(k-1) \approx 2 \cdot \log(2) \cdot 365 \). The solution is \( \approx \sqrt{2 \cdot \log(2) \cdot 365} \approx 22.5 \), and indeed as we have seen above, \( P(E) \) becomes smaller than \( 1/2 \) when \( k \) reaches 23. More generally we can look at the problem of throwing \( k \) balls into \( n \) urns (\( n = 365 \) for the birthday problem) and look at the probability that no two balls fall in the same urn. It can similarly be seen (see Exercise 19 p.90) that \( P(E) \) becomes smaller than \( 1/2 \) when \( k \) becomes larger than a threshold value that is \( \approx \sqrt{2 \cdot \log(2) \cdot n} \).

This analysis has applications for instance in computer science (storing and quickly retrieving data in a table) to estimate how many entries can be stored in a hash table without having too many conflicts (a conflict meaning that two entries are stored at the same position in the hash table, many conflicts make it slower to search for an entry). If \( n \) denotes the number of positions in the hash table, the analysis shows that conflicts typically start to appear when the number of stored objects reaches \( \sqrt{2 \log(2) n} \).

**Week 3.**

For \( A \) an \( n \)-element set, a combination (book Sec 3.2 p.92) is a subset of elements from \( A \). For \( 0 \leq k \leq n \) we talk of a \( k \)-combination if the combination has \( k \) elements. For instance \( A = \{a, b, c\} \) has three 2-combinations, which are \( \{ab, ac, bc\} \), but has 6 ordered 2-samples, which are \( \{ab, ac, ba, bc, ca, cb\} \) (each 2-combination yields 2 ordered 2-samples, since there are 2! ways to order the 2 elements). More generally, if \( C_n^k \) denotes the number of \( k \)-combinations from an \( n \)-element set \( A \), and \( (n)_k \)
denotes the number of ordered \( k \)-samples, then we have
\[
(n)_k = C_n^k \ast k!,
\]
since each ordered \( k \)-sample is obtained by choosing a \( k \)-combination (the elements of the sample) and then an order among the \( k \) elements \((k!)\) possibilities. We also have seen that \((n)_k = n!/(n-k)!\), from which we conclude that \( C_n^k = \frac{n!}{k!(n-k)!} \). The coefficient \( \frac{n!}{k!(n-k)!} \) is denoted \( \binom{n}{k} \), these are called the binomial coefficients. Note that \( \binom{n}{0} = \binom{n}{n} = 1 \) (with the convention \( 0! = 1 \), we have \( \binom{n}{0} = n!/(0! \ast n!) = 1 \), and similarly for \( \binom{n}{n} \)).

The binomial coefficients have the following properties:
- Symmetry: \( \binom{n}{k} = \binom{n}{n-k} \) for \( 0 \leq k \leq n \),
- Recurrence (book Theo 3.4 p.93): for \( 0 < k < n \), \( \binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \).

The recurrence then makes it possible to compute a table of the binomial coefficients, for \( n = 1, 2, 3, \text{etc.} \) (cf book p.94).

Note that \( \binom{n}{k} \) also gives the number of words with two letters (say \( a \) or \( b \)) of length \( n \) with \( k \) occurrences of \( a \); indeed the positions of the \( a \)'s in such a word form a subset of size \( k \) of \( \{1, \ldots, n \} \). From this observation we can easily prove the binomial theorem (which generalizes to any \( k \) form a subset of size \( k \) of \( \{1, \ldots, n \} \)). Thus we can easily prove the binomial theorem (which generalizes to any \( k \) form a subset of size \( k \) of \( \{1, \ldots, n \} \)).

\[
\text{for any } n \geq 1, \quad (x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.
\]
Indeed we can think of \((x+y)^n\) as \((x+y)(x+y)\cdots(x+y)\), and when we expand the product into a sum, for each of the \( n \) factors \((x+y)\) we have to choose if we take \( x \) or \( y \), so when we expand we get a sum over all “words” of length \( n \) with letters \( x \) or \( y \), for instance
\[
(x+y)^3 = (x+y)(x+y)(x+y) = xxy + xyx + yxx + yxy + yyx + yyy.
\]
We then group the terms with the same number of \( x \)'s, and from the observation above, there are \( \binom{n}{k} \) terms with \( k \) occurrences of \( x \), which explains the terms \( \binom{n}{k} x^k y^{n-k} \) in the binomial theorem.

We now move to Chapter 4: **Conditional probabilities.**
For \( \Omega \) a probability space, \( F \) an event such that \( P(F) > 0 \), and \( E \) another event, we define
\[
P(E|F) = \frac{P(E \cap F)}{P(F)}.
\]
The quantity \( P(E|F) \) is called the conditional probability that \( E \) occurs, “knowing” that \( F \) occurs.

For instance if we roll two dice, let \( E \) be the event that the sum of the two dice gives 8, and \( F \) the event that the first die gives 3; then what is the conditioned probability \( P(E|F) \)? The event \( E \cap F \) is the event that the first die gives 3 and the second die gives 5, hence \( P(E \cap F) = 1/36 \). And we have \( P(F) = 1/6 \) since the first die has equal chances for all 6 outcomes. Hence \( P(E|F) = \frac{1/36}{1/6} = 1/6 \). We could have just argued with the sample space reduced to the outcome of the second die, by saying that once the first die is 3, out of the 6 possibilities \{1, 2, 3, 4, 5, 6\}
for the second die, only the outcome 5 gives a sum equal to 8, hence a conditioned probability of 1/6.

Conditioned probabilities also help to compute probabilities of events that are the conjunction of several constraints: for instance we can revisit the birthday paradox.

If we have \( k \) people \( P_1, \ldots, P_k \), for \( i \in \{1, \ldots, k\} \) we call \( E_i \) the event that the \( i \) first persons \( P_1, \ldots, P_i \) have distinct birthdays. Note that

\[
P(E_i | E_{i-1}) = \frac{365 - (i - 1)}{365}.
\]

In addition, since \( E_i \subset E_{i-1} \) (which implies \( E_i \cap E_{i-1} = E_i \)), we have the “chain-formula” \( P(E_i) = P(E_i | E_{i-1}) \cdot P(E_{i-1}) \). We thus have

\[
P(E_k) = P(E_k | E_{k-1}) \cdot P(E_{k-1} | E_{k-2}) \cdots P(E_1) = \frac{365 - (k - 1)}{365} \cdot \frac{365 - (k - 2)}{365} \cdots \frac{365}{365}
\]

and we recover the already found expression \( \prod_{j=0}^{k-1} 1 - \frac{j}{365} \) for the probability that \( k \) people have distinct birthdays.

An important class of problems dealing with conditional probabilities are Bayes type problems.

For instance, population at UBC is divided into students (say, 70%), instructors (say, 20%) and admin. staff (say, 10%). Say that 50% of the students, 30% of the instructors, and 40% of the admin. go to UBC by bike. Now, what is the probability of the next person arriving to campus by bike of being a student? The categories are \( S, I, A \) (students, instructors, admin.), the event \( B \) is “comes by bike”. We know \( P(S) = 0.7 \), \( P(I) = 0.2 \), \( P(A) = 0.1 \), and we also have \( P(B|S) = 0.5 \), \( P(B|I) = 0.3 \), \( P(B|A) = 0.4 \); and we now want to compute \( P(S|B) \), which by definition is

\[
P(S|B) = \frac{P(S \cap B)}{P(B)}.
\]

We compute \( P(S \cap B) \) as \( P(S \cap B) = P(B|S) \cdot P(S) = 0.5 \cdot 0.7 = 0.35 \), and similarly \( P(I \cap B) = P(B|I) \cdot P(I) = 0.2 \cdot 0.3 = 0.06 \) and \( P(A \cap B) = P(B|A) \cdot P(A) = 0.1 \cdot 0.4 = 0.04 \). Hence we have
\[
P(B) = P(S \cap B) + P(I \cap B) + P(A \cap B) = 0.35 + 0.06 + 0.04 = 0.45,
\]
and we conclude that
\[
P(S|B) = \frac{0.35}{0.45} = \frac{7}{9}.
\]

In this type of problems, we have a partition of the sample space \( \Omega \) into categories \( C_1, \ldots, C_m \), and we look at a particular event \( E \). We know the probabilities \( P(C_1), \ldots, P(C_m) \) and also the conditional probabilities \( P(E|C_1), \ldots, P(E|C_m) \) that the event \( E \) occurs in each given category (in the previous example the 3 categories are \{students, instructors, staff\}, and the event \( E \) is \{comes by bike\}).

What we now want to know is: “knowing that \( E \) occurs, what is the probability of being in a certain category?”, that is, we want to compute \( P(C_i|E) \) for each \( i \in \{1, \ldots, m\} \). We can use the general formula

\[
P(C_i|E) = \frac{P(C_i \cap E)}{P(E)} = \frac{P(E|C_i) \cdot P(C_i)}{\sum_{j=1}^{m} P(E|C_j) \cdot P(C_j)}.
\]
Week 4.

We now define the independence of events. Given a probability space, two events $E$ and $F$ are called independent if $P(E \cap F) = P(E) \cdot P(F)$; when $P(F) > 0$ this is equivalent to $P(E|F) = P(E)$ (since $P(E|F) = P(E \cap F)/P(F)$), that is, “knowing that $F$ occurs does not affect the probability that $E$ occurs”. Similarly if $P(E) > 0$ independence of $E$ and $F$ is equivalent to $P(F|E) = P(F)$.

More generally, for $n \geq 2$, we say that events $E_1, \ldots, E_n$ are independent if for any subset $1 \leq i_1 < \cdots < i_k \leq n$ of the indices, we have

\[ P(E_{i_1} \cap \cdots \cap E_{i_k}) = P(E_{i_1}) \cdots P(E_{i_k}), \]

for instance for $n = 3$, events $E, F, G$ are independent if all 3 pairs $\{E, F\}, \{E, G\}, \{F, G\}$ are independent and moreover

\[ P(E \cap F \cap G) = P(E) \cdot P(F) \cdot P(G). \]

Note that 3 events can be pairwise independent, but not overall independent, for instance the events $E = \{Alice & Betty have same birthday\}$, $F = \{Carol & Betty have same birthday\}$, and $G = \{Alice & Carol have same birthday\}$. Then clearly $P(E) = P(F) = P(G) = \frac{1}{365}$. And also $P(E|F) = \frac{1}{365}$, so that $P(E|F) = P(E)$, ensuring that $E, F$ are independent (similarly, $E, G$ are independent and $F, G$ are independent). But for instance $P(E|F \cap G) = P(E|3 \text{ birthdays are equal}) = 1$, which is different from $P(E)$, hence $E$ is not independent from $F \cap G$, which should be the case (exercise!) if $E, F, G$ were independent. So we have an example of 3 events that are pairwise independent but are not independent.

We now move to the concept of random variable (not really defined in the book, even if they talk about it from the beginning, important distributions of discrete random variables are given in Section 5.1). Given $\Omega$ a probability space, define a random variable $X$ as a function from $\Omega$ to $\mathbb{R}$. In a physics context, we imagine we perform an experiment whose outcome is random, and we measure a certain parameter of the outcome. As a first simple example, imagine we roll two dice and the “parameter” is the sum of the two dice. In mathematical terms the sample space is the set of 36 pairs $(i, j)$ where $1 \leq i \leq 6$ and $1 \leq j \leq 6$ and the random variable associates to each pair $(i, j)$ the sum $i + j$.

The probability mass function of a random variable $X$ is the function $p(x) = P(X = x)$, and the range of $X$ is the set of $x$ such that $p(x) \neq 0$. For instance for the two-dice example, the range of $X$ is the set of integers from 2 to 12, and the probability mass function is

\[ p(x) = \frac{\text{number of pairs } (i, j) \text{ such that } i + j = x}{36}. \]

We find $p(2) = p(12) = 1/36$, $p(3) = p(11) = 2/36 = 1/18$, $p(4) = p(10) = 3/36 = 1/12$, $p(5) = p(9) = 4/36 = 1/9$, $p(6) = p(8) = 5/36$, $p(7) = 6/36 = 1/6$.

In general, a probability mass function always satisfies $0 \leq p(x) \leq 1$ for all $x$, and $\sum_x p(x) = 1$. We now list some important examples of random variables:

The Bernoulli law of parameter $p$ (with $0 < p < 1$) is the random variable $X$ (also denoted Bern($p$)) with range $\{0, 1\}$ such that $P(Z = 1) = p$ and $P(Z = 0) = 1 - p$ (we associate 1 to “success” and 0 to “failure”).
The **Binomial law** of parameters \( n \geq 1 \) and \( p \) (with \( 0 < p < 1 \)) is the random variable \( X \) (also denoted \( \text{Bin}(n, p) \)) that counts the number of successes when performing successively (and independently) \( n \) experiments, each having probability \( p \) of success. The range of \( X \) is the set of integers from 0 to \( n \), and we have for each \( 0 \leq k \leq n \),

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

where \( \binom{n}{k} \) is the number of ways to choose the \( k \) positions of successes, and then by independence of the experiments, each fixed sequence of successes and failures, with a total of \( k \) successes, has probability \( p^k (1 - p)^{n-k} \). For instance for \( n = 3 \) and \( k = 2 \), the \( \binom{3}{2} \) sequences are \( SSF \), \( SFS \), \( FSS \) each having probability \( p^2 (1 - p) \). Note also that \( X \) can be seen as the sum \( Z_1 + \cdots + Z_n \) of \( n \) independent copies of a \( \text{Bern}(p) \) random variable (where \( Z_i \) is related to the \( i \)th experiment).

The **Geometric law** of parameter \( p \) (with \( 0 < p < 1 \)) is the random variable \( X \) (also denoted \( \text{Geom}(p) \)) that counts how many trials are needed until getting a first success, when repeating an experiment (e.g. tossing a coin) that has probability \( p \) of success at each trial. This time the range of \( X \) is the set of all positive integers. Moreover we have for any \( k \geq 1 \),

\[
P(X = k) = (1 - p)^{k-1} p,
\]

since the event \( \{X = k\} \) means that the first \( k - 1 \) trials are ‘fail’ (each having probability \( 1 - p \)) and the \( k \)th trial is ‘success’ (having probability \( p \)). Similarly we have for any \( k \geq 0 \)

\[
P(X > k) = (1 - p)^k,
\]

since the event \( \{X > k\} \) means that the first \( k \) trials are ‘fail’.

For instance, if you ask random people their birthday until getting a match with your birthday, then the number of people you need to ask is a \( \text{Geom}(\frac{1}{365}) \) random variable. If you have decided in advance to ask \( k \) people, then the probability to have no match is \( (1 - \frac{1}{365})^k \), this is smaller than \( 1/2 \) if and only if \( k \log(1-1/365) \leq \log(1/2) \), that is, \( k \geq \log(2) / \log(365/364) \approx 252.7 \). So the threshold value is larger than in the birthday paradox (where it is \( \approx 23 \)), the point is that here you wait until having a match with *your* birthday, whereas in the birthday paradox, we look for matches among all pairs of people.
Indeed we have $E$ and $X$ that share the same probability space $\Omega$, with $p(x) = P(X = x)$ the probability mass function of $X$, define the expected value (or average value) of $X$ as

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

If $m(\omega)$ denotes the distribution function for $\Omega$, then there is also the following expression for $E[X]$ (not given in the book but convenient to prove properties of expected values):

$$E[X] = \sum_{\omega \in \Omega} m(\omega)X(\omega)$$

For instance if we toss a fair coin 3 times and let $X$ denote the total number of $H$, then $p(0) = p(3) = 1/8$, and $p(1) = p(2) = (\frac{1}{2})^2 = 3/8$, so that $E[X] = 0 * (1/8) + 1 * (3/8) + 2 * (3/8) + 3 * (1/8) = 3/2$. The sample space is $\Omega = \{HHH, HHT, HTT, TTH, TTT\}$, under the uniform distribution (1/8 for each outcome). And using the second expression for $E[X]$ we find

$$E[X] = \frac{1}{8}(3 + 2 + 2 + 1 + 1 + 0) = \frac{3}{2}.$$

That the two expressions for $E[X]$ are the same follows from grouping outcomes that share the same $X$-value:

$$\sum_{\omega \in \Omega} m(\omega)X(\omega) = \sum_x \sum_{\omega \in \Omega: X(\omega) = x} m(\omega)X(\omega) = \sum_x \left( \sum_{\omega \in \Omega: X(\omega) = x} m(\omega) \right) x = \sum_x p(x)x$$

From the second expression of $E[X]$ we easily obtain the following properties:

- For any random variable $X$ and for any constant $c \in \mathbb{R}$, $E[cX] = c \cdot E[X]$,

Indeed we have

$$E[cX] = \sum_{\omega \in \Omega} m(\omega)(cX(\omega)) = c \sum_{\omega \in \Omega} m(\omega)X(\omega) = cE[X],$$

and

$$E[X + Y] = \sum_{\omega \in \Omega} m(\omega)(X(\omega)+Y(\omega)) = \sum_{\omega \in \Omega} m(\omega)X(\omega) + \sum_{\omega \in \Omega} m(\omega)Y(\omega) = E[X]+E[Y].$$

For $X$ a random variable, with probability mass function $p(x) = P(X = x)$, and for $g(x)$ a function from $\mathbb{R} \rightarrow \mathbb{R}$, we also have the following rule to compute $E[g(X)]$, called the law of the unconscious statistician:

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

Indeed, with the notation $p(x) = P(X = x)$, we have

$$E[g(X)] = \sum_{\omega \in \Omega} m(\omega)g(X(\omega)) = \sum_x \sum_{\omega \in \Omega: X(\omega) = x} m(\omega)g(X(\omega))$$

$$= \sum_x \left( \sum_{\omega \in \Omega: X(\omega) = x} m(\omega) \right) g(x) = \sum_x p(x)g(x).$$
For the Bern($p$) random variable $Z$ (Bernoulli random variable), we have
\[ E[Z] = P(Z = 1) \cdot 1 + P(Z = 0) \cdot 0 = p. \]

For the Bin($n,p$) random variable $X$ (binomial random variable), we have $X = Z_1 + \cdots + Z_n$, the sum of $n$ independent Bern($p$) random variables, hence
\[ E[X] = n \cdot E[\text{Bern}(p)] = np. \]

Before computing the expected value of the geometric random variable, let us introduce the concept of “conditional expected value”. For $X$ a random variable and $F$ an event (with $P(F) > 0$) we define the expected value of $X$ knowing $F$ as
\[ E[X|F] := \sum_x P(X = x|F)x. \]

If the sample space $\Omega$ is partitioned as (disjoint union) $\Omega = C_1 \cup \cdots \cup C_m$ then we have the formula (book Theo. 6.5 p.239)
\[ E[X] = \sum_{i=1}^m P(C_i) \cdot E[X|C_i], \]

For the Geom($p$) random variable $X$ we can condition on what happens in the first toss, either “heads” (event noted $H$) or “tails” (event noted $T$). The above formula thus gives
\[ E[X] = E[X|H] \cdot P(H) + E[X|T] \cdot P(T) = E[X|H] \cdot p + E[X|T] \cdot (1 - p). \]

Note that, conditioning on $H$, we have $X = 1$, and conditioning on $T$, we have that $X$ is distributed as $1 + \text{Geom}(p)$ (we have the first $T$ toss, plus the subsequent tosses until having 'heads' for the first time). Hence
\[ E[X] = p + E[1 + X] \cdot (1 - p) = 1 + (1 - p) \cdot E[X], \]

so that $E[X] = 1/p$. For instance if you ask random people their birthday until having a match with your birthday, then the expected value of the number of people you ask is 365.

For $X$ a random variable we use the notation $\mu$ for the expected value of $X$. Now we would like to have a parameter that quantifies if the random variable is concentrated around its expected value or not. We define the variance of $X$ as
\[ V[X] = E[(X - \mu)^2], \]

intuitively the smaller the variance, the more the random variable is concentrated around its expected value (another candidate could be to take $E[|X - \mu|]$, but the variance has nicer mathematical properties, in particular with respect to the sum of independent random variables).

Note that if $X$ is expressed in a certain unit (e.g. meters) then $V[X]$ is expressed in the squared units (e.g. meters$^2$), so it is also convenient to define the so-called standard deviation $D[X]$ as $D[X] = \sqrt{V[X]}$, the notation $\sigma$ is often used for $D[X]$.

Note that the variance is also expressed (again with $\mu = E[X]$) as
\[ V[X] = E[X^2] - \mu^2, \]

indeed $V[X] = E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2$, which equals $E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2$. 


For $X$ a random variable and $c \in \mathbb{R}$ a constant, it is easy to show (book Theo 6.7 p.259) that

$$V[cX] = c^2 V[X],$$

so that if $c > 0$ we have $D[cX] = c D[X].$

Two variables $X, Y$ are said to be independent if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for any pair of values $x, y$. For $X, Y$ a pair of independent random variables, we have the following two important properties:

- Regarding the expected value we have
  $$E[XY] = E[X] \cdot E[Y].$$

- Regarding the variance we have

To prove the identity for the expected value, let $Z = XY$, then we have

$$E[Z] = \sum_z P(Z = z)z = \sum_z \left( \sum_{x,y:xy=z} P(X = x)P(Y = y) \right)z$$

$$= \sum_z \sum_{x,y:xy=z} P(X = x)P(Y = y)xy$$

$$= \sum_x P(X = x)P(Y = y)xy$$

$$= \sum_x P(X = x) \sum_y P(Y = y)y = E[X] \cdot E[Y].$$

Then the variance formula easily follows (book Theo 6.8 p.259). More generally, $n$ random variables $Z_1, \ldots, Z_n$ are said to be independent if for any values $z_1, \ldots, z_n$ we have $P(Z_1 = z_1, \ldots, Z_n = z_n) = \prod_{i=1}^n P(Z_i = z_i)$. We then have

$$V[Z_1 + \cdots + Z_n] = V[Z_1] + \cdots + V[Z_n],$$

(actually it is enough to have $E[Z_iZ_j] = E[Z_i]E[Z_j]$ for any pair $i \neq j$).

Let us compute the variance of the Bernoulli, Binomial, and geometric random variables. For $Z$ the Bern$(p)$ random variable we have $E[Z^2] = p \cdot 1^2 + (1-p) \cdot 0^2 = p$, so that


For $X$ the Bin$(n, p)$ random variable, recall that $X = Z_1 + \cdots + Z_n$ where the $Z_i$ are independent Bern$(p)$ random variables, hence

$$V[X] = n \cdot V[Z] = n \cdot p(1-p),$$

and $D[X] = \sqrt{n} \cdot \sqrt{p(1-p)}$, indicating that the typical distance of $X$ from its expected value $np$ is of the order of $\sqrt{n}$ (we will see this more precisely when covering the central limit theorem).

For $X$ the Geom$(p)$ random variable we can calculate $E[X^2]$ similarly as we did for $E[X]$, by conditioning on the outcome of the first coin toss. We find


Now, $E[(1 + \text{Geom}(p))^2] = E[(1 + X)^2] = E[X^2] + 2E[X] + 1 = E[X^2] + \frac{2}{p} + 1$, hence

$$E[X^2] = p + (1-p) \left( E[X^2] + \frac{2}{p} + 1 \right),$$
which easily implies that $E[X^2] = (2 - p)/p^2$. Hence


Let us now state Chebyshev’s inequality (book Theo. 8.1 p.305), which bounds (in terms of the variance) the probability that a random variable is away by at least $\epsilon$ from its expected value. Precisely for $X$ a random variable, of expected value $\mu = E[X]$, and for any $\epsilon > 0$, we have

$$P(|X - \mu| \geq \epsilon) \leq \frac{V[X]}{\epsilon^2}.$$  

The first step of the proof is to observe that, for any random variable $Y$ with $Y \geq 0$ and for any constant $c > 0$ we have

$$P(Y \geq c) \leq \frac{E[Y]}{c}$$  

(because $E[Y] = \sum_{y \geq 0} yp(y) \geq \sum_{y \geq c} yp(y) \geq c \sum_{y} p(y) = cP(Y \geq c)$). The inequality follows easily, since we have

$$P(|X - \mu| \geq \epsilon) = P((X - \mu)^2 \geq \epsilon^2) \leq \frac{E[(X - \mu)^2]}{\epsilon^2} = \frac{V[X]}{\epsilon^2}.$$  

Now can apply this inequality to the following situation: we toss a biased coin $n$ times (probability $p$ of ‘heads’, $1 - p$ of ‘tails’), and we define $X_n$ (the observed frequency of H’s) as

$$X_n = \frac{\# H's}{n}.$$  

Note that $X_n$ is distributed as $\frac{Bin(n,p)}{n}$, so that

$$E[X_n] = \frac{E[Bin(n,p)]}{n} = \frac{np}{n} = p,$$  

and

$$V[X_n] = \frac{V[Bin(n,p)]}{n^2} = \frac{np(1 - p)}{n^2} = \frac{p(1 - p)}{n},$$  

and we note that the variance converges to 0 as $n \to \infty$. Using Chebyshev’s inequality we have for any $\epsilon > 0$

$$P(|X_n - p| \geq \epsilon) \leq \frac{p(1 - p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2},$$  

where for the second inequality we have used the fact that the maximum of the function $x \to x(1 - x)$ is $1/4$ (at $x = 1/2$). Hence, for any fixed $\epsilon > 0$ we have

$$\lim_{n \to \infty} P(|X_n - p| \geq \epsilon) = 0.$$  

This is the weak law of large numbers (for coin tosses), see also the book (Theo.8.2 p.307).

We can now apply it to the situation of a poll. Say we have two candidates A and B, let $p \in (0, 1)$ be the fraction of the population voting for A (so that there is a fraction $1 - p$ voting for B). Then $p$ is unknown to us and we are conducting a poll to try to estimate $p$ with a good precision. Say we ask 5000 people for which candidate they vote, and compute the observed frequency as $X = \frac{\# say vote A}{5000}$. Then Chebyshev’s inequality ensures that

$$P(|X - p| \geq \epsilon) \leq \frac{1}{4 \cdot 5000 \cdot \epsilon^2}. $$
in particular, if we use it with \( \epsilon = 0.03 \), we have 
\[
P(|X - p| \geq 0.03) \leq \frac{1}{20000 \cdot 0.03^2} = \frac{1}{18} \approx 0.055,
\]
so that by asking 5000 people and computing the observed frequency \( X \) of A-voters, we have at least 94.5% chances that our estimate \( X \) is not away from the true frequency \( p \) by more than 0.03 (for instance if \( p = 0.6 \), we have at least 94.5% chances that our estimate is in \((0.57, 0.63))\). It actually turns out that Chebyshev’s inequality is very crude, and a much better bound (whose proof relies on generating functions of random variables, to be seen later, Chap.10) is Hoeffding’s bound:
\[
P(|X - p| \geq \epsilon) \leq 2e^{-2\epsilon^2}.
\]
For \( n = 5000 \) and \( \epsilon = 0.03 \), Hoeffding’s bound gives \( P(|X - p| \geq 0.03) \leq 2.5 \cdot 10^{-4} \). For instance for \( p = 0.6 \) we find (from the explicit expression of the binomial distribution that \( P(|X - p| \geq 0.03) \approx 1.6 \cdot 10^{-5} \), so that Hoeffding’s inequality is much tighter (even if differing from the exact value by a factor of order 10) than Chebyshev’s inequality.

**Week 6.**

Another important random variable is the Poisson law (book p.187). For every \( \lambda > 0 \) (real positive value), the Pois(\( \lambda \)) random variable is the random variable \( X \) with probability mass function
\[
P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \text{ for every } k \geq 0.
\]
The interest of the Poisson random variable is that it approximates well a binomial distribution \( \text{Bin}(n, p = \lambda/n) \) when \( n \) is large (and thus \( p \) is small). Indeed, for each fixed \( k \geq 0 \) we have
\[
\lim_{n \to \infty} \binom{n}{k} \frac{(\lambda/n)^k}{k!} (1 - \lambda/n)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!}.
\]
To see this, recall that the binomial coefficient \( \binom{n}{k} \) is equal to \( n! / k!(n-k)! \) with the notation \( (n)_k = n(n-1) \cdots (n-k+1) \). Hence the left side (from which we take the limit) is of the form
\[
\frac{(n)_k \lambda^k}{n^k} \frac{1}{k!} (1 - \lambda/n)^{n-k}.
\]
We have \( \frac{(n)_k}{n^k} = \frac{1}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} = 1 \cdot (1 - \frac{1}{n}) \cdots (1 - \frac{k-1}{n}) \) and clearly each of the \( k \) factors in this product converges to 1 as \( n \to \infty \) (remember that \( k \) is considered as a fixed value in this limit calculation). Hence \( \frac{(n)_k}{n^k} \) converges to 1 as \( n \to \infty \). Now we consider the factor \( (1 - \lambda/n)^{n-k} \), which is also \( (1 - \lambda/n)^n / (1 - \lambda/n)^k \). The factor \( (1 - \lambda/n)^k \) converges to 1 (because \( \lambda/n \) converges to 0 and because \( k \) is fixed), and by taking logarithm we have
\[
\log \left( (1 - \lambda/n)^n \right) = n \log \left( 1 - \lambda/n \right) \sim n \cdot (-\lambda/n) = -\lambda,
\]
where we use \( \log(1 + x) \sim x \) when \( x \to 0 \). Hence \( \log \left( (1 - \lambda/n)^n \right) \) converges to \(-\lambda\), and thus \( (1 - \lambda/n)^n \) converges to \( e^{-\lambda} \). Overall we conclude that \( \binom{n}{k} \frac{(\lambda/n)^k}{k!} (1 - \lambda/n)^{n-k} \) indeed converges to \( e^{-\lambda} \frac{\lambda^k}{k!} \), which is the distribution of Pois(\( \lambda \)).
Typically we can apply the Poisson approximation for random variables that count the number of occurrences of some event over a certain period of time, where the given event occurs at a certain rate by time unit. For instance, consider a shop where every minute a new customer enters with probability 1%. Then the number of customers that enter in a given hour is a Bin(60, 0.01) random variable (since there are 60 minutes in one hour), which is well approximated by Pois(0.6). And thus the probability that no customer enters in a given hour is $\approx e^{-0.6} \approx 0.549$, to be compared with the exact value $(1 - 0.01)^{60} \approx 0.547$ (see also book p.190 for a comparison table).

Regarding the expected value of the Pois($\lambda$) random variable $X$ we have

$$E[X] = \sum_{k \geq 0} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k \geq 1} \frac{\lambda^k}{(k-1)!} = e^{-\lambda} \lambda \sum_{k \geq 1} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda} \lambda e^\lambda = \lambda,$$

which is consistent with $E[\text{Bin}(n, \lambda/n)] = n \ast (\lambda/n) = \lambda$ and the fact that Pois($\lambda$) approximates Bin($n, \lambda/n$) for $n$ large. Similarly, in view of computing the variance we have

$$E[X(X-1)] = \sum_{k \geq 0} k(k-1) e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k \geq 2} \frac{\lambda^k}{(k-2)!} = e^{-\lambda} \lambda^2 \sum_{k \geq 2} \frac{\lambda^{k-2}}{(k-2)!} = e^{-\lambda} \lambda^2 e^\lambda = \lambda^2,$$

and then $E[X^2] = E[X(X-1) + X] = E[X(X-1)] + E[X] = \lambda^2 + \lambda$, from which we conclude that

$$V[X] = E[X^2] - E[X]^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda.$$

Again this is consistent with the fact that $V[\text{Bin}(n, \lambda/n)] = np(1-p) \approx np = \lambda$ when $n$ is large (and thus $p$ is small).

The next table summarizes the important distributions we have seen:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$P(X = k)$</th>
<th>$E[X]$</th>
<th>$V[X]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin($n, p$)</td>
<td>$\binom{n}{k} p^k (1-p)^{n-k}$</td>
<td>$np$</td>
<td>$np(1-p)$</td>
</tr>
<tr>
<td>Geom($p$)</td>
<td>$p(1-p)^{k-1}$</td>
<td>$1/p$</td>
<td>$1 - p/p^2$</td>
</tr>
<tr>
<td>Poisson($\lambda$)</td>
<td>$e^{-\lambda} \frac{\lambda^k}{k!}$</td>
<td>$\lambda$</td>
<td>$\lambda$</td>
</tr>
</tbody>
</table>

We now discuss the so-called gambler’s ruin problem (book Sec.12.2 p.486). Say that there are two players A and B. A has initially $i$$8$ and B has initially $j$$8$. Let $M = i + j$ be the total wealth of the two players. At each round a coin is tossed (with $P(H) = p$ and $P(T) = 1-p$). If the coin gives $H$, then A receives $1$ from B, and if the coin gives $T$ then A gives $1$ to B. The game stops when either A is ruined or B is ruined. In the first case B ends with all the $M$ (B wins) in the second case A ends with all the $M$ (A wins). The following chart shows a scenario (with initial wealths $i = 3$ and $j = 2$) leading to A winning.
For $i \in \{0, \ldots, M\}$ let $P_i$ be the probability that A wins after starting with $i$ $\$ (and B starting with $(M - i)$ $\$); and similarly let $Q_i$ be the probability that A gets ruined after starting with $i$ $(\text{and B starting with } (M - i) \$).

First easy observations are that $P_0 = 0$ (situation where A is ruined from the beginning), $P_M = 1$ (situation where B is ruined from the beginning), and also $P_i \geq P_{i-1}$ for $i \in \{1, \ldots, M\}$ (intuitively the wealthier initially the more likely to win; more precisely for the same scenario of coin tosses, if A wins with initial wealth $i-1$, then A also wins with initial wealth $i$). Define $\ell_i = P_i - P_{i-1}$ for $i \in \{1, \ldots, M\}$. In view of computing the $P_i$ we first obtain linear relations between them, using conditional probability. Let $H$ be the event that the first toss is 'Heads' and $T$ be the event that the first toss is 'Tails'. Note that, for $0 < i < M$ we have

$$P_i = P(A \text{ wins}|H) \cdot p + P(A \text{ wins}|T) \cdot (1 - p) = P_{i+1} \cdot p + P_{i-1} \cdot (1 - p).$$

For $p = 1/2$ (fair coin) this gives $P_i = (P_{i+1} + P_{i-1})/2$, so that $2P_i = P_{i+1} + P_{i-1}$ and thus $P_i - P_{i-1} = P_{i+1} - P_i$, that is, $\ell_i = \ell_{i+1}$. Hence $\ell_1 = \ell_2 = \ldots = \ell_M$, and since $\ell_1 + \ldots + \ell_M = P_M - P_0 = 1$, we have $\ell_i = 1/M$ for all $i \in \{1, \ldots, M\}$. Since $P_1 = \ell_1 + \cdots + \ell_i$, we conclude that for $p = 1/2$ (fair coin) we have

$$P_i = \frac{i}{M} \text{ for } i \in \{0, \ldots, M\}.$$

Similarly we can solve for $Q_i$ (the linear relations are the same, that is $Q_i = Q_{i+1} \cdot p + Q_{i-1} \cdot (1 - p)$, but the boundary conditions are inversed), we now have $Q_0 = 1$ and $Q_M = 0$, giving for $p = 1/2$:

$$Q_i = \frac{M - i}{M} = 1 - P_i \text{ for } i \in \{0, \ldots, M\}.$$

Note that $P_i + Q_i = 1$ implies that the game ends with probability 1 (so the situation where the game never stops has probability 0).

Let us now solve the biased case, that is, $p \neq 1/2$. The relation $P_i = pP_{i+1} + (1 - p)P_{i-1}$, easily yields (after rewriting $P_i$ as $pP_{i+1} + (1 - p)P_i$) the relation $(1 - p)(P_i - P_{i-1}) = p(P_{i+1} - P_i)$. Hence, we have

$$\ell_{i+1} = \alpha \cdot \ell_i, \text{ where } \alpha := \frac{1 - p}{p}.$$

This implies $\ell_i = \alpha^{i-1}\ell_1$ for $i \in \{1, \ldots, M\}$, and thus

$$P_i = \ell_1 + \cdots + \ell_i = \ell_1(1 + \alpha + \cdots + \alpha^{i-1}) = \ell_1 \frac{\alpha^i - 1}{\alpha - 1}.$$
where in the last equality we have to remember that $p \neq 1/2$ (and thus $\alpha \neq 1$).
Since $P_M = 1$ we have $\ell_i a^{M-1} = 1$ and thus $\ell_i = (\alpha - 1)/(\alpha^M - 1)$, from which we conclude that
\[
P_i = \frac{\alpha^i - 1}{\alpha^M - 1},
\]
and by similar techniques we find
\[
Q_i = \frac{\alpha^{M-i} - 1}{\alpha^M - 1} = \frac{\alpha^M - \alpha^i}{\alpha^M - 1} = 1 - P_i.
\]
Hence in the biased case we also have $P_i + Q_i = 1$ (and thus the probability that the game never stops is 0).

Note that for $i = M/2$ (say $M$ is even and the initial fortune is equally distributed to the two players) and $M$ large, we have, if $\alpha > 1$ (i.e., $p < 1/2$),
\[
P_{M/2} = \frac{\alpha^{M/2} - 1}{\alpha^M - 1} \approx \frac{\alpha^{M/2}}{\alpha^M} = \alpha^{-M/2},
\]
so that the chance of $A$ to win is exponentially small. Note that the expected gain $\delta$ of $A$ in any given round is $\delta := p \cdot 1 + (1-p) \cdot (-1) = 2p - 1$, which is negative when $p < 1/2$. By the law of large numbers, the wealth of $A$ after $N >> 1$ rounds will be approximately $M/2 + \delta \cdot N$, so that typically $A$ will be ruined after $N \approx -M/(2\delta)$ rounds (since this value of $N$ is solution of $M/2 + \delta \cdot N = 0$).

Similarly, if $p > 1/2$, the expected gain $\delta = 2p - 1$ is positive, and $A$ typically wins after approximately $M/(2\delta)$ rounds.

**Week 7.**

Recall that the sample space $\Omega$ is the set of all possible outcomes of the experiment we consider. Up to now $\Omega$ was always finite or infinite countable (e.g. $\Omega = \mathbb{N}$). From now on we start to look at *continuous* sample spaces (book Sec.2.1), where $\Omega$ is typically a subset of $\mathbb{R}$, or more generally a subset of $\mathbb{R}^d$ for some $d \geq 1$ (e.g. for $d = 2$ we look at a subset of the plane). In the discrete setting the probability of an event was $P(E) = \sum_{\omega \in E} m(\omega)$, with $m(\omega)$ the probability distribution on the sample space (having the property that $\sum_{\omega \in \Omega} m(\omega) = 1$).

In the continuous setting the distribution function is replaced by a so-called *probability density* $h(\omega) \geq 0$, and the probability of an event $E \subset \Omega$ is given by
\[
P(E) = \int_{\omega \in E} h(\omega) \, d\omega,
\]
and $h(\omega)$ has to satisfy $\int_{\omega \in \Omega} h(\omega) \, d\omega = 1$ (to satisfy the axiom $P(\Omega) = 1$).

For instance, if $\Omega$ is the interval $[0, 1]$ we can consider the uniform probability density on $[0, 1]$, given by $h(\omega) = 1$ for all $\omega \in [0, 1]$ (note that $\int_{\omega \in \Omega} h(\omega) \, d\omega = \int_0^1 1 \cdot d\omega = 1$, as required). And more generally, if $\Omega$ is an interval $I = [a, b]$, the uniform probability density on $I$ is given by $h(\omega) = 1/(b - a) = 1/\text{length}(I)$ for $\omega \in [a, b]$ (again we have $\int_{\omega \in I} h(\omega) \, d\omega = \int_I \frac{1}{\text{length}(I)} \, d\omega = 1$, as required). And then if $J = [c, d]$ is a subinterval of $I = [a, b]$, we have
\[
P(J) = \int_c^d \frac{1}{b - a} \, d\omega = \frac{d-c}{b-a} = \frac{\text{length}(J)}{\text{length}(I)}.
\]
Similarly, if \( \Omega \) is a subset of \( \mathbb{R}^2 \), then the uniform probability density on \( \Omega \) is
\[
h(\omega) = \frac{1}{\text{Area}(\Omega)},
\]
and for each subset \( E \subset \Omega \) we have
\[
P(E) = \frac{\text{Area}(E)}{\text{Area}(\Omega)}.
\]
For instance, in the game of darts (book p.56), the sample space \( \Omega \) is the unit disk, of area \( \pi \), and we assume that the dart lands at a position that is uniformly random on \( \Omega \), thus for \( a \in [0, 1] \) the probability that the darts lands at distance at most \( a \) from the centre of the disk is
\[
P(\text{distance at most } a) = \frac{\text{Area}(\text{disk radius } a)}{\text{Area}(\text{disk radius } 1)} = \frac{\pi a^2}{\pi} = a^2.
\]
Another (more involved) example is Buffon’s needle, where a needle of unit length falls randomly on a floor made of wooden strips of width 1. We then look at the probability of the event that the needle crosses a lines separating two strips. Assuming the distance \( d \) from the middle of the needle to the closest separating line is uniformly random in \( [0, 1/2] \) and the angle \( \theta \) between the needle-direction and the strip direction is uniformly random in \( [0, \pi/2] \), we find (see book p.44-45 for details) that the needle crosses the closest line if and only if
\[
d < \frac{1}{2} \sin(\theta),
\]
and therefore if we denote by \( E \) the set of pairs \((d, \theta)\) such that \( 0 \leq d \leq 1/2, 0 \leq \theta \leq \pi/2, \text{ and } d < \frac{1}{2} \sin(\theta) \), then
\[
P(\text{crosses}) = \int_{E} \frac{dd\,d\theta}{2\pi/2} = \frac{4}{\pi} \int_{E} dd\,d\theta = \frac{4}{\pi} \int_{0}^{\pi/2} \int_{0}^{\sin(\theta)/2} dd\,d\theta = \frac{4}{\pi} \int_{0}^{\pi/2} \frac{1}{2} \sin(\theta)d\theta = \frac{2}{\pi}.
\]
We can now look at the concept of random variable in the continuous setting (book Sec.2.2 and Sec.5.2). Recall that a random variable is a function from the sample space \( \Omega \) to \( \mathbb{R} \). In the discrete setting, the range of values that \( X \) can take is either finite or infinite countable (e.g. for Poisson laws the range is \( \mathbb{N} \)), and we have the probability mass function \( P(x) \) defined as \( p(x) = P(X = x) \) for any \( x \) in the range (with the condition \( \sum_{x} p(x) = 1 \)), and then for any set \( S \) of values we have
\[
P(X \in S) = \sum_{x \in S} p(x).
\]
In the continuous setting, we say that \( X \) is a continuous random variable if there exists a function \( f(x) \) from \( \mathbb{R} \) to \( \mathbb{R}^+ \) such that for any subset \( S \subset \mathbb{R} \),
\[
P(X \in S) = \int_{x \in S} f(x)dx,
\]
where we require that \( \int_{\mathbb{R}} f(x)dx = 1 \) (indeed, since \( X \) is real-valued we must have \( P(X \in \mathbb{R}) = 1 \)). Beware that \( f(a) \) is not \( P(X = a) \) (in the continuous setting, \( P(X = a) = 0 \) for all values \( a \)): for \( \epsilon > 0 \) small, we have
\[
P\left(X \in \left[a - \frac{\epsilon}{2}, a + \frac{\epsilon}{2}\right]\right) = \int_{a-\epsilon/2}^{a+\epsilon/2} f(x)dx \approx \epsilon \cdot f(a),
\]
so that \( f(a) \) indicates how likely \( X \) is to be close to \( a \) (also, \( f(a) \) can be larger than 1, another difference with the discrete setting). The function \( f(x) \) is called the probability density function (p.d.f.) of \( X \).
Another important function related to a random variable \( X \) is the cumulative function \( F(a) \), defined by

\[
F(a) := P(X \leq a) = \int_{-\infty}^{a} f(x) dx.
\]

Note that we have for each interval \([a, b]\):

\[
P(X \in [a, b]) = P(X \leq b) - P(X < a) = P(X \leq b) - P(X \leq a) = F(b) - F(a).
\]

By the fundamental theorem of calculus we also have the following very useful relation (book Theo.2.1 p.61) which allows us to compute \( f(a) \) from \( F(a) \) (we have to be careful that the relation holds only when \( f \) is continuous at \( a \)):

\[
(3) \quad F'(a) = f(a).
\]

As a first example, let \( X \) be the random variable whose p.d.f is \( f(x) = e^{-x} \) for \( x \geq 0 \) and \( f(x) = 0 \) for \( x < 0 \) (note that \( \int_{-\infty}^{\infty} f(x) dx = \int_{0}^{\infty} e^{-x} dx = 1 \), as required). Then the cumulative function \( F \) of \( X \) satisfies \( F(a) = 0 \) for \( a < 0 \) (since the support of \( X \) is \( \mathbb{R}_{+} \)) and for \( a \geq 0 \) we have \( F(a) = \int_{0}^{a} e^{-x} dx = -e^{-x}|_{0}^{a} = 1 - e^{-a} \). In this first example, note that no sample space is mentioned, which will often be the case in the distributed setting (random variables will often be directly given by the p.d.f., and the sample space is somehow hidden in the background).

However we may still consider random variables related to an explicit sample space. For instance, in the dart example (where the sample space \( \Omega \) is the unit disk and the density distribution on the unit disk is uniform), consider the random variable \( X \) that gives the distance of the dart from the centre. As we have seen above we have for \( a \in [0, 1] \):

\[
P(X \leq a) = \frac{\text{Area(disk radius } a)}{\text{Area(disk radius 1)}} = \frac{\pi a^2}{\pi} = a^2,
\]

hence the cumulative function \( F(a) \) of \( X \) satisfies \( F(a) = 0 \) for \( a < 0 \), \( F(a) = a^2 \) for \( a \in [0, 1] \), and \( F(a) = 1 \) for \( a > 1 \). Using (3) we deduce that the p.d.f. of \( X \) is \( f(a) = 0 \) for \( a \notin [0, 1] \) and \( f(a) = 2a \) for \( a \in [0, 1] \).

We now look at how the expected value is defined in the continuous setting (book Sec.6.3). For \( X \) a continuous random variable with p.d.f. \( f(x) \), the expected value of \( X \) is defined as

\[
E[X] = \int_{-\infty}^{+\infty} xf(x) dx.
\]

For instance, in the dart example above we find that the expected value is equal to \( \int_{0}^{1} x \cdot 2x dx = \int_{0}^{1} 2x^2 dx = \frac{2}{3} x^3|_{0}^{1} = \frac{2}{3} \), i.e., the average distance of the dart from the centre is \( 2/3 \).