Data Mining Homework 1

Alessio Maiola

1 Exercise 1

1.1 Define probability space

The sample space Ω for the random process that uniformly obtains a random permutation of a 52 cards deck is composed of all the possible 52! permutations of the deck.

$$\Omega = \{p_1, p_2, \ldots, p_{52!}\}$$

Where each permutation is a different ordered list of all the cards in the deck. For example:

$$p_1 = [\mathrm{H1}, \mathrm{H2}, \ldots, \mathrm{H10}, \mathrm{HJ}, \mathrm{HQ}, \mathrm{HK}, \mathrm{D1}, \ldots, \mathrm{DK}, \mathrm{C1}, \ldots, \mathrm{CK}, \mathrm{S1}, \ldots, \mathrm{SK}]$$

where H, D, C, S stand for the suits (Hearts, Diamonds, Clubs, Spades), and the values range from 1 to 10, with the additional face cards J, Q, and K representing Jack, Queen, and King, respectively.

Being every permutation chosen uniformly at random,

$$Pr(p_i) = \frac{1}{|\Omega|} = \frac{1}{52!}, i \in \{1, 2, \ldots, 52!\}$$

Furthermore, then event space $\mathcal{F}$ is constituted by all the possible subsets of Ω. Therefore it is clear that $\mathcal{F} = 2^{\Omega} = \{E | E \subseteq \Omega\}$. We can extend the probability function to the events of $\mathcal{F}$, by simply summing the probabilities of the simple events (the permutations) that are included in them.

$$\forall E \in \mathcal{F} \text{ we have that } Pr(E) = Pr(\bigcup_{\omega \in E} \omega) = \sum_{\omega \in E} Pr(\omega) = |E| Pr(\omega) = \frac{|E|}{|\Omega|}$$

This also defines the probability function $Pr : \mathcal{F} \rightarrow \mathbb{R}$.

1.2 Find the probabilities of the events

(a) E = "The first four cards include at least one club".
This exercise is easy to solve using combinatorics. First, let us define the complementary event $\overline{E}$ = "First four cards include no club". Being the complementary, it is clear that $Pr(E) = 1 - Pr(\overline{E})$. Now, to compute $Pr(\overline{E})$, we need

to find the proportion of ways to choose 4 cards from the 39 cards that are not clubs compared to the total number of ways to choose 4 cards from the entire deck of 52 cards.

$$Pr(E) = 1 - Pr(\overline{E}) = 1 - \frac{\binom{39}{4}}{\binom{52}{4}} \simeq 1 - 0.304 = 0.696$$

(b) E="The first seven cards include exactly one club".
This time we compute the combinations of 6 cards among the 39 that are not clubs and also consider the combinations of exactly one card among the 13 club cards. Of course, the total possible combinations are the ways to choose 7 cards from the deck of 52.

$$Pr(E) = \frac{\binom{39}{6}\binom{13}{1}}{\binom{52}{7}} \simeq 0.317$$

(c) E="The first three cards are all of the same suit".
The number of ways to choose three cards with a specific suit is the combinations of 3 cards among the 13 cards of each suit. Moreover, we multiply the number of combinations by the number of ways to choose 1 suit from the 4 available suits. The total possible combinations are computed in the same way as before.

$$Pr(E) = \frac{\binom{4}{1}\binom{13}{3}}{\binom{52}{3}} \simeq 0.0518$$

(d) E="The first three cards are all sevens".
We need to calculate the number of ways to choose 3 sevens from the 4 available sevens in the deck. The total possible combinations are computed in the same way as before.

$$Pr(E) = \frac{\binom{4}{3}}{\binom{52}{3}} \simeq 1.81 * 10^{-4}$$

(e) E = "The first five cards form a straight".
The possible ways to make a straight with 5 cards are 10:

$$\{1,2,3,4,5\}, \{2,3,4,5,6\}, \{3,4,5,6,7\}, \{4,5,6,7,8\}, \{5,6,7,8,9\},$$
$$\{6,7,8,9,10\}, \{7,8,9,10,J\}, \{8,9,10,J,Q\}, \{9,10,J,Q,K\}, \{10,J,Q,K,1\}$$

Each card in a straight can be chosen from all 4 suits, leading to 4^5 possibilities for each straight. However, we must exclude the cases where all the cards are of the same suit, as these would be straight flushes. There are 4 such possibilities for each straight (one for each suit). The total possible combinations are computed in the same way as before.

$$Pr(E) = \frac{10 * (4^5 - 4)}{\binom{52}{5}} \simeq 0.00392$$

2 Exercise 2

2.1 Define the sample space

The sample space is defined by $(2 * 7)^2 = 196$ events. This is because each kid can be born in any of the seven days of the week and he/she can be either a boy or a girl. Therefore, for each kid we have the following possible outcomes, obtained combining the sets of sexes (S) and the set of days (D):

$$S = \{b, g\}, D = \{1, 2, 3, 4, 5, 6, 7\}$$
$$O = S \times D = \{\langle b, 1\rangle, \langle b, 2\rangle, \ldots \langle b, 7\rangle, \langle g, 1\rangle, \langle g, 2\rangle, \ldots, \langle g, 7\rangle\}$$
$$|O| = |S| * |D| = 2 * 7 = 14$$

Since the kids are two, the sample space for the problem is defined by O^2, which is the cartesian product of the possible outcomes for each one of the two kids (the i-th element of the tuple refers to the i-th kid). To improve readibility and distinguish the cartesian products between S and D and O with itself, both the symbols $\langle\rangle$ and $()$ have been used for the tuples of the cartesian product.

$$\Omega = O^2 = O \times O = \{(\langle b, 1\rangle, \langle b, 1\rangle), (\langle b, 1\rangle, \langle g, 1\rangle), \ldots, (\langle b, 7\rangle, \langle g, 7\rangle),$$
$$(\langle g, 1\rangle, \langle b, 1\rangle), (\langle g, 1\rangle, \langle g, 1\rangle), \ldots, (\langle g, 7\rangle, \langle g, 7\rangle)\}$$
$$|\Omega| = |O|^2 = 14^2 = 196$$
$$\forall \omega \in \Omega, \text{ we have that } Pr(\omega) = \frac{1}{|\Omega|} = \frac{1}{196}$$

2.2 Probability that the other kid is a girl, knowing the sex of one

To solve this problem, we need a more compact notation to describe events. We define the events:

$$\langle s, *\rangle = \bigcup_{i=1}^{7} \langle s, i\rangle, s \in \{b, g\}$$
$$\langle *, d\rangle = \bigcup_{i \in \{g,b\}} \langle i, d\rangle, d \in \{1, 2, \ldots, 7\}$$

With this shorter notation, we can describe the events we are interested in as follows:

$$E = \text{"both kids are girls"} = (\langle g, *\rangle, \langle g, *\rangle) = (\langle g, *\rangle, \langle *, *\rangle) \cap (\langle *, *\rangle, \langle g, *\rangle)$$
$$F = \text{"one of the kids is a girl"} = (\langle g, *\rangle, \langle *, *\rangle) \cup (\langle *, *\rangle, \langle g, *\rangle)$$

Now we can compute the probability of these events, exploiting the independence of the events related to the two kids. Moreover, we note that $E \subset F$, hence $Pr(E \cap F) = Pr(E)$.

$$Pr(E) = Pr((\langle g,*\rangle, \langle g,*\rangle)) = Pr((\langle g,*\rangle, \langle *,*\rangle))Pr((\langle *,*\rangle, \langle g,*\rangle)) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$$

$$Pr(F) = Pr((\langle g,*\rangle, \langle *,*\rangle) \cup (\langle *,*\rangle, \langle g,*\rangle)) =$$

$$Pr((\langle g,*\rangle, \langle *,*\rangle)) + Pr((\langle *,*\rangle, \langle g,*\rangle)) - Pr((\langle g,*\rangle, \langle g,*\rangle)) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$$

$$Pr(E|F) = Pr((\langle g,*\rangle, \langle g,*\rangle)|(\langle g,*\rangle, \langle *,*\rangle) \cup (\langle *,*\rangle, \langle g,*\rangle)) = \frac{Pr(E \cap F)}{Pr(F)} = \frac{Pr(E)}{Pr(F)} = \frac{1}{3}$$

2.3 Probability that the other kid is a girl, knowing the sex and birthday of one

Again, we are interested in the event defined by E. However, this time we not only know the sex, but also the birthday of one of the kids.

$G =$ "one of the kids is a girl born on Sunday" $= (\langle g,7\rangle, \langle *,*\rangle) \cup (\langle *,*\rangle, \langle g,7\rangle)$

$$Pr(G) = Pr((\langle g,7\rangle, \langle *,*\rangle) \cup (\langle *,*\rangle, \langle g,7\rangle)) =$$

$$Pr((\langle g,7\rangle, \langle *,*\rangle)) + Pr((\langle *,*\rangle, \langle g,7\rangle)) - Pr((\langle g,7\rangle, \langle g,7\rangle)) = \frac{1}{14} + \frac{1}{14} - \frac{1}{196} = \frac{27}{196}$$

This time, for the conditional probability we need to calculate the probability of the intersection of the events E and G.

$$Pr(E \cap G) = Pr((\langle g,*\rangle, \langle g,*\rangle) \cap [(\langle g,7\rangle, \langle *,*\rangle) \cup (\langle *,*\rangle, \langle g,7\rangle)]) = Pr((\langle g,7\rangle, \langle g,*\rangle) \cup (\langle g,*\rangle, \langle g,7\rangle)) =$$

$$Pr((\langle g,7\rangle, \langle *,*\rangle))Pr((\langle *,*\rangle, \langle g,*\rangle)) + Pr((\langle *,*\rangle, \langle g,7\rangle))Pr((\langle g,*\rangle, \langle *,*\rangle)) - Pr((\langle g,7\rangle, \langle g,7\rangle)) =$$

$$\frac{1}{14} * \frac{1}{2} + \frac{1}{14} * \frac{1}{2} - \frac{1}{196} = \frac{13}{196}$$

$$Pr(E|G) = Pr((\langle g,*\rangle, \langle g,*\rangle)|(\langle g,7\rangle, \langle *,*\rangle) \cup (\langle *,*\rangle, \langle g,7\rangle)) = \frac{Pr(E \cap G)}{Pr(G)} = \frac{13}{27}$$

3 Exercise 3

We are interested in the following events:

$$S = \text{"The instructor is sick"}$$
$$T = \text{"The test is positive"}$$
$$Pr(S) = \frac{1}{100000}$$
$$Pr(T|S) = \frac{999}{1000}$$

To compute the requested probability, which is $P(S|T)$, we first need to compute $P(T)$.

$$P(T) = P(T \cap S) + P(T \cap \overline{S}) = P(S)P(T|S) + P(\overline{S})P(T|\overline{S}) = \frac{1}{100000} * \frac{999}{1000} + \frac{99999}{100000} * \frac{1}{1000} = \frac{100998}{10^8}$$

$$P(S|T) = \frac{P(S \cap T)}{P(T)} = \frac{P(S)P(T|S)}{P(T)} = \frac{999}{100998} \simeq 0.00989$$

Therefore, the paradox lies in the fact that, despite the high accuracy of the test, the probability of being sick after a positive test result remains low due to the rare incidence of Data-miningitis in the population.

4 Exerckse 4

4.1 Knowing the range

Both Aris and Gianluca choose random pages in the range [1,n], indicated with A and G respectively. Moreover, the pages they choose need to be different, hence $Pr(A = G) = 0$. If we knew a number X between A and G, the probability to identify the person who chose the smallest number would be 1. It is clear, taking into account the only possible outcomes.

$$\Omega = \{A < X < G, G < X < A\}$$

Obviously, the probability of the events is uniform.

$$\forall \omega \in \Omega, \text{ we have that } Pr(\omega) = \frac{1}{|\Omega|} = \frac{1}{2}$$

To elaborate a strategy, calculating the following probabilities is helpful:

$$\begin{aligned} Pr(A < X) &= Pr(X < A) = \frac{1}{2} \\ Pr(A < G|A < X) &= \frac{Pr(A < X < G)}{Pr(A < X)} = 1 \\ Pr(G < A|A > X) &= \frac{Pr(G < X < A)}{Pr(A > X)} = 1 \end{aligned}$$

Suppose that Aris reveals us his number (the person chosen to reveal the number is actually irrelevant for the result). Our strategy would be choosing Aris if $A < X$, Gianluca otherwise. So let us compute the probability of winning the game:

$W =$ "$A < G$ if $A < X, G < A$ otherwise"

$$Pr(W) = Pr(A < G|A < X)Pr(A < X) + Pr(G < A|A > X)Pr(A > X) = \frac{1}{2} + \frac{1}{2} = 1$$

Of course, choosing an appropriate X between A and G requires to know the limits of the range, that are the chosen numbers. For this reason, we are certain to win in this case.

4.2 Random number

Following the same strategy without knowing A and G yields surprising results. Let us just suppose to choose a random number $X \in [1, n]$. For simplicity, we suppose that $X \neq A$ and $X \neq G$. The sample space is constituted by all the possible orderings of these 3 numbers.

$\Omega = \{A < G < X, X < A < G, A < X < G, G < X < A, G < A < X, X < G < A\}$

Again, the probability of the events is uniform.

$$\forall \omega \in \Omega, \text{ we have that } Pr(\omega) = \frac{1}{|\Omega|} = \frac{1}{6}$$

Let us evaluate again our strategy:

$$Pr(A < X) = Pr(X < A) = \frac{1}{2}$$
$$Pr(A < G|A < X) = \frac{Pr(A < X < G) + Pr(A < G < X)}{Pr(A < X)} = \frac{\frac{2}{6}}{\frac{1}{2}} = \frac{2}{3}$$
$$Pr(G < A|A > X) = \frac{Pr(G < X < A) + Pr(X < G < A)}{Pr(A > X)} = \frac{2}{3}$$
$$Pr(W) = Pr(A < G|A < X)Pr(A < X) + Pr(G < A|A > X)Pr(A > X) = \frac{2}{3} * \frac{1}{2} + \frac{2}{3} * \frac{1}{2} = \frac{2}{3}$$

Our strategy allows us to correctly identify the person who chose the smallest page number $\frac{2}{3}$ of the times, despite choosing X randomly!

4.3 No assumption on X

Until now, we maintained some simplifying assumption on the choice of X, which now will removed. Therefore, now it is possible that X equals A or G.

$$Pr(A = X) = Pr(A = G) = \frac{1}{n}$$
$$Pr(A < X) = \frac{n-1}{2n} = 1 - Pr(A \geq X)$$

The probability of the possible orderings is still uniform, removing the equality events.

$$Pr(G \neq X \cap A \neq X) = 1 - \frac{2}{n}$$
$$Pr(A < X < G) = \frac{Pr(G \neq X \cap A \neq X)}{6} = \left(1 - \frac{2}{n}\right) * \frac{1}{6}$$

Moreover, the probabilities of the orderings between A and G can be combined with the equalities, as the events are independent.

$$Pr(A < G = X) = Pr(G = X)Pr(A < G) = \frac{1}{n} * \frac{1}{2} = \frac{1}{2n}$$

Now we can evaluate our strategy again:

$$Pr(A < G|A < X) = \frac{Pr(A < G = X) + Pr(A < G < X) + Pr(A < X < G)}{Pr(A < X)} =$$

$$\frac{\frac{1}{2n} + 2 * (1 - \frac{2}{n})\frac{1}{6}}{\frac{n-1}{2n}} = \frac{2}{3} + \frac{1}{3(n-1)}$$

$$Pr(G < A|A \geq X) = \frac{Pr(G < A = X) + Pr(G < X < A) + Pr(X < G < A) + Pr(X = G < A)}{Pr(A \geq X)} =$$

$$\frac{2 * \frac{1}{2n} + 2 * (1 - \frac{2}{n})\frac{1}{6}}{\frac{n+1}{2n}} = \frac{2}{3}$$

$$Pr(W) = Pr(A < G|A < X)Pr(A < X) + Pr(G < A|A \geq X)Pr(A \geq X) =$$

$$(\frac{2}{3} + \frac{1}{3(n-1)}) * \frac{n-1}{2n} + \frac{2}{3} * \frac{n+1}{2n} = \frac{2}{3} + \frac{1}{6n}$$

Therefore, choosing X completely randomly, the possibilities to pick the person who chose the smallest page number are even higher. However, the additional advantage is very low if the number of pages of the book (n) is huge. Nevertheless, this strategy is obviously better than picking a random person (with probability $\frac{1}{2}$) each time.

5 Exercise 5

5.1 Define the probability space

The sample space for the Erdős–Rényi random graph model $G_{n,p}$ is consituted by all the possible graphs that can be formed with n edges. An undirected graph with n nodes can have a number of edges in the interval $[0, \binom{n}{2}]$. Every edge is present with probability p. Therefore, the sample space will contain all the possible subsets of edges (they all define distinct graphs). Obviously $|\Omega| = 2^{\binom{n}{2}}$.

$\Omega = \{g_1, g_2, \ldots, g_{\binom{n}{2}}\}$

$g_1 = \{\}$ defines the empty graph

$g_{\binom{n}{2}} = \{(i, j)\}$ $\forall i, j \in \{1, 2, \ldots, n\}$ such that $i < j$ defines the graph with all the edges

Also here, then event space $\mathcal{F}$ is constituted by all the possible subsets of Ω. Therefore it is clear that $\mathcal{F} = 2^{\Omega} = \{E|E \subseteq \Omega\}$. Also here the probability function can be extended to the events of $\mathcal{F}$, by summing the probabilities of the simple events (the graphs) that are included in them. However, this time the simple events will not have the same probability: as we will see in the next point, the probability of a specific graph will depend on the number of edges it contains.

$$\forall E \in \mathcal{F} \text{ we have that } Pr(E) = Pr(\bigcup_{\omega \in E} \omega) = \sum_{\omega \in E} Pr(\omega)$$

This also defines the probability function $Pr : \mathcal{F} \rightarrow \mathbb{R}$.

5.2 Probability of each element

It is obvious from the definition that the probability of each graph depends on the number of edges it contains. This is because every edge exists with probability p. If we considered the events in Ω as sets of edges, we could compute the number of edges of each graph by using the cardinality of the event. In other words, $\forall\omega \in \Omega$ we would have that $|\omega| = \#$ of edges in the graph. Moreover, it is clear that another important factor is the number of nodes, that bounds the total number of edges that can be present in the graph (they cannot exceed $\binom{n}{2}$, which are all the possible pairs of 2 distinct nodes).

$$\forall\omega \in \Omega \text{ we have that } Pr(\omega) = p^{|\omega|}(1-p)^{\binom{n}{2}-|\omega|}$$

5.3 Two disjoint cycles with length $\frac{n}{2}$

First, let us notice that all possible configurations for any value of n have the same number of edges, which is n. Therefore:

$E =$ "probability that the graph has two node-disjoint cycles of length $\dfrac{n}{2}$ and no other edge"

$$Pr(E) = \sum_{\omega\in E} Pr(\omega) = \sum_{\omega\in E} p^n(1-p)^{\binom{n}{2}-n} = |E| * p^n(1-p)^{\binom{n}{2}-n}$$

We have to compute the number of distinct graphs that have the described characteristics. Let us consider first the number of combinations of the nodes that generate a graph with these properties. They are a half of the possible ways to choose $\frac{n}{2}$ nodes among all the n. This is because half of the combinations are simply equal. For example, with $n = 6$, choosing to combine the nodes in the sets "abc, def" generates the same graph as splitting them in the opposite way "def, abc".

Moreover, each subgraph of $\frac{n}{2}$ can be permuted in different ways. Some of the permutations, however, are equal. Consider $n = 8$ and the subgraph of 4 edges containing the cycle "abcd". By simply rotating the graph and starting reading it from a different node, we would have the exact same sequence of nodes, shifted a number of times. For example, the sequence would become "bcda" after one shift, "cdab" after two etc. All this $\frac{n}{2}$ permutations are equal in both subgraphs. Additionally, each direction of reading the graph would provide the same cycle reversed. Therefore, "dcba" describes the exact same subgraph as before, hence the number of distinct permutations is halved.

After these considerations, we can conclude that:

$$|E| = \frac{1}{2}\binom{n}{\frac{n}{2}}\left(\frac{(\frac{n}{2})!}{2*\frac{n}{2}}\right)^2 = \frac{1}{8}\binom{n}{\frac{n}{2}}\left(\left(\frac{n}{2}-1\right)!\right)^2$$

$$Pr(E) = \frac{1}{8}\binom{n}{\frac{n}{2}}\left(\left(\frac{n}{2}-1\right)!\right)^2 p^n(1-p)^{\binom{n}{2}-n}$$

5.4 Two disjoint cycles with any length

This version is more general than the previous, allowing also odd values for n and the presence of cycles of different lenghts, with no other edge and no isolated nodes. However, the same reasoning as before can be applied also here. We have to notice, that in the case of cycles with different length, there is no reason to divide by 2 the number of possible combinations. However, we can leave the division in the formula at the cost of considering them twice. Of course, the permutations have now to be considered for the two subgraph separately, based on their lengths.

$F =$ "probability that the graph has two node-disjoint cycles, no isolated node and no other edge"

$$Pr(F) = \sum_{i=3}^{n-3} \frac{1}{8} \binom{n}{i} (i-1)!(n-i-1)! \; p^n (1-p)^{\binom{n}{2}-n}$$

Matter of fact, this formula is so powerful that it can be used to model even the case where there may be isolated nodes! It is somehow more challenging, as the number of edges is not guaranteed anymore to be n. In such cases, we could have two cycles of length i and j and n-i-j isolated nodes.

$G =$ "probability that the graph has two node-disjoint cycles and no other edge"

$$Pr(G) = \sum_{i=3}^{n-3} \sum_{j=3}^{n-i} \frac{1}{8} \binom{n}{i, j, n-i-j} (i-1)!(j-1)! \; p^{(i+j)} (1-p)^{\binom{n}{2}-(i+j)}$$

5.5 Expected degree of a node

This is easy to calculate using the linearity of expectation. Each node can have at most $n-1$ edges towards all the other nodes of the graph and each edge exists with probability p.

$X =$ "degree of a node"

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th edge adjacent to the node is present} \\ 0 & \text{otherwise} \end{cases} \qquad i \in \{1, 2, \ldots, n-1\}$$

$$\mathbb{E}[X_i] = 1 \cdot p + 0 \cdot (1-p) = p$$

$$\mathbb{E}[X] = \sum_{i=1}^{n-1} \mathbb{E}[X_i] = \sum_{i=1}^{n-1} p = (n-1)p$$

5.6 Expected number of edges

The same reasoning as before can be applied to this case as well. Remind that the maximum number of edges of $G_{n,p}$ is $\binom{n}{2}$.

$$X = \text{"number of edges"}$$

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th edge is present} \\ 0 & \text{otherwise} \end{cases} \qquad i \in \{1, 2, \dots, \binom{n}{2}\}$$

$$\mathbb{E}[X_i] = 1 \cdot p + 0 \cdot (1-p) = p$$

$$\mathbb{E}[X] = \sum_{i=1}^{\binom{n}{2}} \mathbb{E}[X_i] = \binom{n}{2} p$$

5.7 Expected number of papillon subgraphs

Also here the same reasoning can be applied. Nevertheless, we also need to compute the number of distinct papillon subgraphs in $G_{n,p}$. Of course we have to consider all the possible ways of choosing 5 nodes for the papillon among the n nodes: $\binom{n}{5}$. Moreover, we should compute the number of permutations that create distinct papillons. It is clear that any permutation that inverts the two nodes on the left wing or of the right wing originates the same papillon. Moreover, all permutations that invert the pair of nodes on the left wing and the pair of nodes of the right wing don't generate a new papillon.

For instance, think of a papillon defined by "ab c de", where the first two letters represent the left wing and the last two the right one. It would be equal to the papillon described by "ba c de". Moreover, it would also be equal to the papillon described by "ab c ed" and lastly to the one described by "de c ab". All possible combinations of these inversion indeed doesn't modify the papillon. Therefore only $\frac{1}{2^3}$ of the permutations generates a different papillon.

Finally, we need to compute the probability that a specific papillon subgraph exists in the graph, which can be done by calculating the probability that in the given 5 nodes exactly 6 edges are present.

$$N = \text{"Number of distinct papillons in } G_{n,p}\text{"} = \binom{n}{5}\frac{5!}{8} = \binom{n}{5} 15$$

$$X = \text{"Number of papillon subgraphs"}$$

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th papillon subgraph is present} \\ 0 & \text{otherwise} \end{cases} \qquad i \in \{1, 2, \dots, N\}$$

$$\mathbb{E}[X_i] = Pr(X_i = 1) = p^6 (1-p)^{\binom{5}{2}-6} = p^6 (1-p)^4$$

$$\mathbb{E}[X] = \sum_{i=1}^{N} \mathbb{E}[X_i] = \binom{n}{5}(15) p^6 (1-p)^4$$

6 Exercise 6

First of all I manually downloaded the file. I accessed its content using the cut instruction, showing the first field of the file using the flag -f1. The default character used for separating the field is the one used in the file (tab). Afterwards, I sorted the file to use the uniq function, which eliminates duplicates and counts them using the -c flag. Finally, a last sorting based on numerical values (the count of the previous uniq instruction) in reverse order (flags -n and -r) returned the beers in descending order by the number of reviews. I isolated the first ten beers with the highest number of reviews using the head instruction with flag -n 10.

```
cut -f1 beers.txt | sort | uniq -c | sort -n -r | head -n 10
```

Listing 1: File analysis

Eventually, I discovered that even downloading the file was not necessary, as it is possible to apply commands to the output of the wget and curl functions without storing it in a file. I used the wget with the -qO- flags, which indicate quiet output, so no logging in the terminal about the download, and no output path respectively. Moreover the –user and –password flags allow to log in and access the resource. Similarly the curl with the -su flag (silent download and username and password for autentication) allows to access a web resource and apply commands directly to the output, without storing it in a file. Elegant.

```
wget -qO- --user datamining2021 --password Data-Min1ng-2021 http://aris.me/contents/teaching/data-mining-2024/protected/beers.txt | cut -f1 | sort | uniq -c | sort -n -r | head -n 10

curl -su datamining2021:Data-Min1ng-2021 http://aris.me/contents/teaching/data-mining-2024/protected/beers.txt | cut -f1 | sort | uniq -c | sort -n -r | head -n 10
```

Listing 2: File download and analysis

7 Exercise 7

After downloading the file, I preprocessed each row using the strip and split methods. The first removes white spaces at the beginning and at the end of the line (included the newline character at the end of the line), the latter separates the beer from the mark using the separating character "\t". I used the name of the beers as keys for a dictionary and I stored for each one the sum of the marks and the count of the reviews. After deleting the file, I filtered out of the dictionary the beers with less than 100 reviews and stored the average overall score in the dictionary replacing the previous tuple. I sorted in decreasing order the beers left in the dictionary based on their values (which is now their average score) and I printed the top 10 beers with at least 100 reviews. Here is my code:

```

import os

file_name = "beers.txt"

os.system(f"curl -o {file_name} -su datamining2021:Data-Min1ng-2021
              ↪  http://aris.me/contents/teaching/data-mining-2024/
              ↪ protected/beers.txt")
beers = {}

with open(file_name) as file:
    for line in file:
        line = line.strip().split("\t")
        beer = line[0]
        mark = int(line[1])
        if beer not in beers:
            beers[beer] = [mark, 1]
        else:
            beers[beer][0] += mark
            beers[beer][1] += 1

os.remove(file_name)

#filter out beers with <100 reviews
for beer in list(beers.keys()):
    if beers[beer][1] < 100:
        beers.pop(beer)
    else:
        total = beers[beer][0]
        count = beers[beer][1]
        beers[beer] = total/count

#sort the beers by their average review mark (descending)
beers_by_avg_score = sorted(beers.keys(), key=lambda beer: beers[
              ↪ beer], reverse=True)

print("Top 10 Beers with at least 100 reviews:")
for beer in beers_by_avg_score[:10]:
    print(f"{beer}: {beers[beer]}")
```

Listing 3: Top 10-scored beers with $\geq$ 100 reviews

8 Other scripts

In addition to the above code, I also wrote some simple simulations for Exercise 1 and Exercise 4, which I included in the delivery. They are so basic and straightforward that I believe they don't require any additional comment.