8. First passage and continuous random variables

Last time

- Markov chains
- Master equation and exact enumeration

Goals for today

- Exact enumeration for first-passage times
- Continuous random variables
- Histograms and probability density function
- Cumulative distribution function
- Central limit theorem

Exact distribution of first-passage times

- Suppose a simple symmetric 1D walk starts at 1
- Let τ be hitting time of origin, i.e. first time to reach 0
- What is expected (mean) time $\langle \tau \rangle$?
- What is probability distribution of τ?
- Solve by exact enumeration

Master equation

- $P_i^0 = \delta_{1,i}$ – prob. concentrated at 1
- Master equation

$$P_i^{t+1} = \tfrac{1}{2}P_{i-1}^t + \tfrac{1}{2}P_{i+1}^t \quad \forall 1 < i < L$$

- Absorbing at 0 so nothing returns from 0 to 1:

$$P_1^{t+1} = \tfrac{1}{2}P_2^t$$

- Reflecting at L:

$$P_L^{t+1} = \tfrac{1}{2}P_{L-1}^t + \tfrac{1}{2}P_L^t$$

Absorption

- $\frac{1}{2}P_1^n$ jumps to 0 at nth step
- Store this as $\mathbb{P}(\tau = n)$
- Set $P_0^n := 0$.
- Alternative viewpoint: Probability remains at site 0 (instead of leaving system).

Code

- Use `OffsetArrays.jl` to have array with index starting at 0:

```
using OffsetArrays

function first_passage_distribution(L=20, T=100)

    P = OffsetArray([0.0; 1.0; zeros(L-1)], 0:L)
    next_P = similar(P)

    absorption_prob = Float64[]
```

```
for t in 1:T
      for i in 1:L-1
          next_P[i] = 0.5*(P[i-1] + P[i+1])
      end

      next_P[L] = 0.5 * (P[L-1] + P[L])
      next_P[0] = 0.5 * P[1]

      push!(absorption_prob, next_P[0])
      next_P[0] = 0.0

      next_P, P = P, next_P
  end
```

```
function calculate_and_plot(L, T)
    absorption_prob = first_passage_distribution(L, T)

    plot!(absorption_prob[1:2:end],
        xscale=:log10, yscale=:log10, alpha=0.4, label="$L", lw=3)
end
```

Mean of distribution

- Calculate mean as

$$\langle \tau \rangle = \sum_n n \, \mathbb{P}(\tau = n)$$

- How does it behave as function of L?
- Or define mean hitting time $T_i(L)$ starting from site i with boundary at L.
- Obtain system of linear equations for $T_i(L)$.
- Get *infinite mean* for $L = \infty$

Code for mean

- Calculate mean for **probability mass function** (PMF)

  (what we have so far called the “probability distribution”)

```
function mean_of_distribution(pmf)
    return sum(n * pmf[n] for n in 1:length(pmf))
end
```

Continuous random variables

What is a continuous random variable?

- Recall Monte Carlo simulation of π: throw darts at unit disc
- Obtain value that is **random variable** (result of random process)
- Takes **continuous values**: any real number between 0 and 1.
- So called **continuous random variable**

Summary statistics

- **Mean** and **variance** make sense, just as for discrete random variables.
- How describe **probability distribution** of continuous random variable?
- For discrete random variable *count* number of times each value occurred
- Impossible for continuous random variables
- Uncountably infinite possibile values for outcome

We can't count

- For (many) continuous random variables X we have $\mathbb{P}(X = x) = 0 \quad \forall x$
- Never expect to repeat outcomes in a simulation
- Counting is useless!
- But values still concentrate around π (mean / expectation) as in discrete case
- How replace counting?

Probability density function (PDF)

- Idea: Calculate $\mathbb{P}(a \leq X \leq b)$
- I.e. prob. that outcome *lies in certain range*
- For discrete r.v.s this is the *sum* of probabilities
- Analogous idea for continuous r.v.s: *integral*
- So "expect"

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\, dx$$

for some function f_X

- NB: This is *not* always true

Probability density function II

- f_X is the **probability density function of** X
- $f_X(x)\,dx$ is prob. that $X \in [x, x + dx]$
- f_X is not a probability; it's a *density* of probability

Calculating a PDF: histograms

- It's "easy" to calculate approximations of the PDF
- Fix **bin width** h
- Bin edges $x_n := x_0 + h\,n$
- *Count* points in $[x_n, x_{n+1})$
- Do this for several such intervals to get **histogram**

Histograms II

- Draw bar whose *area* is proportional to frequency in that bin
- Sum of areas = 1
- How choose bin width?
- Choose to give “best” result. Several interpretations
- Alternative: **kernel density estimate**: for each x, count number of points near x

Histograms in Julia

- Three options:

  1. Make your own!
  2. `histogram(data)` function in `Plots.jl`:
     - Draws histogram
     - Does not allow access to data in histogram
  3. `fit(Histogram, data)` in `StatsBase.jl`:
     - Need `StatsPlots.jl` to plot
     - Returns data

fit(Histogram, data)

```
using StatsBase

data = rand(100)

h = fit(Histogram, data, nbins=50)

using StatsPlots
plot(h)
```

Cumulative distribution function (CDF)

- Histograms lose information: lump data together in single bin
- Cumulative distribution function does not lose information:

$$F(x) := \mathbb{P}(X \leq x)$$

- Empirical CDF: Step function that increases at each data point

Normal distribution

- PDF of standard normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{x^2}{2}\right)$$

- Famous bell curve
- CDF cannot be written in terms of standard functions
- Introduce new “error function”, erf
- Quadratic on log-linear (log y-axis)

Why is the normal distribution so ubiquitous?

- **Central limit theorem**:
  Sum of independent random variables converges to a normal distribution
- Limiting shape of “centre” of distribution (not tails)
- Summands (things being summed) can be *different*

Why is the CLT true?

- Dice example (PS2): means increase linearly; standard deviations increase *slower*
- So everything concentrates around mean with zero (relative) width in limit
- CLT: centre around mean and *rescale*; obtain limiting normal shape
- Says how positive and negative deviations tend to cancel each other
- PDF does *not* always “converge”: **weak convergence**

Does the Central Limit Theorem always hold?

- No!
- Only if mean and variance are finite
- e.g. Sample from a Pareto distribution (power-law tail)

```
α = 4
data = [sum(rand(Pareto(α, 1.0), 100)) for i in 1:10000]
histogram(data)  # satisfies CLT


α = 1.5
data = [sum(rand(Pareto(α, 1.0), 100)) for i in 1:10000]
histogram(data)  # doesn't satisfy CLT
```

- Then convergence to other distributions: Lévy stable distributions
- Long tail often corresponds to some kind of "memory

Review

- Exact first-passage distribution and diverging (infinite) mean hitting time
- Continuous random variables
- Probability density function (PDF)
- Central Limit Theorem