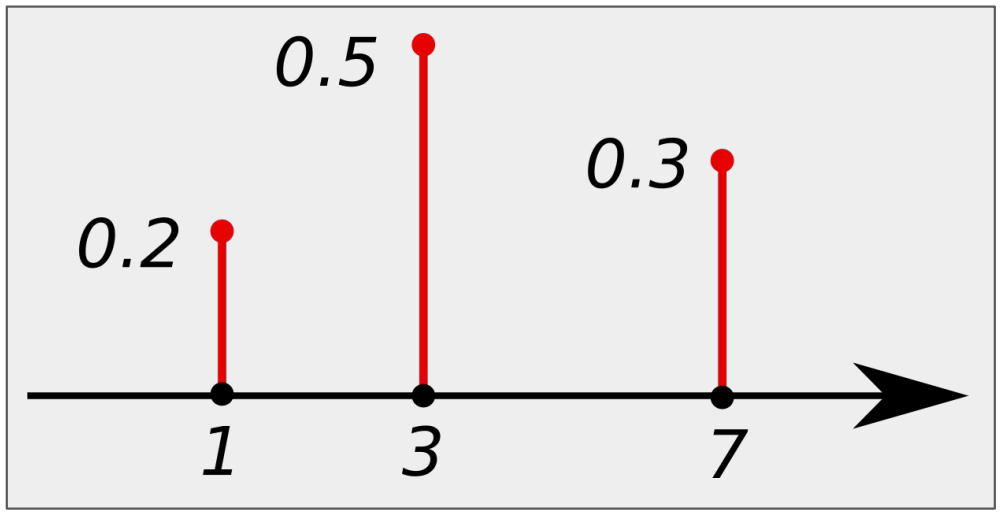
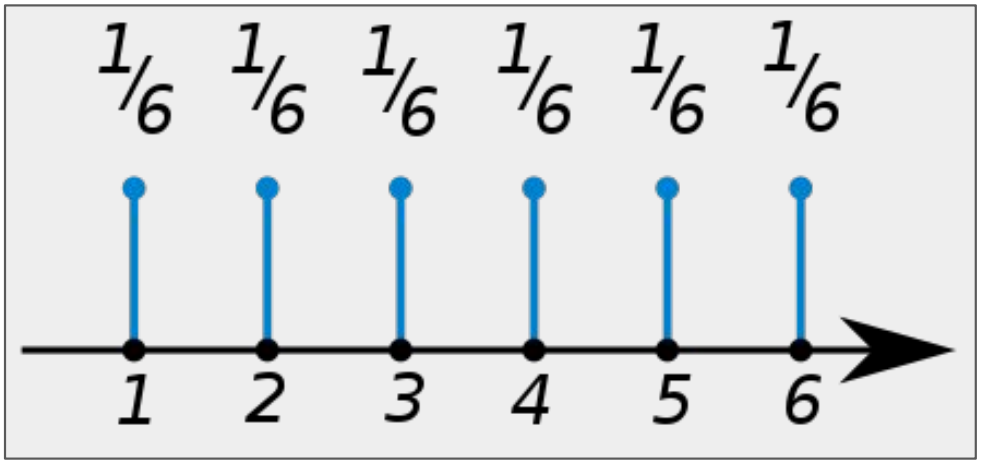
Tour de Distributions!

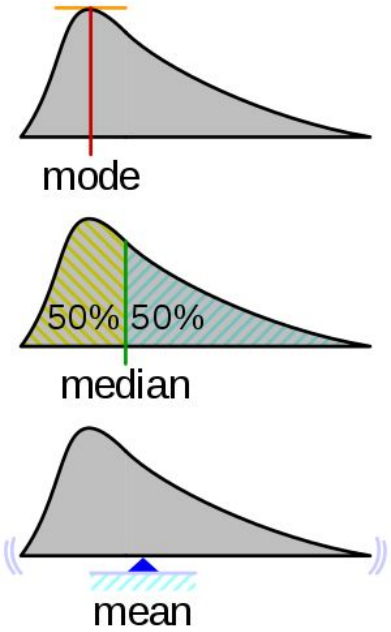
DC-DS-0715

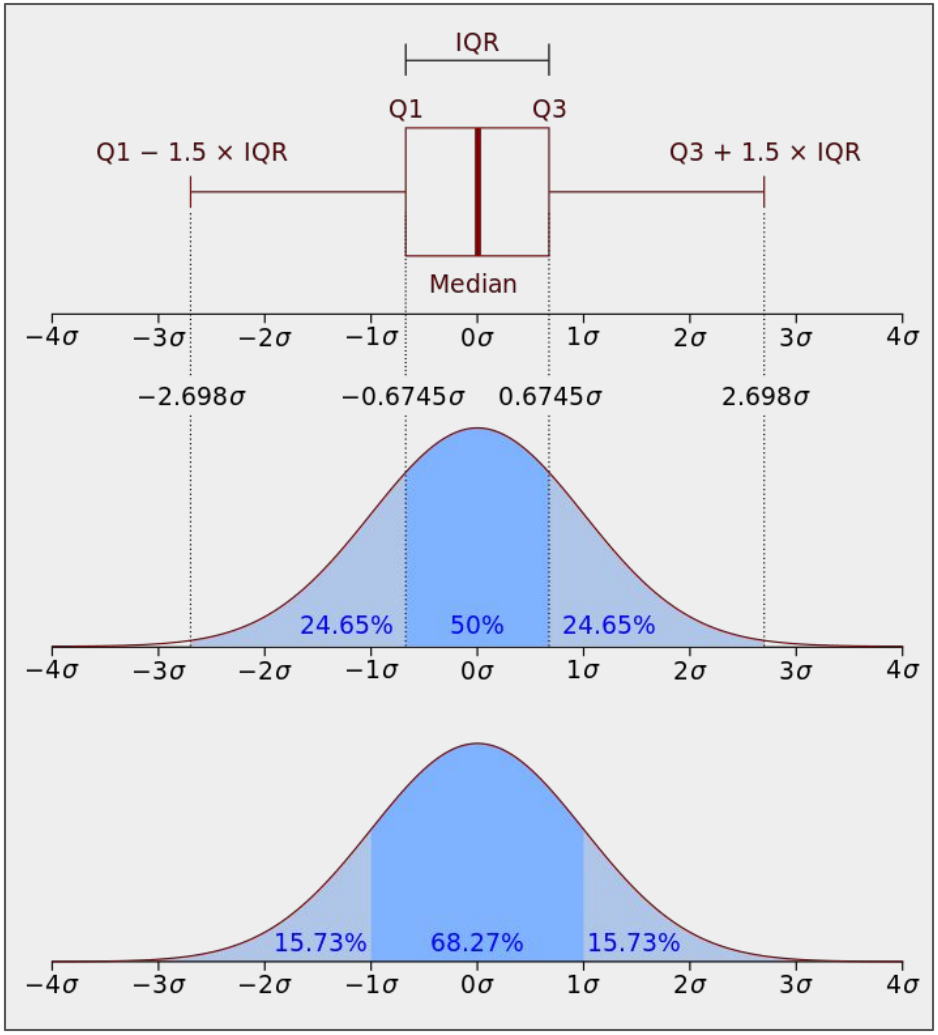
Probability Mass Function



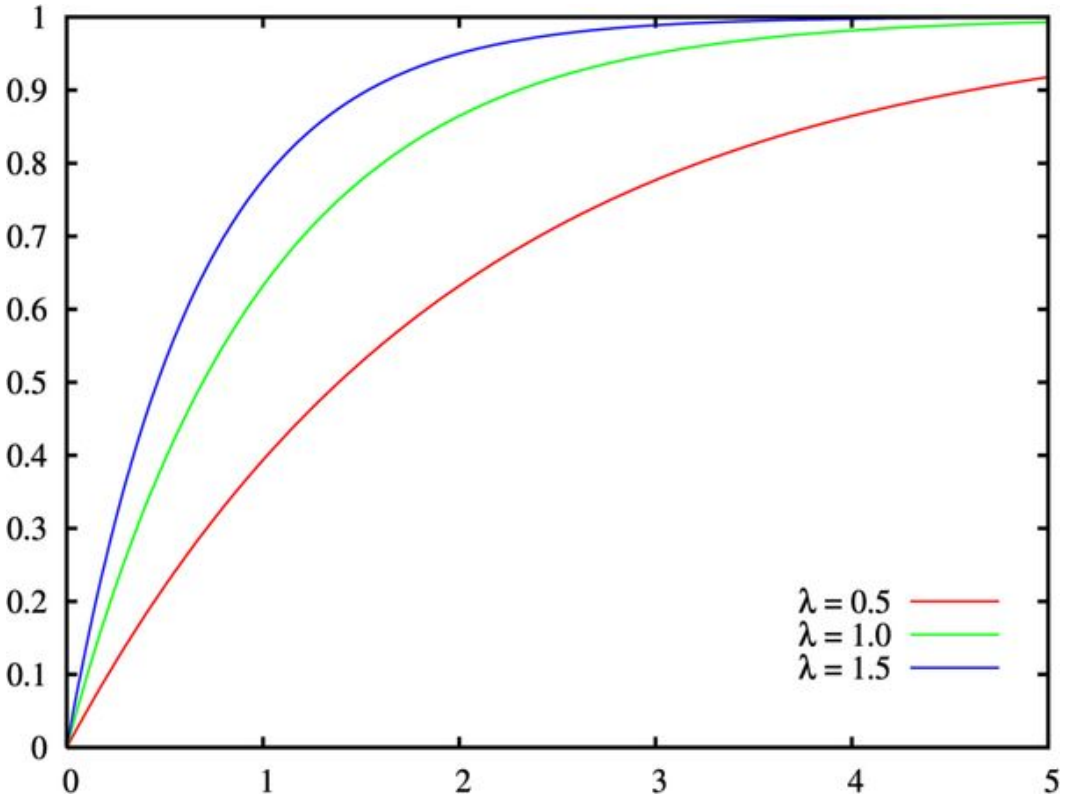


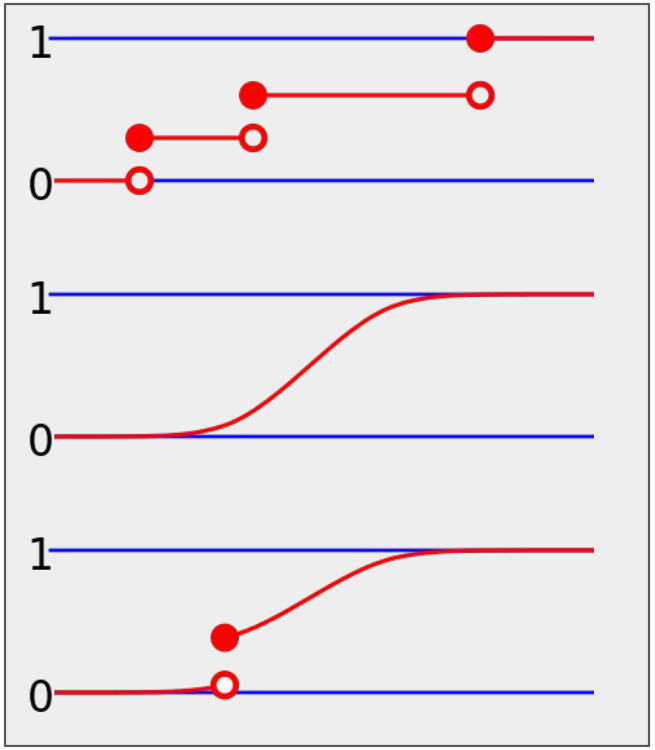
Probability Density Function



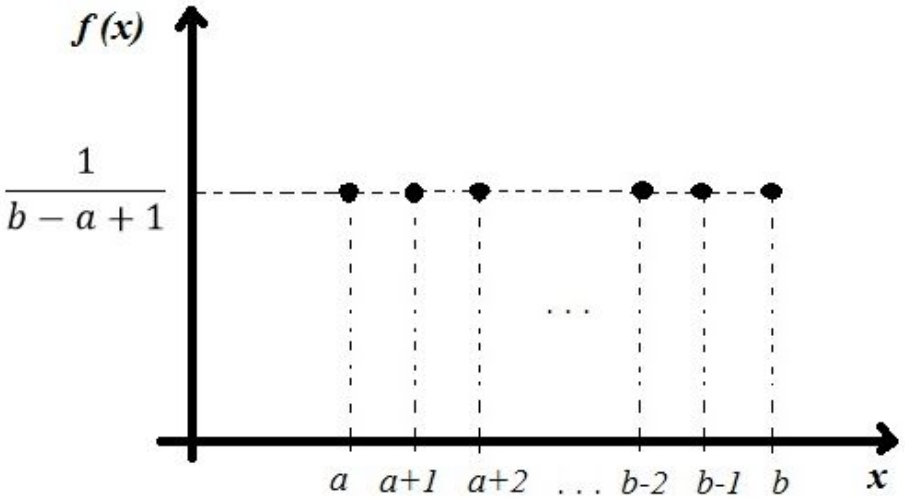


Cumulative Density Function

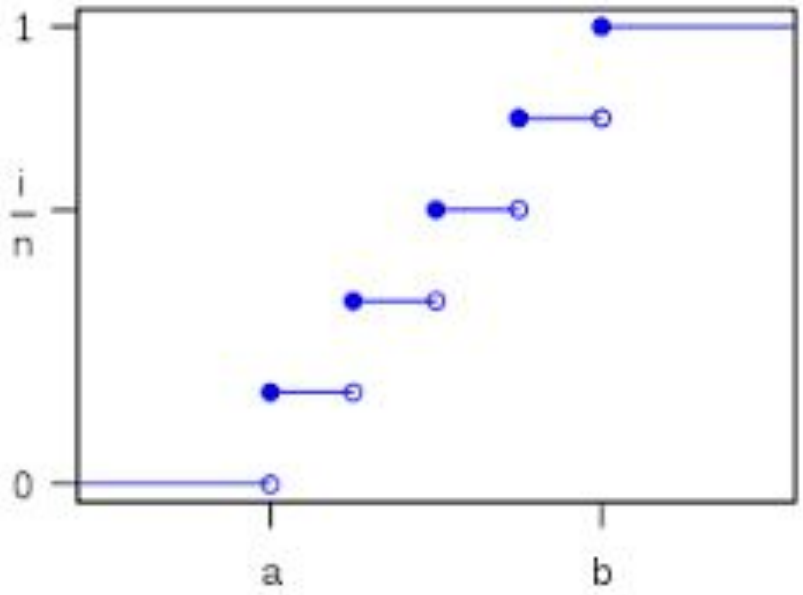




Uniform: Discrete

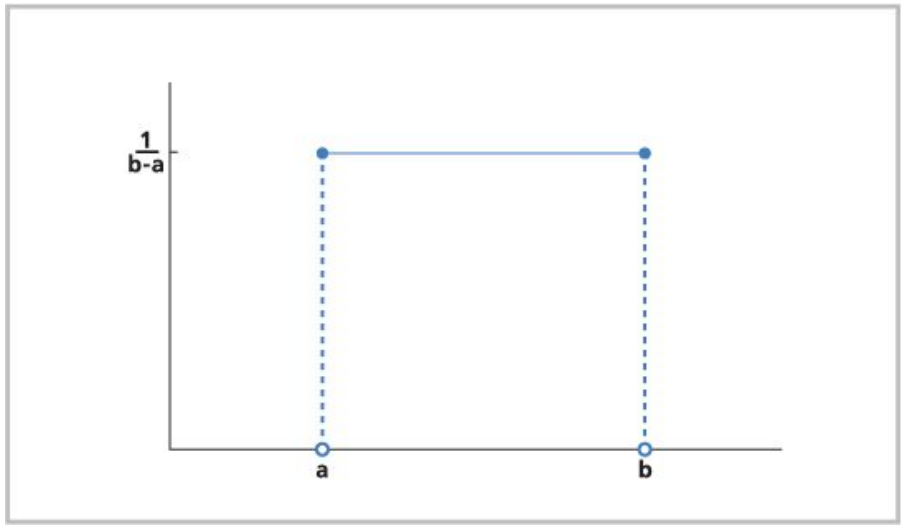


Probability Mass Function

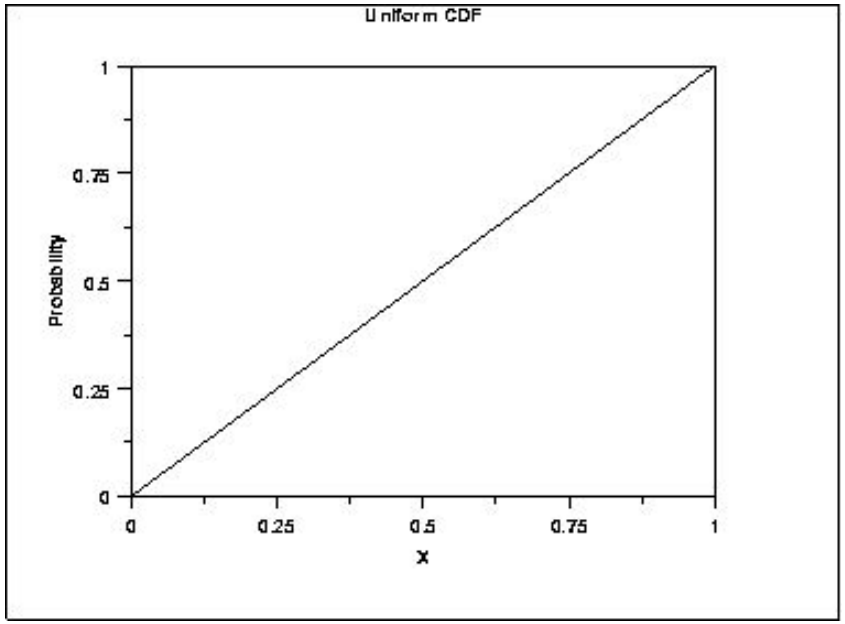


CDF

Uniform: Continuous



Probability Distribution Function



CDF

Uniform

- Probabilities for all outcomes are the same
- $$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$
- $$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \le x \le b \\ 1 & \text{for } x > b \end{cases}$$

$$\mu = \frac{1}{2}(a+b)$$
$$\sigma^2 = \frac{1}{12}(b-a)^2$$

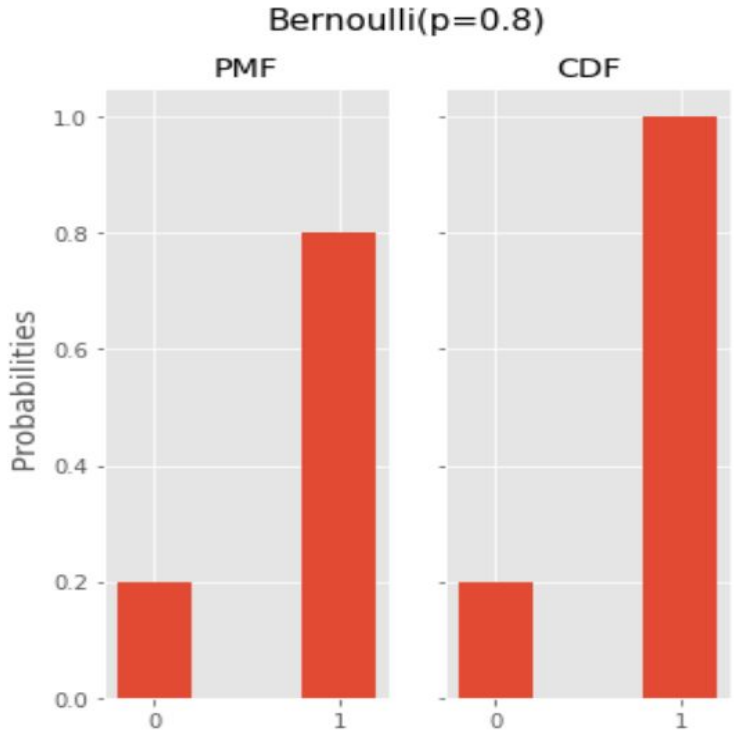
Bernoulli and Binomial Distributions

- Key parameter, p, probability of success
- Examples: coin flip, rain/shine, win/lose
- Relates to:

  probit model (normal distribution)

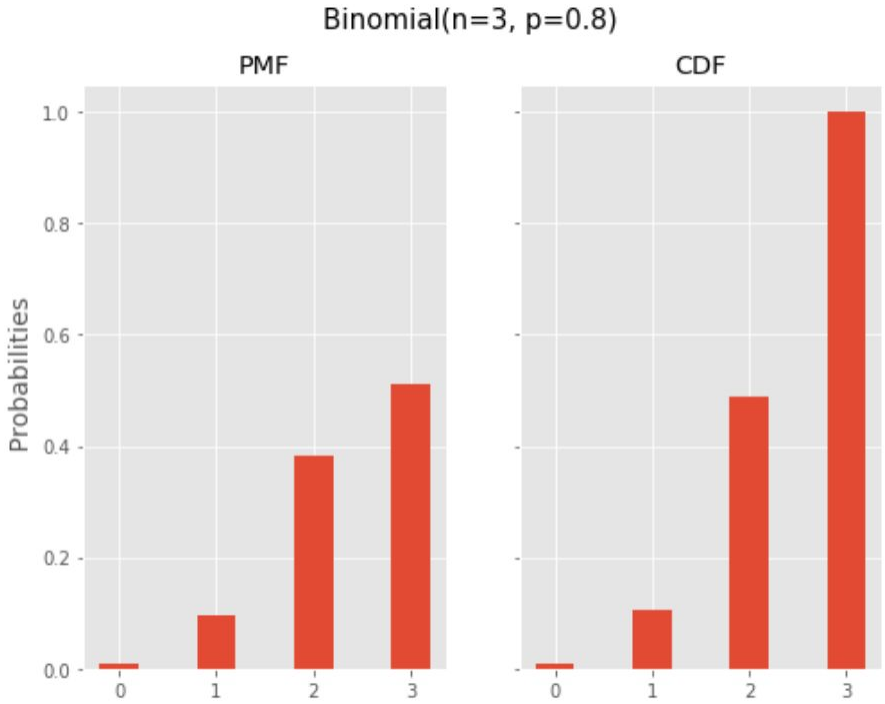
  logistic model (extreme value distribution)

  - 0,1 choice/outcome
  - P(1{cancer})= f(exposure to chemicals, diet, genetics)
  - P(1{default})=g(income, education, credit history)



Bernoulli and Binomial Distributions

- Key parameter: (p) probability of success, (n) number of trial, (k) number of successes



Bernoulli and Binomial Distributions:

$$(-p+1+p)^n = \sum_{k<n} \binom{n}{k} p^k (1-p)^{n-k}$$

$$(x+y)^n = \sum^{k<n} \binom{n}{k} x^k y^{n-k}$$

PDF: $P(p,n,k) = \binom{n}{k} p^k (1-p)^{n-k}$

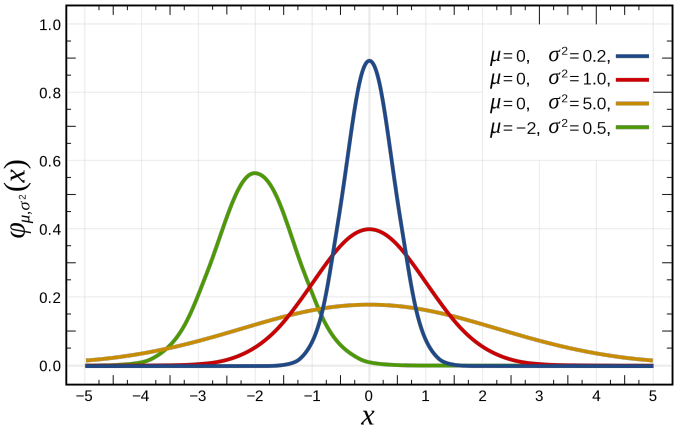
Mean: np

Var: np(1-p)

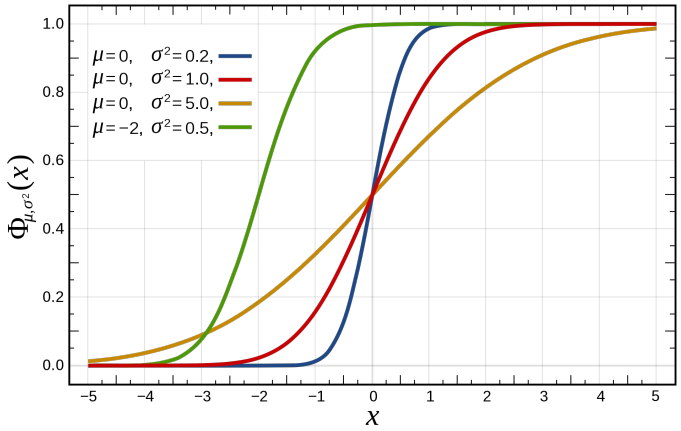
CDF: $C(p,n,x) = \sum_{k<x} \binom{n}{k} p^k (1-p)^{n-k}$

Normal/Gaussian and Standard Normal Distributions

Normal standard distribution: mean = 0, std = 1 (often result of normalized data)



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

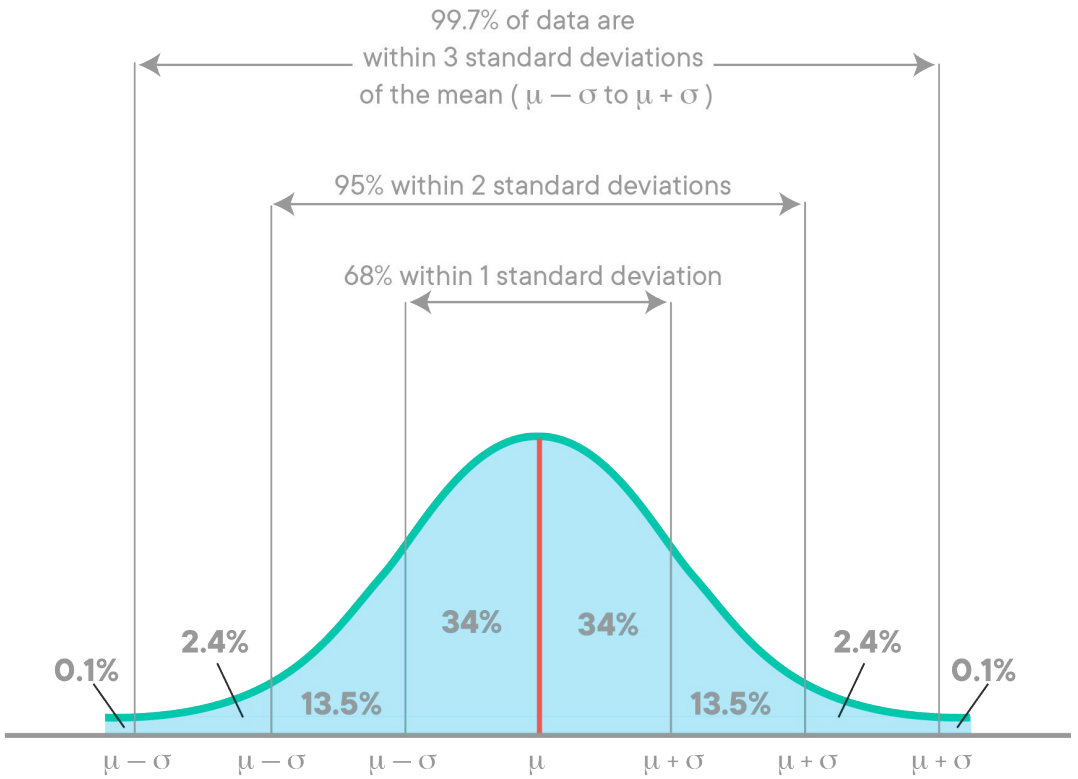


$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

$$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

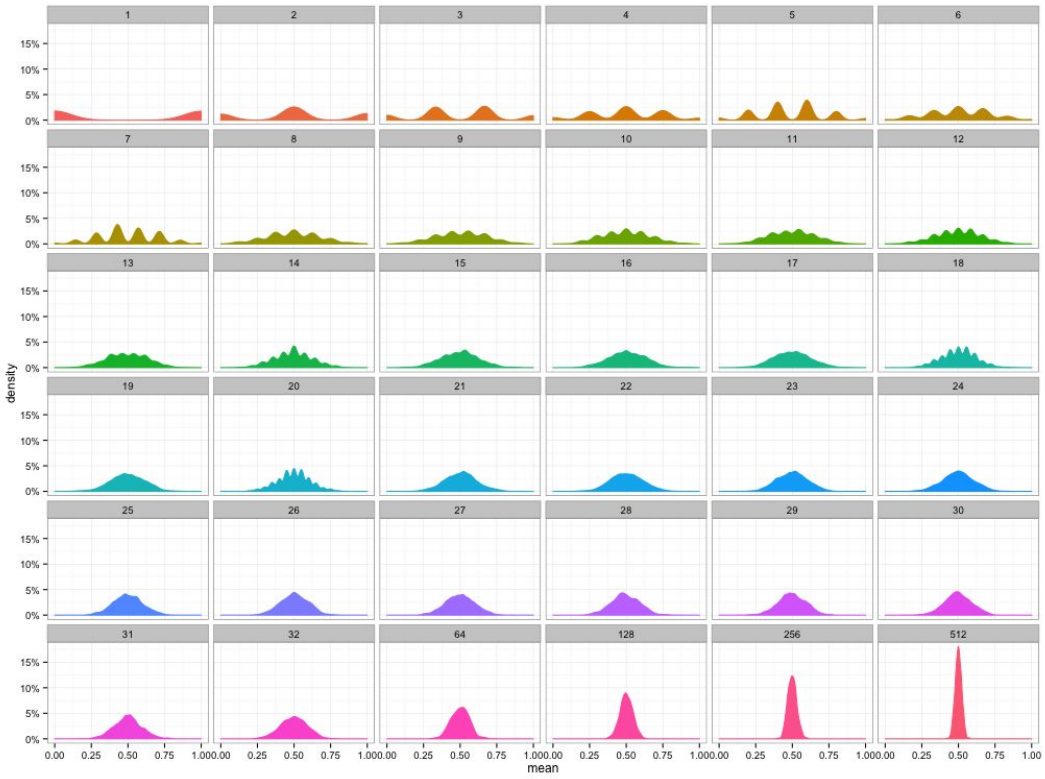
Normal Distribution

- Mean, median, and mode are equal
- Area under bell curve equals 1



Central Limit Theorem

When you add a large number of **independent, random** variables, their sum will tend towards a Gaussian (Normal) distribution.



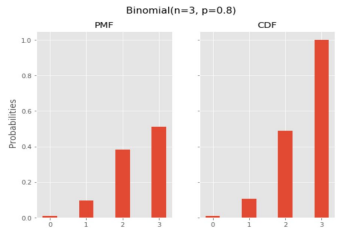
Poisson Distribution

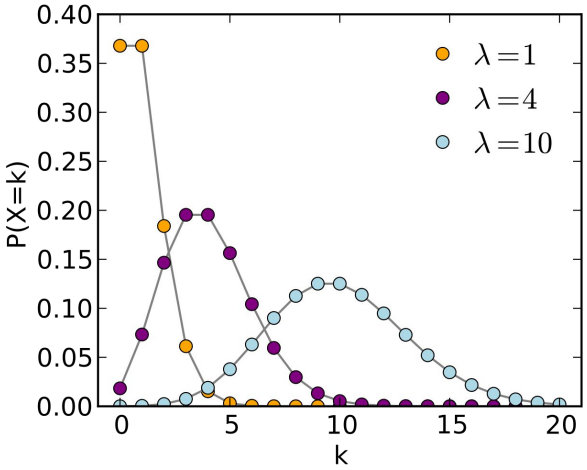
- event can be counted in **whole numbers**
- occurrences are **independent**
- average frequency of occurrence for the time period in question is **known**

https://www.umass.edu/wsp/resources/poisson/

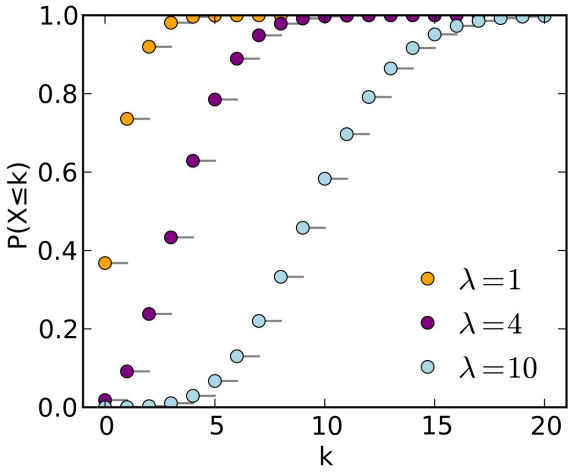
- the poisson distribution counts discrete occurrences among a **continuous domain.**
- *compare:* the binomial distribution counts discrete occurrences among **discrete trials**.

CogitoErgoCogitoSum (https://math.stackexchange.com/users/52938/cogitoergocogitosum), Difference between Poisson and Binomial distributions., URL (version: 2014-12-06): https://math.stackexchange.com/q/1050237

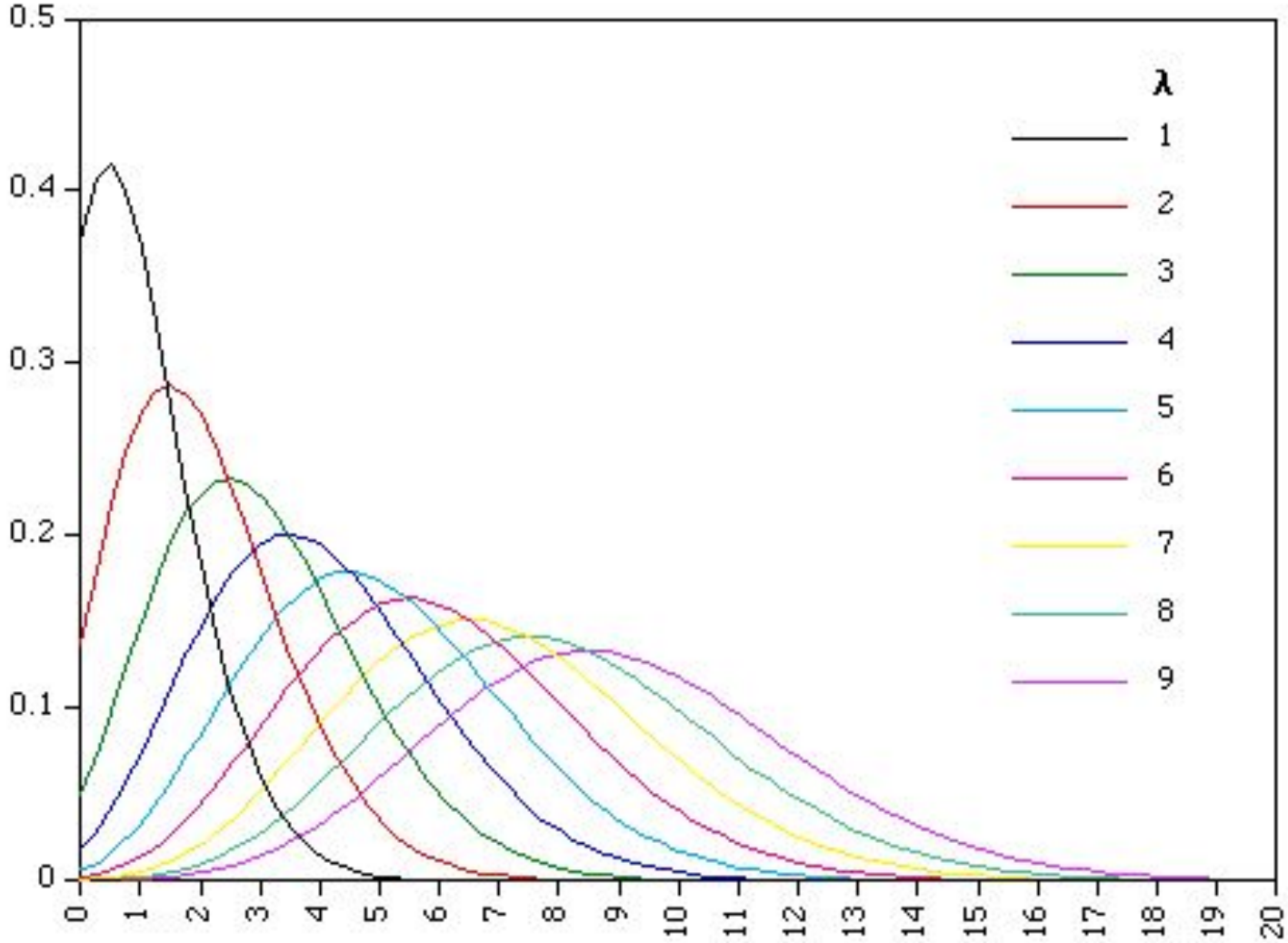




$$\frac{\lambda^k e^{-\lambda}}{k!}$$



$$\sum_{i=1}^{k} \frac{\lambda^k e^{-\lambda}}{k!}$$

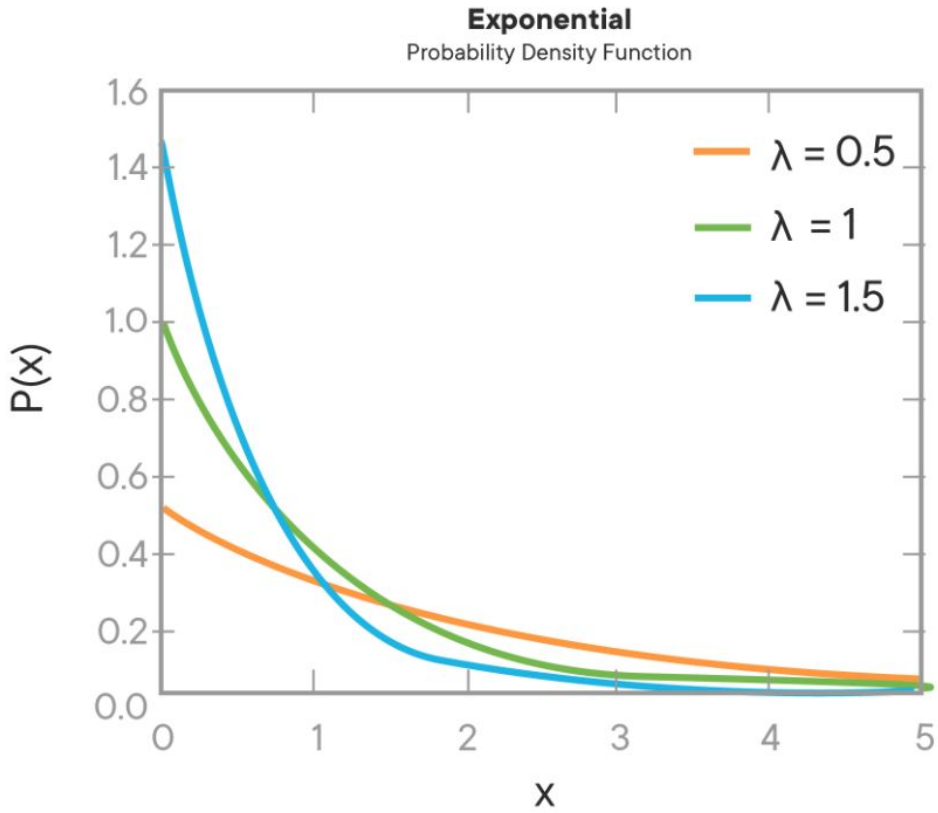

0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
λ
1
2
3
4
5
6
7
8
9

Geometric and Exponential Distributions

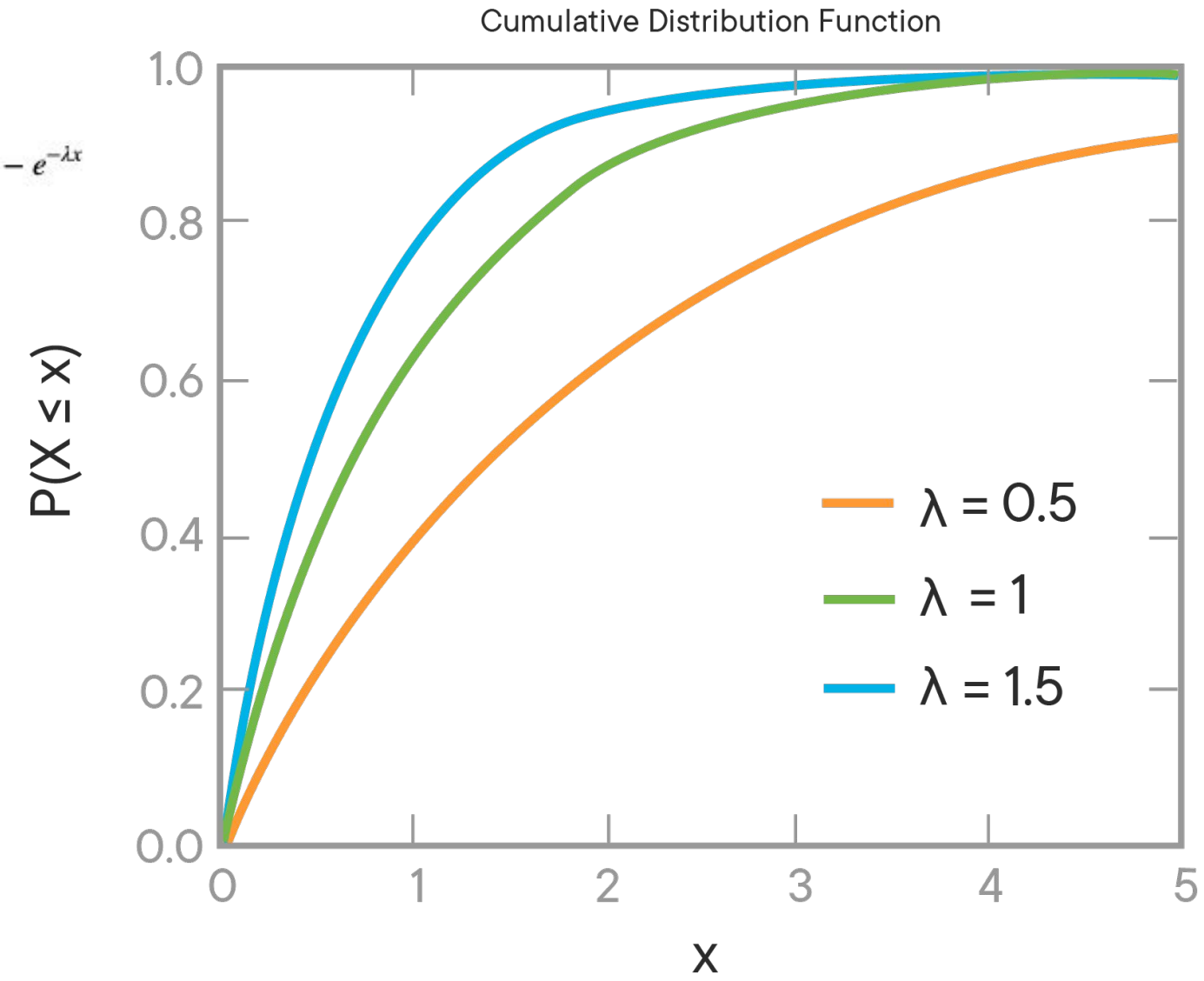
$$\lambda = \frac{1}{\mu}$$

$$PDF(x) = \lambda e^{-\lambda x}$$

The **Exponential Distribution** describes the probability distribution of the amount of time it takes before an event occurs. In a way, it solves the inverse of the problem solved by the Poisson Distribution



$$CDF(x) = 1 - e^{-\lambda x}$$



Fun fact:

$\sigma = \mu$

Guess the Distribution- 1

A doctor often recommends that patients of his take allergy medication to help alleviate their symptoms. Some of his patients report having relief with medication while others do not.

Guess the Distribution - 2

A nutritionist is interested in examining the body mass index (BMI) in a population of 60 year old males. She collects data on 100 60 year old males and finds that the average BMI value for these males is 29 and that the distribution has a spread that is symmetrical on both sides of the mean.

Guess the Distribution - 3

Paul has the option between a high deductible plan and a low deductible plan for health insurance. If Paul chooses the low deductible plan he will pay the first 1000 dollars of the any medical costs. The low deductible plan costs 8000 dollars.If Paul chooses the high deductible plan he will pay the first 2500 dollars of any medical costs. The high deductible plan costs 7500 dollars. Paul found a table of data on the frequency of medical costs.

Cost	Probability
0	30%
1000	25%
4000	20%
7000	20%
15000	5%

Guess the Distribution - 4

Hospital staff is curious about the amount of staff they need on hand in the evenings. Staff collects data on the number of patients arriving in an emergency room between 10 and 11 pm each day.