Lecture 31 — Probability, Convergence, & Ergodicity

Jeff Zarnett

jzarnett@uwaterloo.ca

Department of Electrical and Computer Engineering
University of Waterloo

December 9, 2019

Suppose we are doing an experiment like rolling dice.

We associate the outcomes of these experiments with is called a random variable.

A random variable X has many instances, each with some probability.

If you roll a fair die X can have instances of $1, 2, 3, 4, 5, 6$, each with a probability of $1/6$ (because we defined this die to be perfectly fair).

In a die roll, the events (values of the roll) are discrete—you can't roll 2.1 or 5.5.

Time is generally modelled as continuous, but in reality there are limits to how accurately we can measure time on a computer.

The resolution depends on the hardware.

Building on this, we can think of the outcome of an experiment as an "event".

We can ask ourselves all kinds of questions like: what is the probability of rolling four 1s on a 20-sided-die in a row?

Looking at this from the perspective of application programmers, we say the number of users at time t can be represented by a random variable U.

Then we can ask ourselves what is the probability of U being greater than 1000?

Probability is usually defined as some sort of experiment with sample space Ω; some subset E of Ω is an event. (In full generality, sample spaces are measure spaces.)

A standard example: you have a perfectly fair coin.

Possible outcomes are heads, tails, and very rarely, edge (it could happen).

Each outcome is an event with a certain probability—i.e., chance of happening—usually written $P\{E_n\}$.

If the coin coming up heads is an event E_1 and it coming up tails E_2 are these independent?

No, they are not.

They are mutually exclusive—if the coin comes up heads it cannot possibly also come up tails.

Events are (measurable) sets and we can do unions and intersections and such.

The formal definition of mutual exclusivity is $E_1 \cap E_2 = \emptyset$.

Independence, on the other hand, means the probability of E_1 does not change if event E_2 has occurred.

Conditional probability is also important: what is the probability of event E given that event F has occurred?

The notation is $P\{E \mid F\}$ and it has a definition (as long as $P\{F\}$ is not zero):

$$P\{E \mid F\} = \frac{P\{E \cap F\}}{P\{F\}}$$

If events E and F are independent, then $P\{E \cap F\}$ is equal to $P\{E\} \cdot P\{F\}$.

That is, the probability of E is not affected by the occurrence of F.

Suppose you flip a coin four times and it comes up heads each time (event F).

What is the probability that the next coin flip will be tails (event E)?

Sadly, a lot of people think that after four heads, now tails is "due" and the chance of it being tails is higher than the usual 50-50 (minus the edge case)[1].

This is not true—the next coin flip is independent of all the coin flips that went before it.

---

[1]Wikipedia will tell you about the Monte Carlo Casino in 1913.

If we want $P\{F \mid E\}$, but have $P\{E \mid F\}$ can we get it?

Yes, if we know $P\{F\}$ and $P\{E\}$.

Bayes law is:

$$P\{F \mid E\} = \frac{P\{E \mid F\} \cdot P\{F\}}{P\{E\}}$$

For now let us focus on discrete random variables (i.e., those with a countable number of values).

For a random variable X we can define a probability mass function (p.m.f.):

$$p_x(a) = P\{X = a\}, \text{where} \sum_x p_x(x) = 1$$

and the cumulative distribution function is:

$$F_x(a) = P\{X \leq a\} = \sum_{x \leq a} p_x(x)$$

There are five distributions we want to discuss:

1. Bernoulli
2. Binomial
3. Geometric
4. Poisson
5. Exponential

Bernoulli represents a coin flip: so the coin has a probability p of coming up heads and $1 - p$ of coming up tails.

So the random variable X associated evaluates to 1 (heads) with probability p and 0 (tails) with probability $1 - p$.

Thus the p.m.f. is $p_x(1) = p$ and $p_x(0) = 1 - p$.

The Binomial distribution builds upon the the Bernoulli distribution.

A coin with probability p of coming up heads flipped n times (assuming coin flips are independent).

The random variable X, if it's Binomial, represents the number of heads when flipping a Bernoulli coin n times.

So X can take on values of $\{0, 1, ..., n\}$.

The Geometric distribution builds on Bernoulli.

Again suppose we have the coin with probability p of coming up heads.

If we want to ask how many (independent) coin flips it will take until it comes up heads, the Geometric distribution answers this question.

X is the number of flips until we get the result we want.

Suppose we have a server farm that has n disks, each of which independently dies (fails) with probability p in a year.

So what is the appropriate model for each of these questions?

1. How many disks die in the first year?
2. Given disk d, how long until this disk dies?
3. After one year, is a specific disk d alive or dead?

The answers are (1) Binomial, (2) Geometric, and (3) Bernoulli.

Now let us turn to the Poisson distribution; this is very common in computing.

This is our somewhat "natural" idea of randomness.

This distribution arises when we have a mixture of a very large number of sources, each with a very small probability.

So this is a lot like the number of arrivals to a website or packets at a router in some unit time.

If a random variable X has a Poisson distribution, the p.m.f. is:

$$p_x(i) = \frac{e^{-\lambda}\lambda^i}{i!}, \text{ where } i = 0, 1, 2....$$

We should also consider the exponential distribution.

This is a distribution where the probability mass function (p.m.f.) drops off exponentially.

If $x \geq 0$, then $f_x(x) = \lambda e^{-\lambda x})$.

The cumulative distribution function, i.e., $F_x(x) = P\{X \leq x\}$ is the integration of this function from negative infinity to x, equal to $1 - e^{-\lambda x}$ (again, for $x \geq 0$).

What we care about here is actually the inverse of this, $\overline{F_x}(x)$, the probability that the value is greater than x.

These are the outliers or the "tail" of the distribution.

Note that we see a constant drop-off of $e^{-\lambda}$ with each unit increase of x.

Heavy tails are not very nice.

Recall the notation here $E[X]$ is the expected value of a random variable X.

Expected value is, not kidding, just a "weighted average".

Let's say you're entering into a game where you pay $3 to play.

If you have a 50% chance of losing all your money, a 10% chance of winning $10, and a 40% chance of getting your $3 back, should you play?

Do the math: your expected return is
$(0.5 \times 0) + (0.1 \times 10) + (0.4 \times 3) \rightarrow 0 + 1 + 1.2 = 2.2$.

So you expect that by playing this game, for every \$3 entry, you get back \$2.20 (on average).

It doesn't seem like a good investment, does it?

Econ 101 logic would tell you the answer is: how much do you value the fun of playing and the hope and excitement of winning?

If it's equal to or more than \$0.80 then it is worth it.

Proofs of these exist in the textbook(s) but let's not belabour the point by proving it.

Just take my word for it that the expectation for a Geometric distribution with probability p is $1/p$.

If a Poisson distribution applies to something, the expectation $E[\lambda] = \lambda$.

The weighted average is nice, but it's not the whole story.

If your numbers are 5, 6, and 7, then their average is 6.

But the average of the numbers 0, 6, and 12 is also 6.

But we should know already that averages can be misleading.

An average arrival rate of 1 request per second could still, with some probability, produce 3 arrivals in a single second.

The formal definition of variance of a random variable X is the expected squared difference of X from its mean:

$$Var(X) = E[X^2] - (E[X])^2$$

Just take my word for it at the moment that the following table is correct:

Distribution	Mean	Variance
Bernoulli	p	$p(1-p)$
Binomial	np	$np(1-p)$
Geometric	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Poisson	λ	λ

Think back to calculus class.

Remember when we talk about limits: $\lim_{x \to \infty} f(x)$.

We would like to see that our random variables converge.

You might flip a coin four times and all four times it comes up heads.

That doesn't match our expectation that we should have about half heads and half tails.

We have convergence if, given enough samples and enough sample paths, it will converge to the 0.5 we expect.

There may be some sample paths that don't converge (e.g., continually coming up heads), but they have a "probability mass" of zero (i.e., they are incredibly unlikely).

There are in fact uncountably many "bad paths", each with probability zero (but that's okay). Zero probability doesn't mean it can't happen, mind you.

An image of what convergence looks like:

We won't concern ourselves with systems where there is no convergence.

A small but important digression on the subject of sampling, measurement, and testing.

You have an idea of what an average is, but there are two different types of average—the time average and ensemble average.

Let us just focus on having a single First-Come-First-Serve queue.

Every second, a new job arrives with probability p and if there is any work to do, the job being worked on is completed with probability q (and $q > p$).

As a definition, let $N(v)$ equal the number of jobs in the system at a time v.

In the story, Tim and Enzo are trying to simulate the FCFS system to determine what is the average number of jobs in the system.

Tim decides he's going to run it as one really long simulation.

He simulates the queue over a very long period, logging as he goes, taking a million samples.

Then he takes the average value over those samples to get the average number of jobs.

Enzo does something slightly different: instead of having one super long simulation, he does 1000 shorter simulations.

He waits until the simulation has run for 1000 seconds and then samples the queue at exactly that point, obtaining one value.

This experiment is restarted with a new random seed.

So after obtaining a thousand samples, he averages these, and Enzo produces another average number of jobs.

Tim
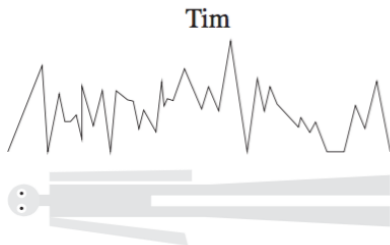
Enzo

So—who has done this correctly, Tim or Enzo?

Time Average vs Ensemble Average

The time average has potential problems because we are only looking at a single sequence and maybe something very unusual has happened here in this single run.

The ensemble average is more likely what we talk about when we talk about the system being at "steady state" (i.e., past the initial conditions).

So we kind of like the Enzo approach. Tim's approach still has some merit though.

Both the Tim and Enzo approaches here require caring about the initial conditions.

Enzo needs to make sure that the initial conditions (startup costs etc) have attenuated before the measurement point.

Tim needs to ensure that the initial conditions impact a sufficiently small portion of all his measurements.

If we have a nicely behaved system, the time average and the ensemble average are the same (so both Tim and Enzo can be correct).

What is a nicely behaved system? The word for this is ergodic.

That probably did not help, so what is an ergodic system?

It is a system that is positive recurrent, aperiodic, and irreducible.

Irreducibility means a process should be able to get from one state to any other state (where state is the number of jobs in the system).

This means the initial state of the system does not matter.

So if we started at 0 jobs or 10 we could still get to any state in the system (jobs at 2 or 27)...

Positive recurrence means that given an irreducible system, any state i is revisited infinitely often, and the time between visits to that state are finite.

So we can define a certain state as being a "restart".

The logical choice in the case of a queue or similar is the idea of the queue being empty.

Every time the queue gets down to zero jobs, it's a "restart" of sorts.

This is what makes Tim's view and Enzo's view potentially the same.

A single long run (Tim's view) is just like a number of independent runs (Enzo's view).

Every time we get down to zero jobs in the queue, it's a restart.

The aperiodicity condition is required for the ensemble average to make sense or exist.

That is to say, the state of the system should not be related to the the time.

Otherwise the way Enzo chooses to sample, i.e., $t = 1000$, is potentially going to skew the result.

Restart   Restart   Restart   Restart   Restart
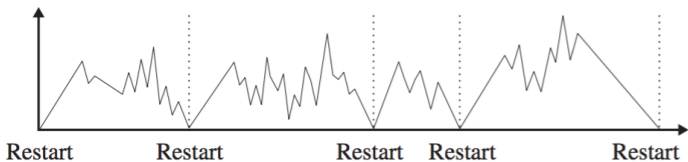
So both Tim and Enzo are correct, given an ergodic system.

We could compute either the time or ensemble average.

$$\text{Time Average} = \lim_{t \to \infty} \frac{\sum_{i=1}^{A(t)} T_i}{A(t)}$$

where $A(t)$ is the number of arrivals by time t and T_i is the time in the system of arrival i. The average is taken over one sample path.

$$\text{Ensemble Average} = \lim_{t \to \infty} E[T_i]$$

where $E[T_i]$ is the average time in the system of job i, where the average is taken over all sample paths.