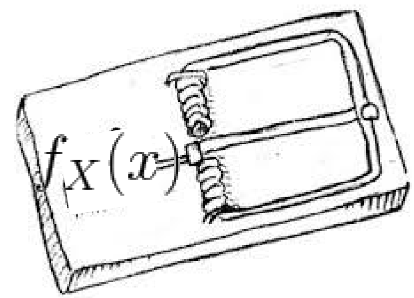
CRV Review: 2-4

- Concept of PDF
- Formal definition of a pdf
- How to create a continuous random variable in python
- Plot Histograms
- Plot PDFs

Continuous Random Variables

Common Trap



- $f_X(x)$ does not yield a probability
  - $\int_a^b f_X(x)dx$ does
  - x may be anything ($\mathbb{R}$)
    - thus, $f_X(x)$ may be > 1

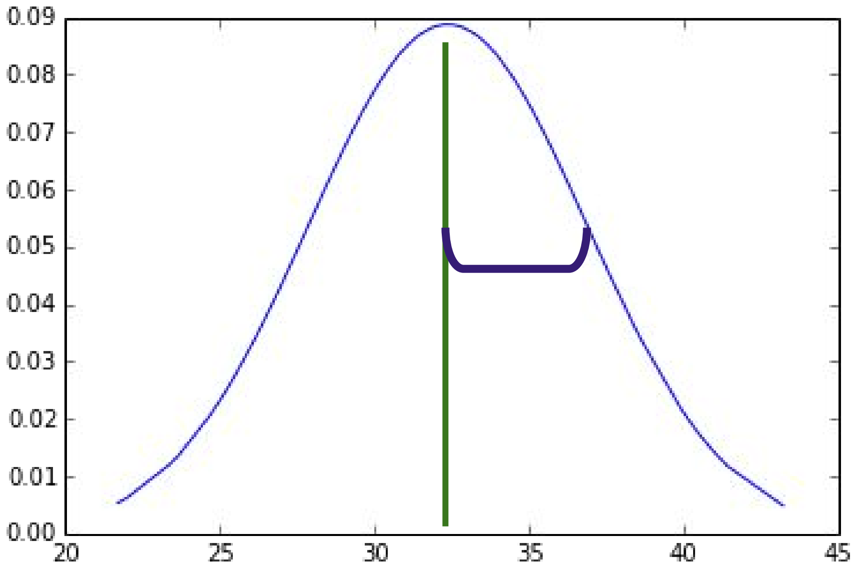
Continuous Random Variables

Some Common Probability Density Functions

Continuous Random Variables

Common *pdf*s: Normal(μ, σ^2)

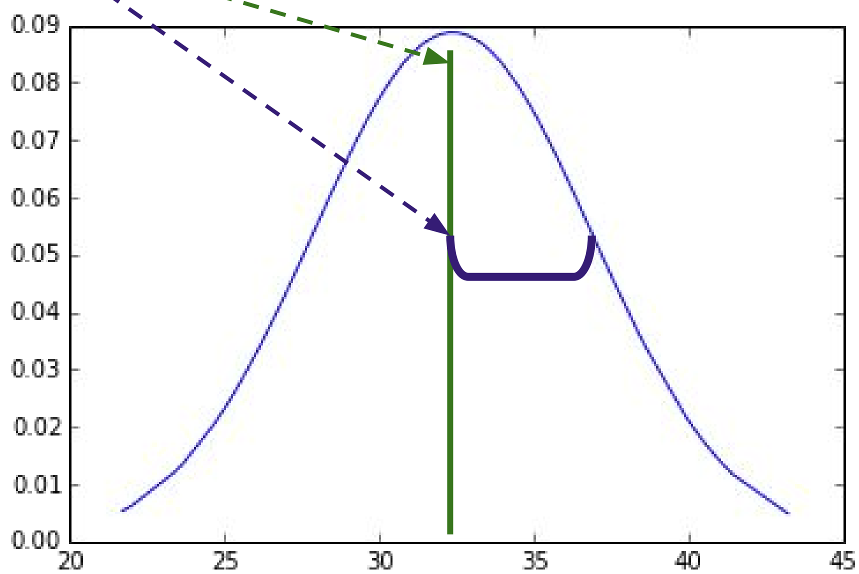
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Continuous Random Variables

Common *pdf*s: Normal(μ, σ^2)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



μ: mean (or "center")

= expectation

σ^2: variance,

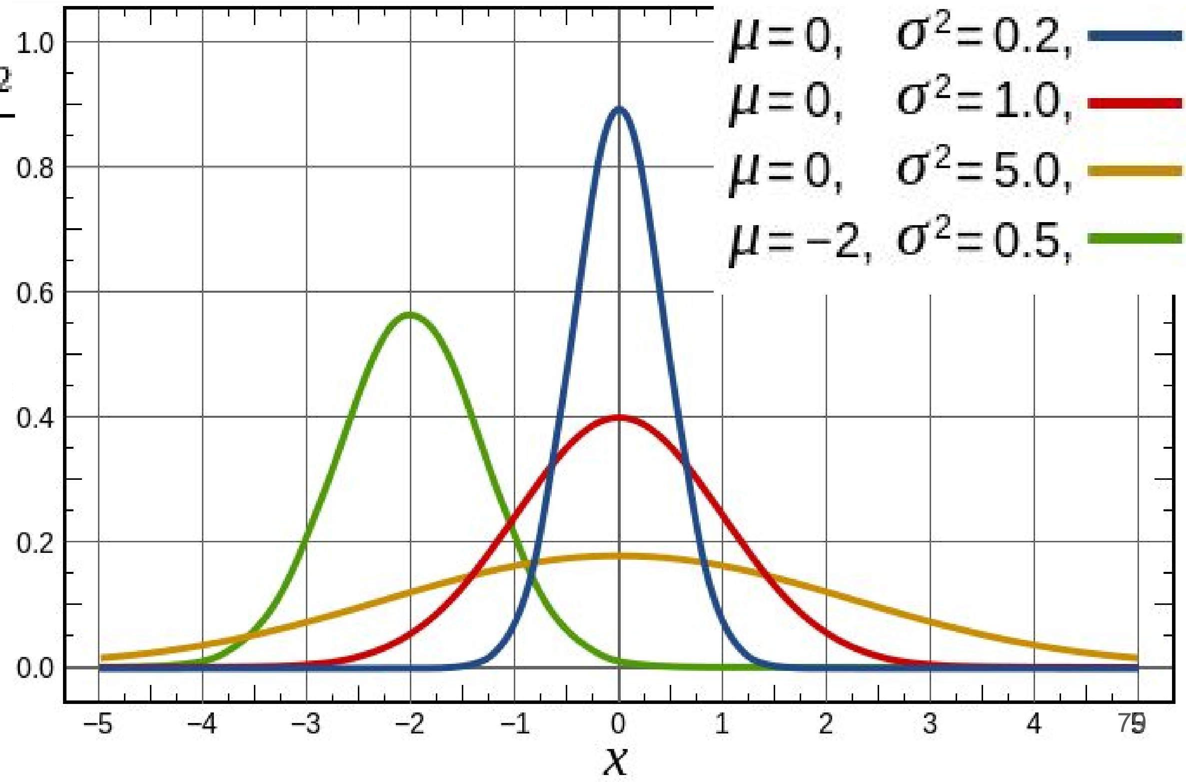
σ: standard deviation

Continuous Random Variables

Common *pdf*s: Normal(μ, σ^2)

Credit: Wikipedia

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



μ: mean (or "center")

= expectation

σ^2: variance,

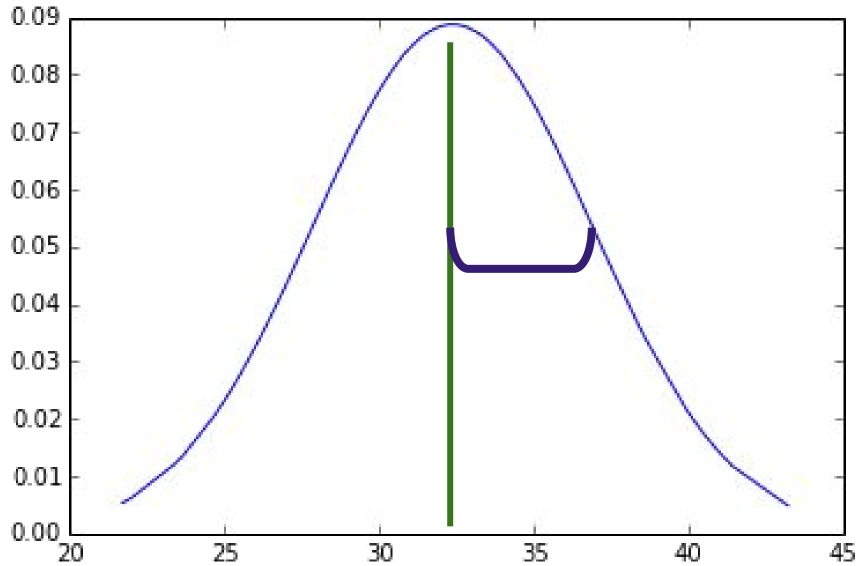
σ: standard deviation

Continuous Random Variables

Common *pdf*s: $\text{Normal}(\mu, \sigma^2)$

$X \sim \text{Normal}(\mu, \sigma^2)$, examples:

- height
- intelligence/ability
- **measurement error**
- averages (or sum) of lots of random variables



Continuous Random Variables

Common *pdfs*: Normal(0, 1) ("standard normal")

How to "standardize" any normal distribution:

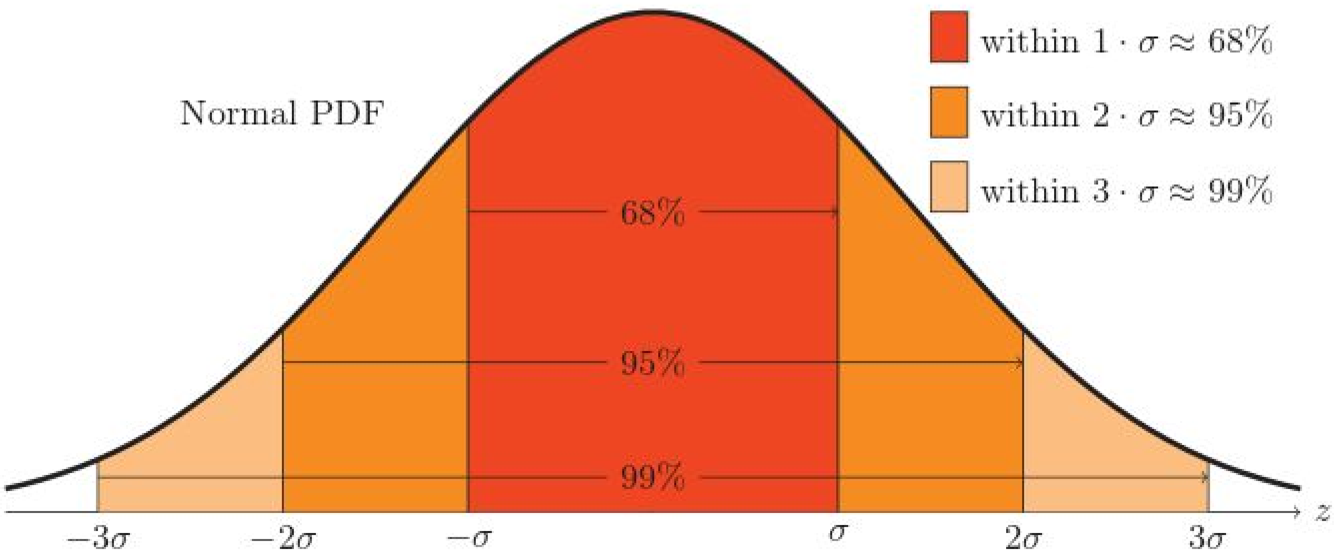
- subtract the mean, μ (aka "mean centering")
- divide by the standard deviation, σ

$z = (x - \mu) / \sigma$, (aka "z score")

Continuous Random Variables

Common *pdf*s: $\mathrm{Normal}(0, 1)$

$$P(-1 \le Z \le 1) \approx .68, \quad P(-2 \le Z \le 2) \approx .95, \quad P(-3 \le Z \le 3) \approx .99$$

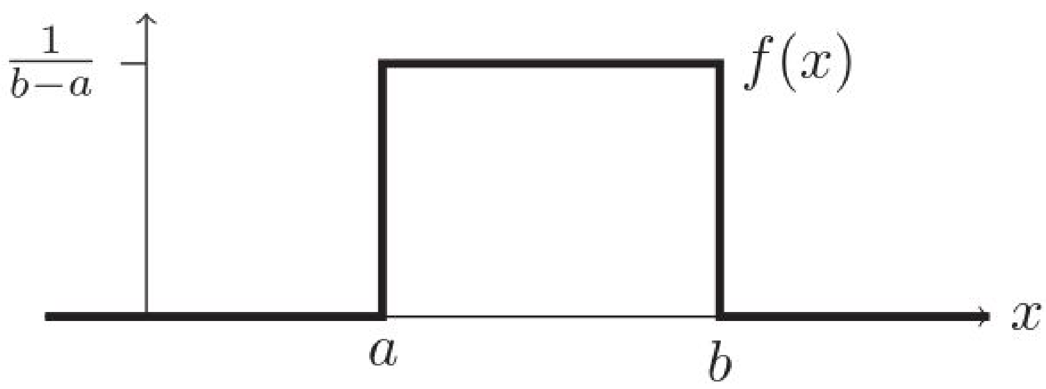


Credit: MIT Open Courseware: Probability and Statistics

Continuous Random Variables

Common *pdf*s: Uniform(a, b)

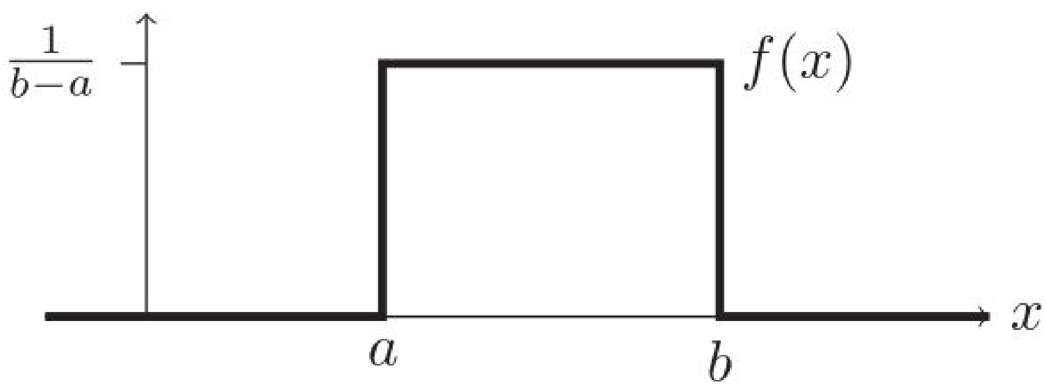
$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$



Continuous Random Variables

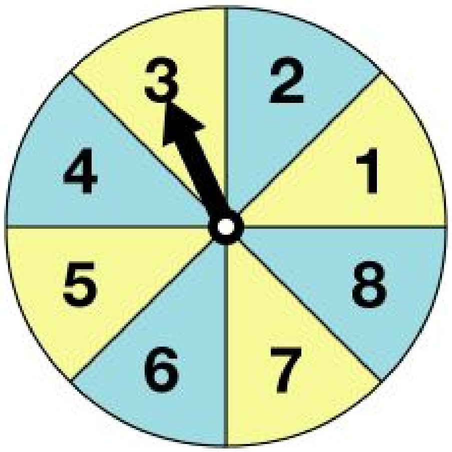
Common *pdf*s: $\text{Uniform}(a, b)$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$



$X \sim \text{Uniform}(a, b)$, examples:

- spinner in a game
- random number generator
- analog to digital rounding error

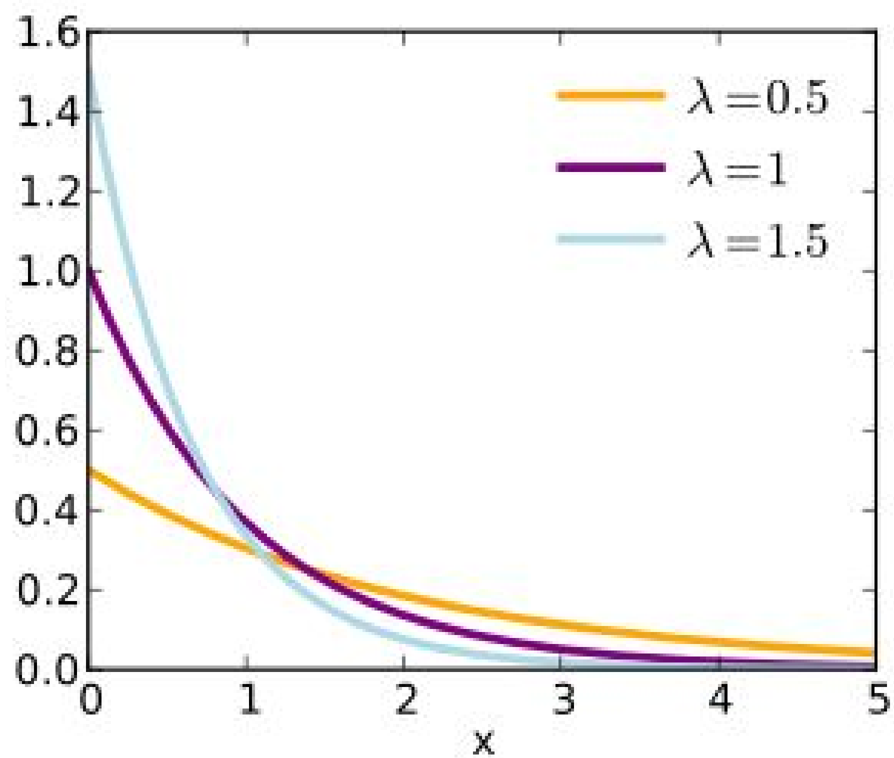


Continuous Random Variables

Common *pdf*s: Exponential(λ)

Credit: Wikipedia

$$f_X(x) = \lambda e^{-\lambda x}, x > 0$$



λ: rate or inverse scale

β: scale ($\lambda = \frac{1}{\beta}$)

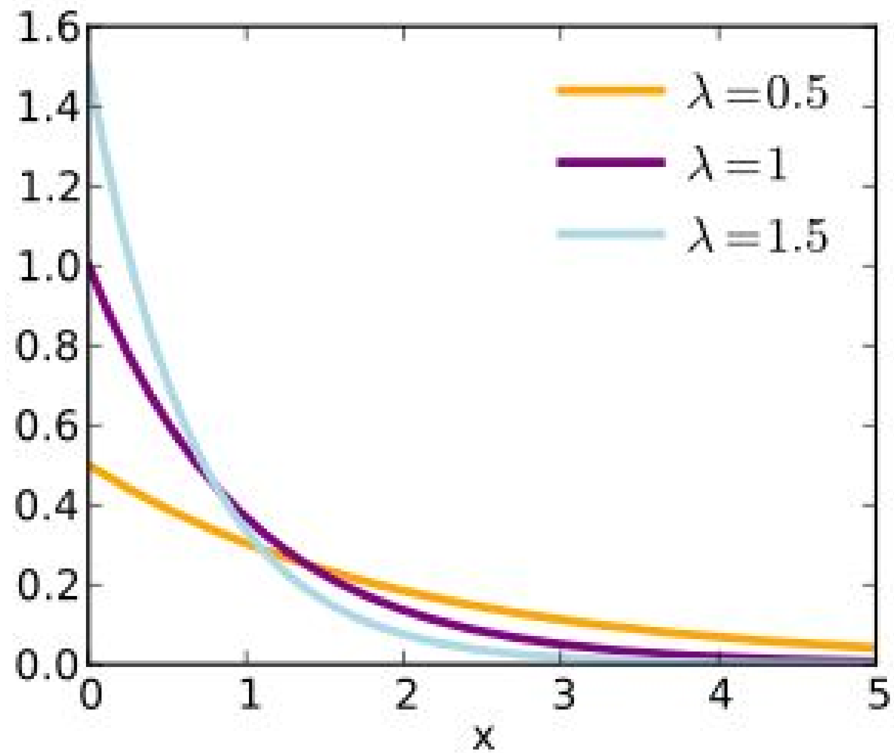
Continuous Random Variables

Common *pdf*s: Exponential(λ)

Credit: Wikipedia

$X \sim \text{Exp}(\lambda)$, examples:

- lifetime of electronics
- waiting times between rare events (e.g. waiting for a taxi)
- recurrence of words across documents



Continuous Random Variables

How to decide which pdf is best for my data?

Look at a *non-parametric* curve estimate:
(If you have lots of data)

- Histogram
- Kernel Density Estimator

Continuous Random Variables

How to decide which pdf is best for my data?

Look at a *non-parametric* curve estimate:
(If you have lots of data)

- Histogram
- **Kernel Density Estimator**

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

K: kernel function, *h:* bandwidth

(for every data point, draw *K* and add to density)

Continuous Random Variables

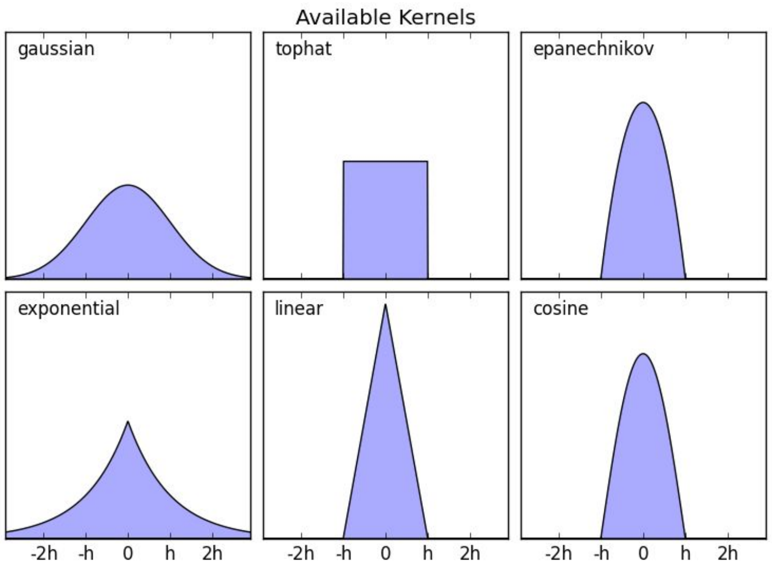
How to decide which pdf is best for my data?

Look at a *non-parametric* curve estimate:
(If you have lots of data)

- Histogram
- **Kernel Density Estimator**

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

K: kernel function, *h:* bandwidth



(for every data point, draw *K* and add to density)