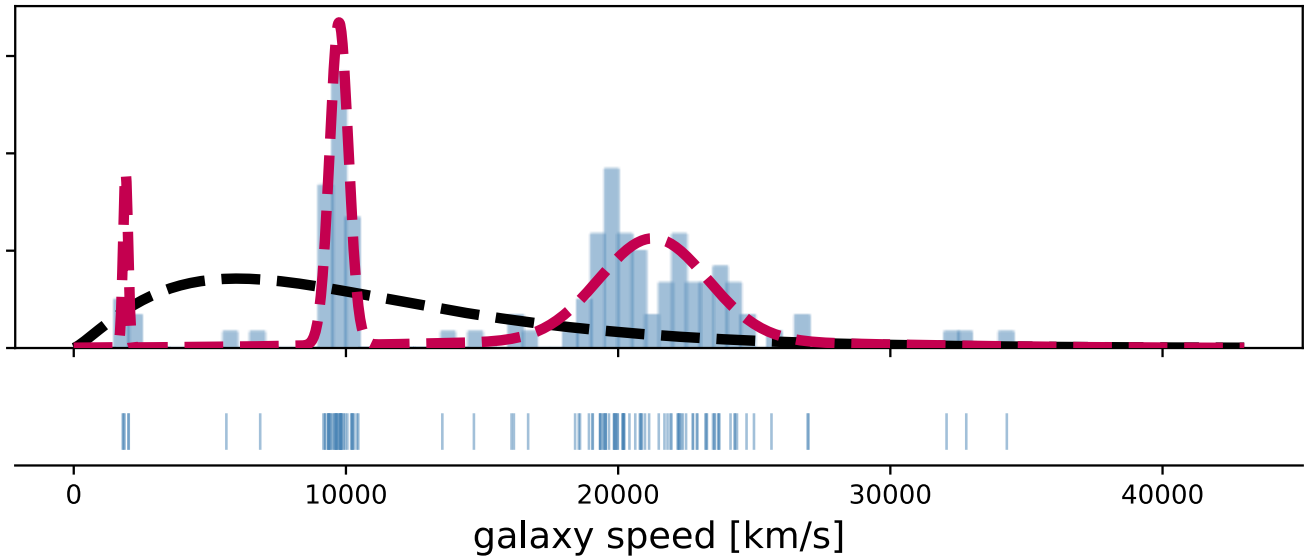
PART II: HANDLING PROBABILITY MODELS

Expanding our model-building repertoire



This chart shows the distribution of the speeds of 120 galaxies, from a survey of the Corona Borealis region.
Postman, Huchra, Geller (1986)

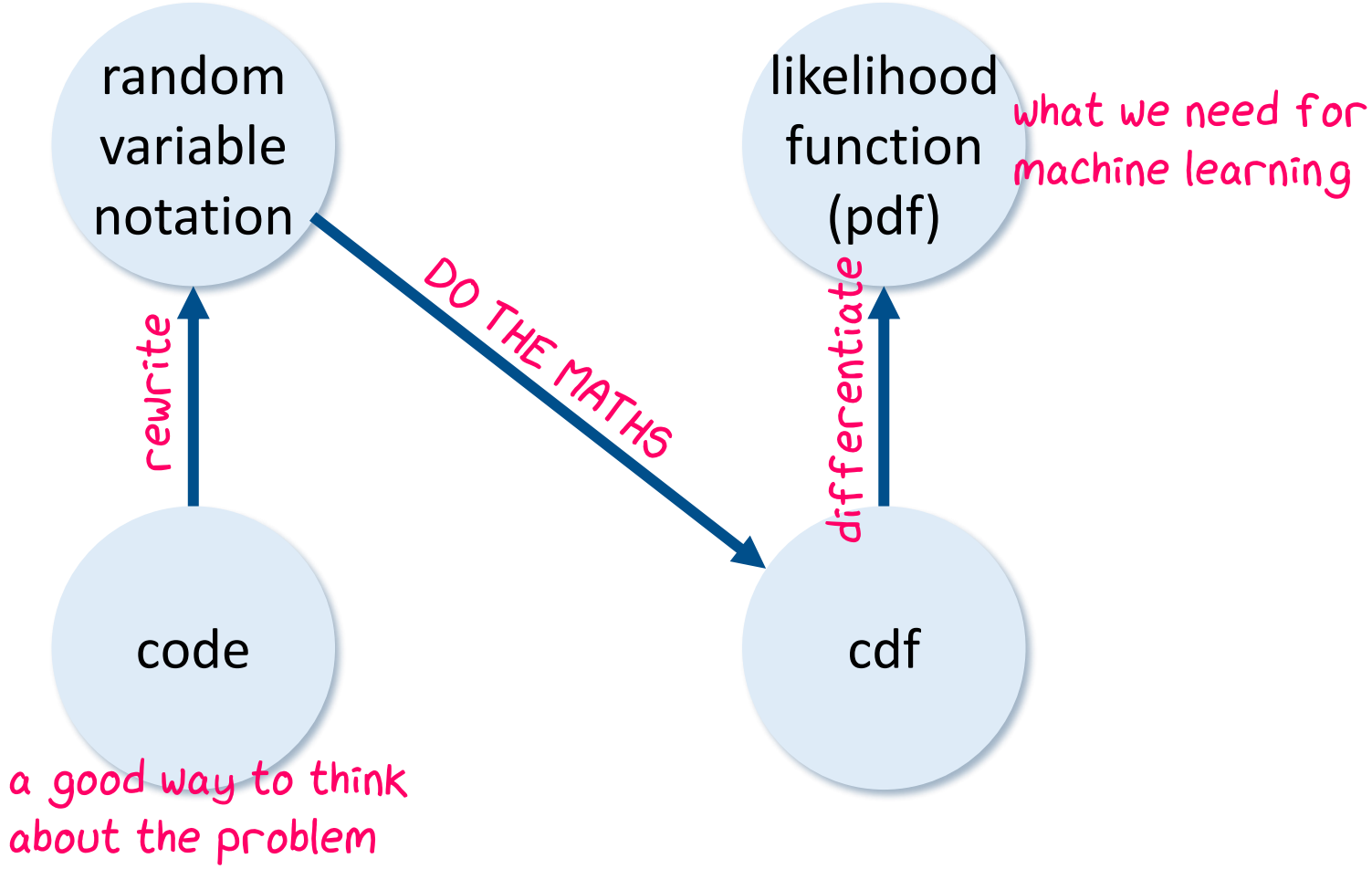
theoretical distribution for uniform universe

fitted Gaussian mixture model
(a mixture of three Gaussians)

Can we come up with a better model for this dataset?

Bespoke probability distributions part I: from code to likelihood (for continuous random variables)

There are four ways to specify a distribution:



Exercise 5.3.2

Find the pdf of the random variable generated by this code:

```
u = np.random.uniform()
x = - np.log(u) / λ          # λ>0
```

STRATEGY

- Try to write our probability in terms of simple standard random variable *for which we can look up the cdf*
- Break it down so that the random variables are on the left *so we can use the textbook cdf*

Step 1: random variable notation

$U \sim \text{Uniform}[0,1]$

$X = -\frac{1}{\lambda} \log U$

Step 2: X is a continuous r.v., so find its cdf

$$\text{cdf for } X = \mathbb{P}(X \le x) = \mathbb{P}\left(-\tfrac{1}{\lambda} \log U \le x\right) \text{ by defn of } X$$

$$= \mathbb{P}\left(U \ge e^{-\lambda x}\right) \text{ by algebra}$$

$$= 1 - \mathbb{P}\left(U < e^{-\lambda x}\right)$$

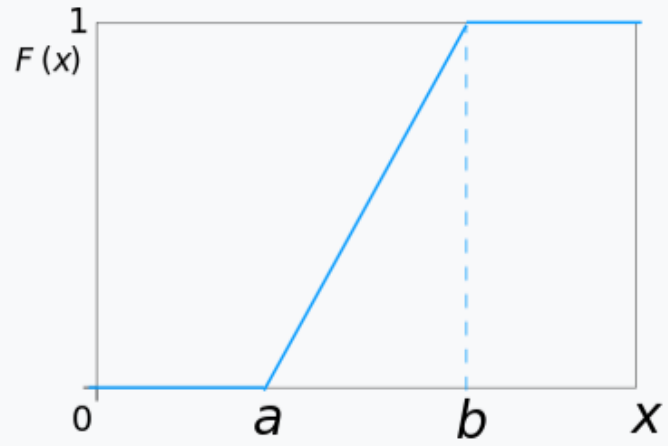
$$= 1 - e^{-\lambda x} \quad \text{assuming } x \ge 0, \text{ given } \lambda > 0$$

Step 3: differentiate cdf to get pdf

$$\frac{d}{dx}\left(1 - e^{-\lambda x}\right) = \lambda e^{-\lambda x}$$

Wikipedia: Uniform distribution

Cumulative distribution function



Notation	$\mathcal{U}_{[a,b]}$
Parameters	$-\infty < a < b < \infty$
Support	$x \in [a, b]$
PDF	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
CDF	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$

$$P(u[0,1] < e^{-\lambda x}) = \begin{cases} 0 & \text{for } e^{-\lambda x} < 0 \\ \frac{e^{-\lambda x} - 0}{1 - 0} & \text{for } 0 \le e^{-\lambda x} \le 1 \\ 1 & \text{for } e^{-\lambda x} > 1 \end{cases}$$

Exercise 5.3.5 (Gaussian mixture model)

Find the likelihood function $\Pr_X$ for the Gaussian mixture model

$$K \sim \text{Cat}([p_1, \ldots, p_m])$$
$$X \sim \text{Normal}(\mu_K, \sigma_K^2)$$

STRATEGY

- The Law of Total Probability is useful for sequentially-generated probability models
  since we're given both terms in the equation

Step 1: random variable notation — The question tells us

Step 2: X is a continuous r.v., so find its cdf

$$\mathbb{P}(X \le x) = \sum_k \mathbb{P}(X \le x \mid K = k)\, \mathbb{P}(K = k) \quad \text{by Law of Total Probability}$$

$$= \sum_k \Phi_{\mu_k, \sigma_k}(x) \cdot p_k \quad \text{where } \Phi_{\mu,\sigma} \text{ is cdf for } N(\mu, \sigma^2)$$

Step 3: differentiate cdf to get pdf

$$\frac{d}{dx}\mathbb{P}(X \le x) = \sum_k p_k \frac{d}{dx} \Phi_{\mu_k, \sigma_k}(x)$$

$$= \sum_k p_k \varphi_{\mu_k, \sigma_k}(x) \quad \text{where } \varphi_{\mu,\sigma} \text{ is pdf for } N(\mu, \sigma^2)$$