Probability An Introduction

December 11, 2023

Abstract
We shall try to introduce the power and importance of probability theory at an undergraduate level as interestingly as possible!

1 The Setup

We will need an abstract language to deal with several things. We shall get to their definitions via some concrete examples
Let us say we have a population say of $1234567(N)$ people, and each one of them has a weight, an age, certain number of brothers, certain number of sisters and so on. Say we pick some at random. Here are the associated 'random' terminologies!

Preliminaries

- SAMPLE SPACE - is our population. This can be any set whatsoever!
- EVENT - is the set we choose. This is 'as of now' any subset of the sample space.
- RANDOM VARIABLE - weight, age, number of siblings. This is any function from the sample space to the Reals.(or anything really!)
- PROBABILITY - In this case $p(person) = 1/N$. Any distribution of a total mass of 1 on the elements of the sample space. More formally, any fuction $p : \Omega \to \mathbb{R}$
  1. $\sum_{\omega \in \Omega} p(\omega) = 1$
  2. $0 \leq p(\omega) \leq 1$
  3. $p(A \sqcup B) = p(A) + p(B)$

Remarks. Here we mention some properties of the above abstractions.

1. PROBABILITY should be viewed as MASS DISTRIBUTIONS. As we shall see later, it is possible to ask where is the center of mass? what is its moment of inertia? How does it move under deformation? and so on, corresponding to the questions of what its expected value, variance and variable transformation is.
2. RANDOM VARIABLES are AGGREGATING masses. Say there is a function $f : \Omega \to$ some set 'foo'. Then by associating with each element of foo the mass of all elements that mapped to it, we create a new mass distribution . In some sense coarser than the previous one. In this sense, The gender of that person can also be considered a random variable.
3. Such a 'setting' where one has a set and a mass distribution on it is called as a 'PROBABILITY SPACE'. It's just like saying that any place where ordering food makes sense is a food place (hotel). Here we say that any setting where asking probabilities makes sense as a probability space.
4. In this sense, Random variables help us define new probability spaces from earlier ones via functions and aggregations.
5. Random variables from $\Omega \to \Omega$ can also be thought of as transforming masses instead of an entirely new distribution. Its about whether you like to think about functions *actively* or *passively*.

Examples. Now for a nice collection of examples. In each of these, the sample space and mass function are as important as the experiment which they are relevant to.

- **Tossing coins** Ofcourse we toss coins. Say n times. Let the coin show up heads a q fraction of times. Then we are looking at a sample space of size 2^n with the probaility mass function being,

$$p(\text{a certain sequence of tosses}) = q^{\text{no. of heads}}(1-q)^{\text{no. of tails}}$$

  This can be seen as a way of constructing new probability spaces from old ones.

  Exercise. Say Ω_1, p_1 and Ω_2, p_2 are two probability spaces. Then the new set $\Omega_1 \times \Omega_2$ and the function $p((\omega_1, \omega_2)) = p_1(\omega_1) \times p_2(\omega_2)$ is a valid probability space. Also comment on the link between this and the tossing coins example above.

- **Toss multi-coins!** Basically throwing balls into bins (why?) where each bin has a certain probability of being chosen. This is also basically a product probability space.
  Now say the probability space for one toss is Ω where each bin is labelled with a natural number (Say the prize money corresponding to it at a carnival). Now this allows us to construct a random variable $X : \Omega \to \mathbb{N}$ the number corresponding to each bin. Which allows us to contruct a mass dstribution on naturals corresponding to probability of getting that number. Consider Ω^2 and the random variables $Y : \Omega^2 \to \mathbb{N}^2$ and $\overline{X} : \Omega^2 \to \mathbb{R}$ given by

$$Y(\omega_1, \omega_2) = (X(\omega_1), X(\omega_2))$$

$$\overline{X}((\omega_1, \omega_2)) = \frac{1}{2}(X(\omega_1) + X(\omega_2))$$

  What do they mean?

- **Keep tossing them....FOREVER!** Now the sample space is actually all infinite strings of H,T. An Infinite product of the space above. The probability of any paritucular element can be seen to tend to zero. However we all know that the set of strings with the 2023rd element as H is exactly half of the total. So some subsets have mass while induvidual elements do not! What kind of a mass distribution is this?

2 Distributions

Definition

Say we have a random variable defined on a probability space. Then the mass distribution on the range of the random variable is called the distribution of the random variable.

Whenever our probability space is 'discrete' we can capture all the information about the space in a single function from it to the reals. What exactly we mean by discrete is that the sample space is countable(What is that? Not very important for us). Basically should not be like the last example. Some distributions just occur way too often and have been extensively studied. Let us go through some of them.

Bernoulli

Its a single coin toss. Random variable is the following function $X : \Omega \to \mathbb{R}$

$$X(\text{heads}) = 1, X(\text{tails}) = 0$$

It can denote sevaral things like wether it will rain or not, will we win or not, or will a 1D randomwalker turn left or right and so on. Probably the most important instance of it is as the indicator random variable. In a certain sample space(Ω), random variables that are 1 for all elements(ω) in a certain subset(A) and 0 outside it are called indicator rvs. Denoted as $\mathbf{1}_{\omega \in A}$

Uncle Paul

Here's the problem statement.

> ATLEAST how big should a group of people be so that there must certainly be k-people who are all either mutually friends or all mutually enemies? We assume all people either know or not know each other. No in-betweens.

This number must exist. Is in itself non-trivial. This problem is called as the ramsey number problem. Seems like nothing to do with probabilities right? let see what **Uncle Paul** does. Consider the set of all friendship situations $\mathcal{C}$(formally, a function from pairs of people to $\{0,1\}$) make it a probability space with uniform mass function. Also the set of all k-sizes subsets of them $\mathcal{S}_k$(also a uniform probability function). Consider the rv $C : \mathcal{C} \to \mathbb{N}$ the number of k-subsets that are all friends or all enemies. Now we want to find the minimum of this rv. This has a very ugly distribution(try writing it down for small numbers). But we use the following slick approach. If we somehow show that the mean of C is less than one for some n, then for that n there must be some colouring where C takes the value zero. But how do we compute the mean of C? For this we extend our probability space from $\mathcal{C}$ to $\mathcal{C} \times \mathcal{S}$. Consider the following rv on the product space.

$$X(c, S) = 1 \text{ iff S is monochromatic under c}$$

where monochromatic means all friends or all enemies. Then,

$$C(c) = \sum_{S \in \mathcal{S}_k} X(c, S)$$

Now we can compute the mean of C as follows.

$$\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} C(c) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \sum_{S \in \mathcal{S}_k} X(c, S) = \frac{1}{|\mathcal{C}|} \sum_{S \in \mathcal{S}_k} \sum_{c \in \mathcal{C}} X(c, S)$$

Now the inner sum is just the number of colourings that keep S monochromatic. Dividing by $|\mathcal{C}|$ it denotes the fraction of colourings that keep S monochromatic. Which is easily seen to be 2^{1-k}. So the required mean is,

$$\sum_{S \in \mathcal{S}_k} 2^{1-k} = \binom{n}{k} 2^{1-k}$$

. This can be shown to be less than 1 for certain n. Meaning for that n, there are colourings that have no monochromatic k-subsets. So that becomes a lower-bound for the ramsey number. This is called the probabilistic method. It is a very powerful tool in combinatorics. It is also a very good example of how probability can be used to solve problems that are not directly related to it. There is in fact a whole book devoted to it written by **Noga Alon** and **Joel Spencer**.

Exercises.

1. The above proof can be cast in a more elegant manner in the language of graphs. DO SO!
2. We used the idea of a mean here. Try to define it for a finite probability space.
3. Consider the space of all random variables. Show that the mean is a linear function on this space. Is it a vector space?

Binomial

The notation $X \sim Bin(n, p)$ means it has the same distribution as the random variable defined on the nth power of the single coin toss probability space, as the count of the number of heads. Now on that space consider the

following rvs.

$$X_i(\omega) = \begin{cases} 1 & \text{if ith element of } \omega \text{ is heads} \\ 0 & \text{otherwise} \end{cases}$$

where ω is some element of the sample space. Then the rv X that counts the number of heads is given by

$$X(\omega) = \sum_{i=1}^{n} X_i(\omega)$$

Each of the X_i can be easily seen to obey bernoulli distribution. And the X_i s are independent. We can feel they should be independent, but what does it precisely mean though? for this we will see another way of getting probability spaces from existing ones.

Conditional Probability

Say the pair (Ω, p) form a probability space. Ω being the sample space and p being the mass function. Now consider a subset A of the sample space. The function p restricted to the set A can still be a mass function, except that the total mass won't be one. But we can always normalize! so we define a new mass function p_A on A as follows.

$$p_A(x) = \frac{p(x)}{p(A)}$$

where $p(A)$ is the sum of the masses of all elements of A. The space (A, p_A) is the conditional probability space of A given Ω.

Exercise. Check that the above does indeed define a mass function.(according to the definition above) This can be seen as

This can be thought of as updating the probabilities. Say earlier we knew that certain contestants(Ω) have certain probabilities(p) of winning. Now some subset of them have been eliminated. Then the *relative* probability of any two of the remaining contestants winning cannot be different as nothing really changed with them! So we update the probabilities by normalizing them with respect to the new sample space(A).

We have also introduced the notation of $p(A)$ as the total mass of A. which we shall use hereon. For any subset B of Ω, $B \cap A$ is a set in the conditional probability space.

$$p(B|A) := p_A(B \cap A)$$

read as the probability of B given A.

Ok...so now to independence.

Independence

Two events A and B are said to be independent if the info of one happenening provides no new info about the other. formally,

$$p(A|B) = p(A)$$

.

Two rvs X and Y are said to be independent if the value of one gives no info about the value of the other. formally,

$$p(X = x|Y = y) = p(X = x)$$

Remarks.

1. First verify that the relation of independence on subsets of sample space is a symmetric one as it is not explicit in the definition.

2. We have seen that rvs define a new mass function on their range based on the mass function of the sample space. So an alternate definition of independence of rvs is as follows. say $\mathcal{P}_X = (range(X), p_X)$ and $\mathcal{P}_Y = (range(Y), p_Y)$ are the probability spaces induced by rvs X and Y repectively. Consider the rv $Z : \Omega \to range(X) \times range(Y)$ as

$$Z(\omega) = (X(\omega), Y(\omega))$$

. Say $\mathcal{P}_Z$ is the probability space formed by Z. X,Y are independent iff $\mathcal{P}_Z = \mathcal{P}_X \times \mathcal{P}_Y$

3. The distribution Z defined from X and Y above is called as the joint distribution of X and Y.

Exercise. Try to give a natural definition of joint distribution of rvs $\{X_i\}_{i=1}^n$ all on the sampe sample space.

So in summary $Bin(n, p)$ is the distribution of sum of n independent bernoulli rvs with parameter p. One last definition! When two rvs of same range induce the same mass distribution they are said to be identically distributed(logically!) Hence

Bin(n,p) is a sum of n independent-identically-distributed(iid) ber(p) rvs

Exercise. Show that if $X \sim Bin(n, p)$ then its distribution on the range is given by

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Examples. Let's walk randomly!
Consider a particle at the origin, and at each time step, the particle moves either one unit to the right or one unit to the left, with equal probability. Let X_n be the random variable corresponding to the position of the particle after n-time steps. Each coin toss is basically a bernoulli rv with parameter $\frac{1}{2}$. So, X_n is basically binomial. Explicitly,

$$p(X_n = t) = p\left(\#\text{ of heads} = \frac{n+t}{2}\right) = \binom{n}{\frac{n+t}{2}} \left(\frac{1}{2}\right)^n$$

Below is a code to plot this distribution for different values of n.

```
import numpy as np
import matplotlib.pyplot as plt

one_step = np.array([1/2,0,1/2])
def walk(n, one_step):
    if n == 1:
        return one_step
    ret = one_step
    while n!=1:
        ret = np.convolve(ret, one_step)
        n -= 1
    return ret

def plot_walk(walk):
    l = (len(walk)-1)/2
    x = np.arange(-l, l+1)
    plt.plot(x, walk)
    plt.show()

plot_walk(walk(200, one_step))
```

Exercise. Run this script with different values of n. explain the several features of the resulting plot.(e.g, What is the differece between odd n and even n?) and there is a function used called as "convolve". What do you think it does?

Some clean observations can be made.

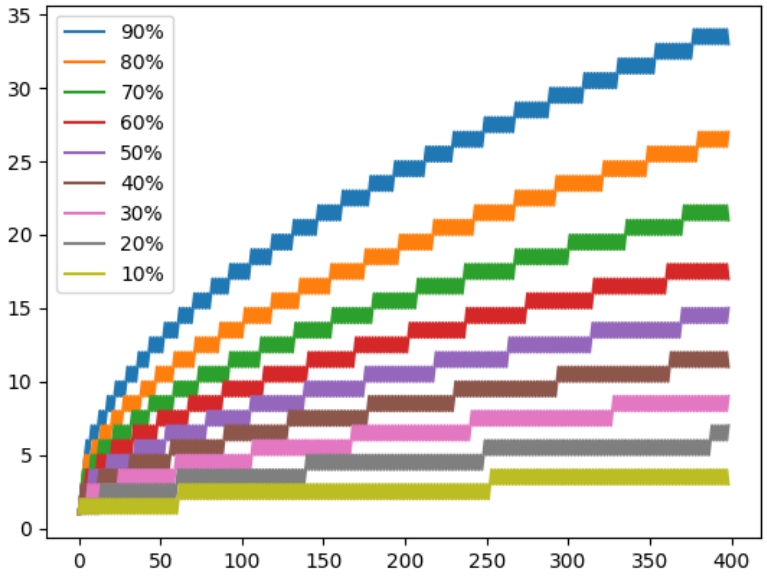
1. The distribution is symmetric. Afterall, the coin was fair!
2. The probability maximum is always near 0. A random-walker mostly gets no where!
3. But the ground he explores keeps increasing with n. This is also expected as given enough time he will explore the whole field.

We can clearly quantify the fact that he doesnt get any where on "average" by noting that the center of mass of the mass distribution is at 0. How do we quantify spread? We can see that the farther away from origin we check the less likeley he is to be found there. Try fixing an interval around the origin, say $(-a, a)$. What is the probability he escapes it eventually? Intuitively, for any finite a it becomes 1. Because for large n he defenitely will explore more. But what if a was a function of n? This gets ud to the following question.

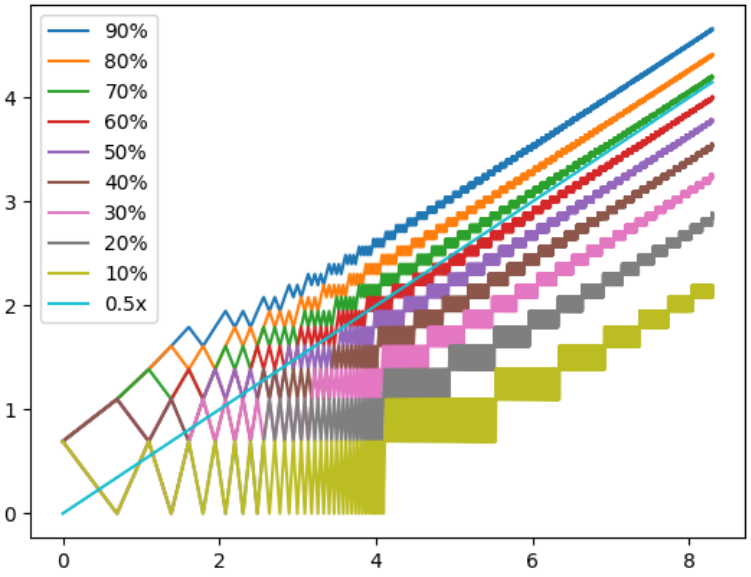
What is the function $a(n, p)$ *such that the probability of the random walker escaping* $(-a(n, p), a(n, p))$ *is* $1 - p$*?*

Exercise. Try to modify the code above to get that number.

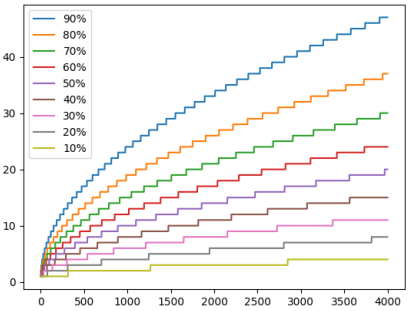
Its clear that for fixed p, the number $a(n, p)$ can be used as a measure of spread. Here are computed values of $a(n, p)$ for n from 1 to 400, and p- values 10%,20%,..,90%.

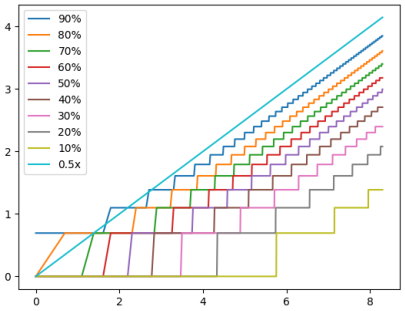


Does it look like $\sqrt{x}$? lets check it with a log plot(this time till n=4000!).



Initially its noise but eventually it stabilizes. Now what if we change the coin, say its fair but 80 percent of the times it jut refuses to answer. lands on its edges or somthing! The variable one step in the above code is set to $[0.1, 0.8, 0.1]$. The same two graphs are obtained for it and they are as follows.





Same results! Heck... it happens even if we make the coin unsymmetric or multivalued, as long as the mean is zero.

Exercise. If the mean is non-zero its true that the walker drifts off. But about his mean he still roams cluelessly. Concretize this. Simulate this for a variety of randomness choices. Say he chooses any natural number step with a probability proportional to to its inverse power 3. Does the same thing happen? what about power 4?

The above is certainly interesting. It is no coincidence that many distributions when added to each other seem to have some regularity. And by the way, when we defined the function $a(n, p)$ we made use of a very usefull quantity. The probability of the random walker being present in an interval. I computed it as follows.

1. Find the x such that p(walker in $[-\infty, x]$ after n steps) $= \frac{1-p}{2}$.
2. Answer is x...!(by symmetry of the distribution)

The quatity found in step 1 is called as the $\frac{1-p}{2}$-quantile of the distribution. We must have seen it being used in

ranking systems as percentile. The 50-quantile is called as the median. A very relevant function is the cumulative distribution function(CDF) of a random variable. It is defined as follows.

CDF

Given a real valued random variable, the *Cumulative Distribution Function* of it is a function $F_X : \mathbb{R} \to [0, 1]$ defined as

$$F_X(x) = p(X \leq x)$$

And it immediately follws that *q-quantile* of the rv is $F_X^{-1}(x)$, given that F_X is monotone. If say it is not invertible then one can use this fancy looking but arbitrary definition.

$$q - quantile = \inf\{x : F(x) = q\}$$

Apart from being crazy, it is also important! Say suppose we make a measurement, say of the distance between some stars or the wieght of some microbe. All experiments are bound to have some random errors. Now consider the random error to be an rv(say ϵ). It is supposed to have some distribution. Typically an experiment is performed multiple times and the mean is chosen as the result. Consider this. What if instead of saying the value is so and so, we say the value is in this interval with a probability(confidence) p? This seems like a more logical approach to quantify our surity.

$$\text{mean} = \frac{1}{n}\sum_{i=0}^{n}(\text{true-value} + \epsilon_i) = \text{true-value} + \frac{1}{n}\sum_{i=0}^{n}\epsilon_i$$

That summation is a sum n iid random variables. By the observation we made earlier, that value lies within $\pm a(n, p)$ with probability p and $a(n, p) \approx f(p)\sqrt{n}$ for large n. Hence,

$$|\text{true-value - mean}| \leq \frac{1}{n}(\sqrt{n}f(p)) = \frac{f(p)}{\sqrt{n}} \text{ with probability p}$$

So we need to make four times as many measurements to be twice as sure of our result(ehich we assume is our mean). We have arrived at this without any assumption on the distribution of the error! But may be we skipped a few steps. Let us fill in some gaps.

Remarks. We have only arrived at the result that the confidence itervals grow as $\sqrt{n}$ for a random walker empirically. We have yet not given even a hint of its proof, although its clear that it is going to be purely mathematical. Even without knowing the proof we were able to use the result to derive some statistical result of some significance. This is the whole endeavour of statistics. To use data and probability theory(or any other tool) to infer precise things about the world.
One more thing to note is that the rv ϵ is a continuous vlaued rv. We applied our result to it without any hesitation. Until we formalize what are continuous distributions, just think of continuous distributions as "discrete" ones witht the values being so closely packed that it resembles the continuum. This is not a very good definition but it will do for now.