Lab 1 - Probability Review

Computing the expected value when probabilities are known

For ease of notation let's define the proability of obtaining the i-th face of the die as $p(X = i)$, therefore the event "we throw the die and get the i-th face" is described by $p(X = i)$ (we use i to describe the possible outcomes of the random variable X).
By definition of expected value we have:

$$\mathbf{E}[X] := \mu := \sum_{i=1}^{F} p(X = i) i \tag{1}$$

Note: This holds for any number of faces (F)

Estimating the expected value when probabilities are unknown

Say that we can throw the die N times, call each single realization x_j with $j = 1, 2, \ldots, N$. We can approximate the expected value using the empirical average:

$\hat{\mu} := \frac{1}{N} \sum_{j=1}^{N} x_j$

What are we doing here?
Let's massage the empirical mean a little bit (we are going to exploit the fact that each realization x_j belongs to the set $\in \{1, 2, \ldots, F\}$). In the following we will slightly change notation on the realizations x_j: we will index them using two indexes in place of one. Imagine you are counting the number of occurences of each single face, a possible "smart" way to do this is to parse the entire string of outcomes x_j with $j = 1, 2, \ldots, N$ and insert each single realization x_j into the proper bin (we have F bins labelled from 1 to F) and assign an increasing index whitin each bin (it is different for different bins). Clearly this operation is using two indeces to uniquely index each single realization as we were doing at the beginning using $j = 1, 2, \ldots, N$. Note that a priori we do not know how many outcomes a particular value (face) has and therefore we do not know how far the index of each bin will go (let's call these numbers N_i). We are now indexing the sequence of outcomes using $x_{i,z}$.

$\hat{\mu} = \frac{1}{N} \sum_{j=1}^{N} x_j = \frac{1}{N} \sum_{i=1}^{F} \left[\sum_{z=1}^{N_i} x_{i,z} \right]$

Note 1: Here we only rearranged the summation. We collect the realizations with the same outcome (i) and count how many of them there are (N_i). For example we can throw the die 10 times and see 5 times the face #1, 2 times #2 and 3 times #5.

Exploiting the fact that in each bin (indexed by i) we have the same value of the outcome we can rewrite the empirical average using indicators variables: $x_{i,z} = \mathbb{1}[x_{i,z} = i] * i$

$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{F} \left[\sum_{z=1}^{N_i} \mathbb{1}[x_{i,z} = i] * i \right] = \sum_{i=1}^{F} \left[\frac{i}{N} \sum_{z=1}^{N_i} \mathbb{1}[x_{i,z} = i] \right] = \sum_{i=1}^{F} \frac{N_i}{N} i$

It should be clear now what is happening under the hood: the sample average is approximating the probability of each possible outcome: $\frac{N_i}{N} \approx p(X = i)$

So that we are approximating the true average $\mathbb{E}[X]$ using the definition of expected value replacing the unknown true distribution $p(X = i)$ with its empirical approximation $\frac{N_i}{N}$.

$$\mu = \sum_{i=1}^{F} p(X=i)i \approx \sum_{i=1}^{F} \hat{p}(X=i)i = \sum_{i=1}^{F} \frac{N_i}{N} i = \hat{\mu} \tag{2}$$

Final remark: Under the assumption we have enough outcomes (and that the sampling mechanism is not "degenerate") we can expect the approximation to become increasingly better (this is a straighforward application of LLN).

Hands-on!

Central Limit Theorem

So far we have understood and empirically verified, thanks to the Law of Large Numbers, that the average of samples obtained from the same distribution converges to the expectation of the underlying distribution.

The Central Limit Theorem allows us to say something more about the distribution of the sample average. In particular, it says that the averaging over infinitely many random variables converges to a Gaussian distribution, i.e.:

$\hat{\mu}(x_1, x_2, \ldots, x_m) \simeq \mathcal{N}(\mu, \frac{\sigma^2}{m}) \quad \text{when } m \to \infty$

Sample average randomness

Remember the sample average is a random variable too. Therefore we can study its distribution, first moment, etc.
Consider the following experiment: throw a die N times and repeat the procedure for T times. We can compute the sample average on each batch composed by N throws (this is a random variable) and then we can look at different realization of it since we compute it T times.

For the sake of completeness let's establish some notation, say $X \sim \mathcal{D}$ and let $\mu := \mathbb{E}[X]$ and $\sigma^2 := \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

Consider now the empirical average $\hat{\mu}$, this is a statistics of X, i.e. a function of the outcomes of X. In other words: $\hat{\mu}(x_1, x_2, \ldots, x_N)$, $\hat{\mu}$ is a function of the outcomes $x_1, x_2, \ldots, x_N$. With different realizations $x_1, x_2, \ldots, x_N$ we get different values of $\hat{\mu}$. Therefore we can write $\hat{\mu}$ as a random variable depending on the sequence of random variables $X_1, X_2, \ldots, X_N$, where X_i are i.i.d. random variables with distribution X (i.e. $X_i \sim X$).

It is legitimate to ask ourself what the distribution of $\hat{\mu}(X_1, X_2, \ldots, X_N) := \frac{1}{N}\sum_{i=1}^{N} X_i$ is. The answer to that simple question is not trivial at all and depends on $\mathcal{D}$. A closed form expression for $\hat{\mu}(X_1, X_2, \ldots, X_N)$ might not be easy to compute. Luckily for us $\hat{\mu}$ is a sum of i.i.d. r.v. and therefore we can apply the Central Limit Theorem we have seen in class (many other versions exist!). We can approximate its distribution using a gaussian random variable whose mean a and variance b are given by the following:

$a = \mathbb{E}[\hat{\mu}(X_1, X_2, \ldots, X_N)] \qquad b = Var[\hat{\mu}(X_1, X_2, \ldots, X_N)]$

Fortunately we can easily compute such quantities using basic properties of expected value and variance:

$a = \mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^{N} X_i\Big] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[X_i] = \mu$
$b = Var\Big[\frac{1}{N}\sum_{i=1}^{N} X_i\Big] = \frac{1}{N^2}\sum_{i=1}^{N} Var\Big[X_i\Big] = \frac{\sigma^2}{N}$

Empirical density functions

General approach for continuous densities

Suppose we have access to m samples $s_1, s_2, \dots, s_m$ from from an unknown continuous distribution. Since the distribution is unknown, we also don't know the shape of the density function. The density function is however a useful representation, since it often makes easy to get an understanding of the distribution, i.e. what are the values we can expect when sampling it. Therefore we are now interested in drawing an approximation of the density function for the unknown distribution. Of course the only information we have about the latter distribution is given by the m samples, so how can we exploit them to draw a meaningful approximation of the density function?

First, we define M regions, which constituting a partition of the density domain. We consider a scalar distribution, therefore we need to define M intervals covering $\mathbb{R}$: $\bigcup_{i=1}^{M} R_i = \mathbb{R}$. There are no further particular restrictions (e.g. they do not have to be equally spaced).

Then, for each sample $s_i,\ i = 1, \dots, m$ in our collection, we check which region it belongs to, and we assign a discrete probability mass of $1/m$ to the corresponding region. This procedure result in a frequentist approximation (a histogram!) of the unknown density function, which is more and more "fine-grained" (i.e. it will look more and more like a continuous function) as M increases. Be aware however that if M is big we need many many samples (indeed it must hold that $m >> M$) to have an accurate estimation of the relative frequency for each region!

Let $f_m(R_i)$ be the relative frequency of the region R_i (i.e. are of the corresponding rectangle in the histogram). Then we know that by the Law of Large numbers

$f_m(R_i) \rightarrow P[s \in R_i]$ as m goes to infinity.

Note: In case the die is fair ($\mathcal{D} \sim \mathcal{U}[F]$) it is pretty straightforward to directly compute all those quantities:
$\mu = \frac{F+1}{2}, \qquad \sigma^2 = \frac{F^2-1}{12}$

This would not be that straightforward for different distributions $\mathcal{D}$, nonetheless the code we developed so far can be applied to any discrete distribution. You can try to see what happens with an unfair die (whose distribution $\mathcal{D}$ is not uniform), you can also use a different way to enumerate its faces (e.g. using prime numbers 2,3,5,7,11,13,17,...)