7. Markov chains and continuous random variables

Last time

- Defining our own type
- Abstract types

- Specifying a method (version) of a function to call based on type: **dispatch**

Goals for today

- Markov chains
- Continuous limit
- Continuous probability distributions
- Probability density function

Markov chains

- Let's think back to modelling recovery:
- At each step, with probability p you recover

Markov chains

- Let's think back to modelling recovery:
- At each step, with probability p you recover
- Call infected state 0, recovery state 1

Markov chains

- Let's think back to modelling recovery:
- At each step, with probability p you recover
- Call infected state 0, recovery state 1
- Then have directed **transition graph**

$$0 \xrightarrow{p} 1$$

Markov chains II

- Introduce some notation:
- X_n is state at time n
- $X_0 = 0$ (infected)

Markov chains II

- Introduce some notation:
- X_n is state at time n
- $X_0 = 0$ (infected)

- Let $P_i^{(n)}$ be $\mathbb{P}(X_n = i)$ = probability that in state i at time n

Markov chains II

- Introduce some notation:
- X_n is state at time n
- $X_0 = 0$ (infected)

- Let $P_i^{(n)}$ be $\mathbb{P}(X_n = i)$ = probability that in state i at time n
- Then

$$P_0^{(n+1)} = (1-p)P_0^{(n)}$$

$$P_1^{(n+1)} = pP_0^{(n)} + 1P_1^{(n)}$$

Markov chains III

- Have

$$P_j^{(n+1)} = \sum_i p_{i\to j} P_i^{(n)}$$

Markov chains III

- Have

$$P_j^{(n+1)} = \sum_i p_{i \to j} P_i^{(n)}$$

- Set $\mathbf{P}_n := \begin{bmatrix} P_0^{(n)} & P_1^{(n)} \end{bmatrix}$

Markov chains III

- Have

$$P_j^{(n+1)} = \sum_i p_{i \to j} P_i^{(n)}$$

- Set $\mathbf{P}_n := \begin{bmatrix} P_0^{(n)} & P_1^{(n)} \end{bmatrix}$
- Then

$$\mathbf{P}_{n+1} = \mathbf{P}_n \mathsf{M}$$

- With **transition matrix** $\mathsf{M} := (p_{i \to j})$
- We have

Markov chains IV

- So
$$\mathbf{P}_{n+1} = \mathbf{P}_n \mathsf{M}^n$$
- Dynamics is just multiply by a matrix at each step!

Markov chains IV

- So

$$\mathbf{P}_{n+1} = \mathbf{P}_n \mathsf{M}^n$$

- Dynamics is just multiply by a matrix at each step!
- Discrete-time **Markov chain**

Continuous random variables

- What is a continuous random variable?
- Random procedure where outcome can take **continuous range of values**
- E.g. `rand()`: outcome 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