Distributions

Why is knowledge of probability distributions important?

- Many ML algorithms are basically trying to find the probability distribution of the underlying data. So understanding the math could help if you observe weird behavior in your models later on.
- It's the foundation for hypothesis testing which makes some assumptions about the probability distributions of the data to make reliable conclusions.
- Some problems can directly be solved with probability without the need for complex ML algorithms if you know how.
- Helps you understand the limitations of ML. Since ML is probabilistic in nature, no prediction is ever guaranteed to happen. There's always a probability for any prediction to really happen.
- Popular interview questions.

Random variable

- A function enabling mapping from a **sample space S to real numbers** (or) it's just a variable whose value is a **numerical outcome** from some random event
- Example experiment: flipping a coin twice
    - Sample space $\boldsymbol{S}$: $\{\{H,H\},\{T,H\},\{H,T\},\{T,T\}\}$
    - *(Discrete)* Random variable X:
        - Let's say for an event: the number of heads I observe.
            - Then, X = 0, 1, or 2
- Example experiment: rolling 6 sided die once
    - Sample space $\boldsymbol{S}$: {1,2,3,4,5,6}
    - *(Discrete)* Random variable Y:
        - Event: Getting a given outcome from a die-roll experiment.
            - Then, Y = 1, 2, 3, 4, 5, 6

Random variable types

Random variable

Discrete

"Countable" outcomes

Examples:

- Die roll
- Deck of cards
- Churn
- Sentiments

Continuous

"Uncountable" outcomes

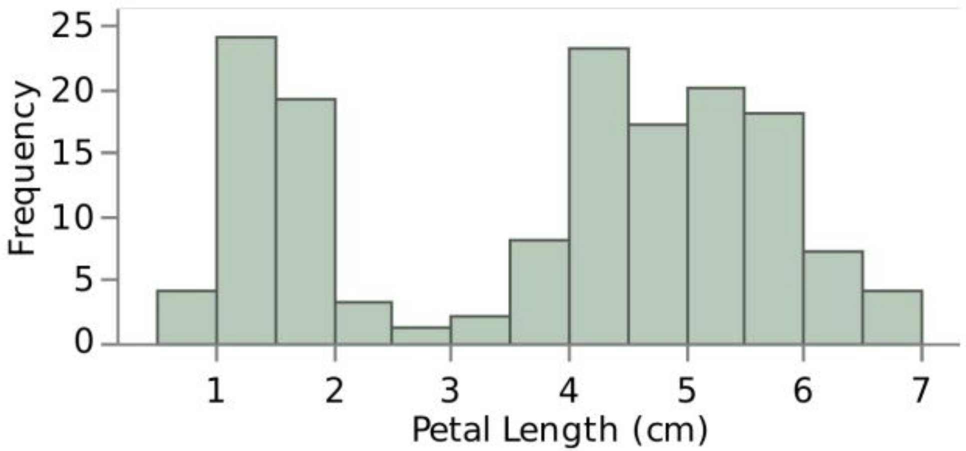
Examples:

- Temperature
- House price
- Age

Distribution

- **Distribution** is a **spread or range of data**. (recap histogram?)

A **distribution** is the set of all values of a variable and how frequently we observe each value.



Probability distributions: Discrete vs Continuous

- A probability distribution is a way to represent the possible values of a random variable (the numerical outcomes) and the respective **probabilities** of those values. - *it's a distribution of probabilities of values that a random variable takes!*

<table>
<tr><th>Discrete probability distributions</th><th>Continuous probability distributions</th></tr>
<tr><td>describes probabilities associated with discrete outcomes of a discrete random variable</td><td>describes probabilities associated with continuous outcomes</td></tr>
<tr><td>

pmf (probability mass function):

f(x) = P(X=x)
Function describing Probability that discrete random variable is exactly equal to some value (A single discrete outcome)

cdf (cumulative distribution function):

F(x) = P(X<=x)
Function describing Probability that discrete random variable is less than or equal to some value

</td><td>

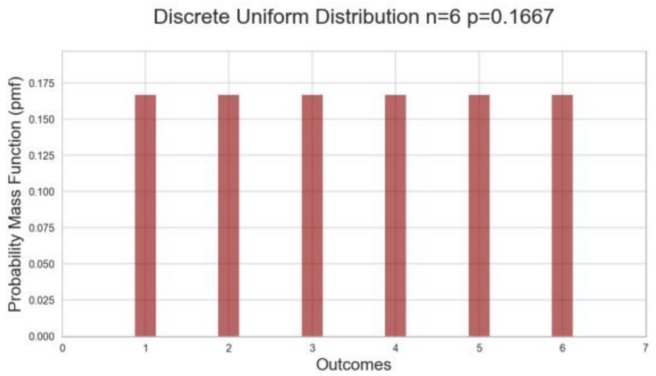
pdf (probability density function):

Function describing relative likelihood that continuous random variable would equal to some value
Since the outcomes are uncountable, we cannot calculate probability for a single outcome, but we estimate for a “range” of values

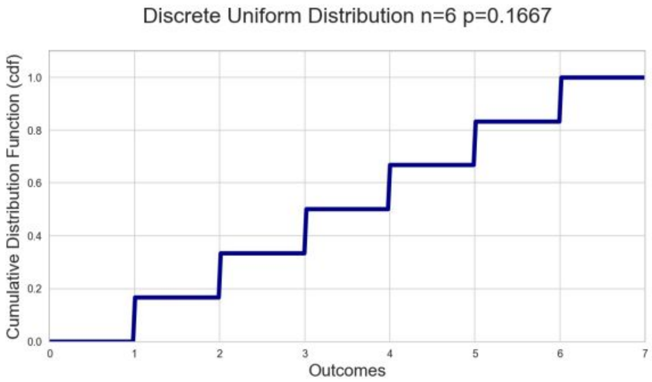
</td></tr>
</table>

Discrete Probability Distributions: PMF vs CDF

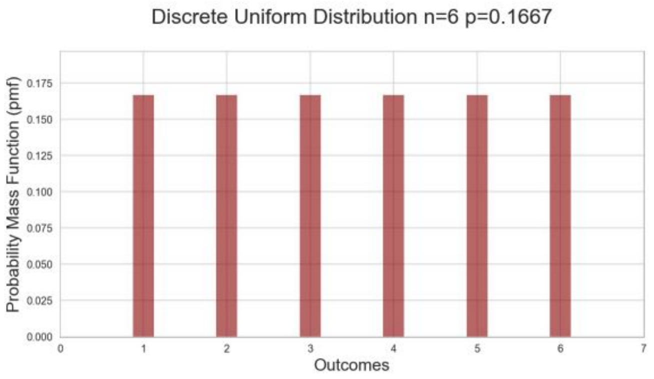
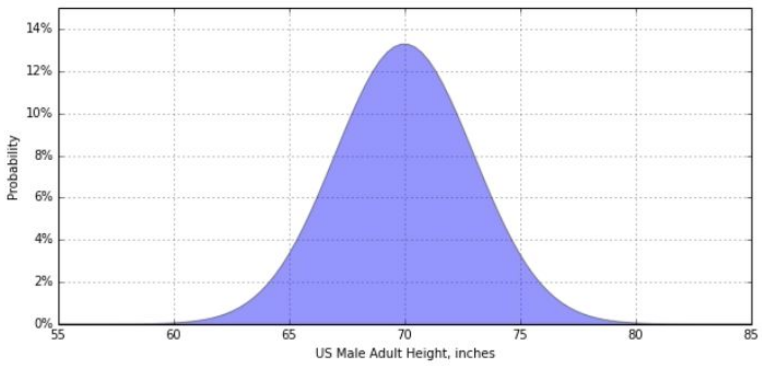
PMF



CDF



Discrete vs Continuous Probability distributions

Discrete probability distributions (PMF)	Continuous probability distributions (PDF)
	

Discrete Distributions: Types

	Uniform	Bernoulli	Binomial	Poisson
definition	We have a discrete set of outcomes and each outcome is **equally likely**	We only have 2 outcomes (**binary**), with a **probability of success**	Extension of Bernoulli - caters for **multiple experiments**, each with a probability of success and **binary** outcomes	Models **count** data (non-decimals, non-negative)
example	Outcomes from a fair die roll	Binary outcome from a coin flip	Flipping a coin multiple times, each with binary outcome	No. of student enrolments for a given cohort
parameters	Probability of each equally likely outcome, “p”	Probability of success, “p” of one of the 2 binary outcomes	1. Probability of success, “p” 2. No. of trials, “n”	rate of event occurring λ (mean number of events)

Continuous Distributions: Types

	Uniform	Exponential	Gamma	Normal	Beta
definition	All continuous values have the **same probability density**	This distribution is used to model **amount of time** until an event	Extension of exponential to **model amount of time** until **multiple events**	The distribution that MOST **real-word processes** are modeled with	Perfect for modeling **probabilities**. Only takes **values between 0 - 1**
example	Food delivered between 20-30 minutes	Amount of time until lesson starts	If one light bulb lasts one year, Gamma can model time taken to replace 5 bulbs	Examination scores	Probability of anything happening
parameters	1. Min value of distribution, “a” 2. Max value of distribution, “b” *(visual in next slide)*	β, average time to an event	1. β average time to an event 2. α (shape) exponential distributions	1. Mean, μ 2. Standard deviation, σ	α, β (shape) parameters

Continuous Uniform Distribution

Examples:

- I am thinking of a number between 1 and 10.
- My food will arrive between 25 - 35 minutes.