8. Continuous random variables

Last time

- Dynamics of discrete stochastic processes
- Markov chains
- Stationary distribution

Goals for today

- Continuous limit
- Continuous probability distributions
- Probability density function

Geometric random variable

- Recall **geometric random variable** X

Geometric random variable

- Recall **geometric random variable** X
- Integer time steps
- Probability to decay / recover at each time step is p

Geometric random variable

- Recall **geometric random variable** X
- Integer time steps
- Probability to decay / recover at each time step is p

- $\mathbb{P}(X = n) = p(1-p)^{n-1}$ for $n = 1, 2, \ldots$

Probability mass function

- For a **discrete(-valued) random variable** X:
- Define **probability mass function** f_X as

$$f_X(n) := \mathbb{P}[X = n]$$

i.e. the probability that X takes the value n

Probability mass function

- For a **discrete(-valued) random variable** X:
- Define **probability mass function** f_X as

$$f_X(n) := \mathbb{P}[X = n]$$

i.e. the probability that X takes the value n

- For geometric random variable X we have

$$f_X(n) := p(1-p)^{n-1}$$

Probability mass function

- For a **discrete(-valued) random variable** X:
- Define **probability mass function** f_X as

$$f_X(n) := \mathbb{P}[X = n]$$

i.e. the probability that X takes the value n

- For geometric random variable X we have

$$f_X(n) := p(1-p)^{n-1}$$

- Satisfies $\sum_{n=1}^{\infty} f_X(n) = 1$

Towards continuous time

- Suppose want jumps not only at integer times
- Jump at times $t_n := n\,\delta$ for $n = 1, 2, \ldots$

Towards continuous time

- Suppose want jumps not only at integer times
- Jump at times $t_n := n\,\delta$ for $n = 1, 2, ...$
- E.g. if $\delta = \frac{1}{2}$, jump twice per day

Towards continuous time

- Suppose want jumps not only at integer times
- Jump at times $t_n := n\,\delta$ for $n = 1, 2, \ldots$
- E.g. if $\delta = \frac{1}{2}$, jump twice per day
- If we want to *reproduce* same dynamics, how should we choose decay probability p?

Towards continuous time

- Suppose want jumps not only at integer times
- Jump at times $t_n := n\,\delta$ for $n = 1, 2, \ldots$
- E.g. if $\delta = \frac{1}{2}$, jump twice per day
- If we want to *reproduce* same dynamics, how should we choose decay probability p?
- p must depend on δ, so call it $p(\delta)$

Towards continuous time

- In order to compare the two different processes, can't look at instantaneous jumps

Towards continuous time

- In order to compare the two different processes, can't look at instantaneous jumps
- Need to look at **cumulative distribution function**:

$$F_X(n) := \mathbb{P}[X \leq n]$$

Towards continuous time

- In order to compare the two different processes, can't look at instantaneous jumps
- Need to look at **cumulative distribution function**:

$$F_X(n) := \mathbb{P}[X \leq n]$$

- "The probability that have decayed *by time* n"

Towards continuous time

- In order to compare the two different processes, can't look at instantaneous jumps
- Need to look at **cumulative distribution function**:

$$F_X(n) := \mathbb{P}[X \leq n]$$

- "The probability that have decayed *by time* n"
- Actually now can talk about any *continuous* (real) time t:

$$F_X(t) := \mathbb{P}[X \leq t]$$

Towards continuous time II

- How can we write the cumulative distribution function in terms of f_X?

Towards continuous time II

- How can we write the cumulative distribution function in terms of f_X?
- We need to find the largest integer value less than t

Towards continuous time II

- How can we write the cumulative distribution function in terms of f_X?
- We need to find the largest integer value less than t
- We obtain

$$F_X(t) = \sum_{j=1}^{\lfloor t \rfloor} f_X(j)$$

Towards continuous time II

- How can we write the cumulative distribution function in terms of f_X?
- We need to find the largest integer value less than t
- We obtain

$$F_X(t) = \sum_{j=1}^{\lfloor t \rfloor} f_X(j)$$

- $\lfloor t \rfloor$ is the **floor** function:
- $\lfloor t \rfloor$:= largest integer $\leq t$
- `floor(t)` in Julia

Towards continuous time III

- Now think about jumps at times $t_n = n\delta$
- Call resulting process Y_δ

Towards continuous time III

- Now think about jumps at times $t_n = n\delta$
- Call resulting process Y_δ
- Also a discrete(-valued) random variable

Towards continuous time III

- Now think about jumps at times $t_n = n\delta$
- Call resulting process Y_δ
- Also a discrete(-valued) random variable

- Its probability mass function is

$$f_{Y_\delta}(n) = p(\delta)\left[1 - p(\delta)\right]^{n-1}$$

Towards continuous time III

- Now think about jumps at times $t_n = n\delta$
- Call resulting process Y_δ
- Also a discrete(-valued) random variable

- Its probability mass function is

$$f_{Y_\delta}(n) = p(\delta)\,[1 - p(\delta)]^{n-1}$$

- This is the probability after n *time-steps*
- But now these jumps occur at *times* $n\delta$

Towards continuous time IV

- We need $F_{Y_\delta}(t)$
- Probability that Y_δ has decayed by *time* t

Towards continuous time IV

- We need $F_{Y_\delta}(t)$
- Probability that Y_δ has decayed by *time* t

- We showed on Problem Set 2 that

$$F_X(n) = 1 - (1-p)^n$$

Towards continuous time IV

- We need $F_{Y_\delta}(t)$
- Probability that Y_δ has decayed by *time* t

- We showed on Problem Set 2 that

$$F_X(n) = 1 - (1-p)^n$$

- For Y_δ we use $t = n\,\delta$, so $n = t/\delta$

Towards continuous time IV

- We need $F_{Y_\delta}(t)$
- Probability that Y_δ has decayed by *time* t

- We showed on Problem Set 2 that

$$F_X(n) = 1 - (1-p)^n$$

- For Y_δ we use $t = n\,\delta$, so $n = t/\delta$
- So

$$F_{Y_\delta}(t) = 1 - [1 - p(\delta)]^{\lfloor t/\delta \rfloor}$$

Continuous limit

- We could impose that at integers n we want to have the same probability as for the original process X

Continuous limit

- We could impose that at integers n we want to have the same probability as for the original process X
- Instead let's look directly at the **continuous limit**
- i.e. $\delta \to 0$

Continuous limit

- We could impose that at integers n we want to have the same probability as for the original process X
- Instead let's look directly at the **continuous limit**
- i.e. $\delta \to 0$
- We need to take $p(\delta)$ in a way that "makes sense"
- I.e. where nothing goes to ∞ or to 0

Continuous limit II

- When $\delta \to 0$ we neerd $p(\delta) \to 0$, since otherwise we will have an infinite number of finite decays in a finite time

Continuous limit II

- When $\delta \to 0$ we neerd $p(\delta) \to 0$, since otherwise we will have an infinite number of finite decays in a finite time
- When $\delta \to 0$, we can approximate

$$1 - p(\delta) \simeq \text{exp}[-p(\delta)]$$

Continuous limit II

- When $\delta \to 0$ we neerd $p(\delta) \to 0$, since otherwise we will have an infinite number of finite decays in a finite time
- When $\delta \to 0$, we can approximate

$$1 - p(\delta) \simeq \text{exp}[-p(\delta)]$$

- So

$$F_{Y_\delta}(t) \simeq 1 - \{\text{exp}[-p(\delta)]\}^{t/\delta}$$

Continuous limit II

- When $\delta \to 0$ we neerd $p(\delta) \to 0$, since otherwise we will have an infinite number of finite decays in a finite time
- When $\delta \to 0$, we can approximate

$$1 - p(\delta) \simeq \exp[-p(\delta)]$$

- So

$$F_{Y_\delta}(t) \simeq 1 - \{\exp[-p(\delta)]\}^{t/\delta}$$

- Hence

$$F_{Y_\delta}(t) \simeq 1 - \{\exp[-p(\delta)]\}^{t/\delta}$$

Continuous limit III

- We find

$$F_{Y_\delta}(t) \simeq 1 - \mathsf{exp}\left[-p(\delta)\frac{t}{\delta}\right]$$

Continuous limit III

- We find

$$F_{Y_\delta}(t) \simeq 1 - \exp\left[-p(\delta)\frac{t}{\delta}\right]$$

- Hence we need $p(\delta) \propto \delta$ when $\delta \to 0$

Continuous limit III

- We find

$$F_{Y_\delta}(t) \simeq 1 - \exp\left[-p(\delta)\frac{t}{\delta}\right]$$

- Hence we need $p(\delta) \propto \delta$ when $\delta \to 0$
- Take

$$p(\delta) =: \lambda\,\delta$$

Continuous limit III

- We find

$$F_{Y_\delta}(t) \simeq 1 - \exp\left[-p(\delta)\frac{t}{\delta}\right]$$

- Hence we need $p(\delta) \propto \delta$ when $\delta \to 0$
- Take

$$p(\delta) =: \lambda\,\delta$$

- λ is called the **rate** of the continuous process
- Decay probability *per unit time*
- Limiting **continuous random variable** with

Exponential distribution

- A random variable Z with the cumulative distribution function

$$F(t) = 1 - \exp(-\lambda t)$$

is called an **exponential random variable**

Exponential distribution

- A random variable Z with the cumulative distribution function

$$F(t) = 1 - \text{exp}(-\lambda t)$$

is called an **exponential random variable**

- $F_Z(t) = \mathbb{P}(Z \leq t)$
- Can write

$$F_Z(t) = \int_{0^t} f_Z(s)\,ds$$

- $f_Z(s)$ is the **probability density function**

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:

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

- 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