ضروریات آمار برای علم داده

مروری بر انواع داده و توزیع‌های آماری و مفهوم ناپارامتری

- انواع داده
  - کمی(عددی)
    - پیوسته
    - گسسته
  - کیفی(غیرعددی)
    - اسمی
    - ترتیبی

متغیر تصادفی

Outcome	X	Probability
NNN	0	3/4*3/4*3/4
NND	1	3/4*3/4*1/4
NDN	1	3/4*1/4*3/4
DNN	1	1/4*3/4*3/4
NDD	2	3/4*1/4*1/4
DND	2	1/4*3/4*1/4
DDN	2	1/4*1/4*3/4
DDD	3	1/4*1/4*1/4

- یک متغیر تصادفی تابعی از فضای نمونه به مجموعه اعداد حقیقی است، به طوری که به هر نقطه از فضای نمونه یک عدد حقیقی را نسبت می‌دهد.
- برای نمایش متغیر تصادفی از حروف بزرگ مانند $X, Y, \ldots$ استفاده می‌شود.
- برای نمایش یکی از مقادیری که متغیر تصادفی اختیار می‌کند از حروف کوچک معادل آن استفاده می‌شود.
- مجموعه مقادیر و یا برد متغیر تصادفی X را با S_X نمایش می‌دهند و آن را تکیه‌گاه X می‌گویند.
- متغیر تصادفی گسسته: متغیر تصادفی که مجموعه مقادیر آن متناهی یا نامتناهی شمارش‌پذیر باشد
- متغیر تصادفی پیوسته: متغیرتصادفی که مجموعه مقادیر آن یک فاصله عددی یا اجتماع چند فاصله عددی باشد.

داده‌ها

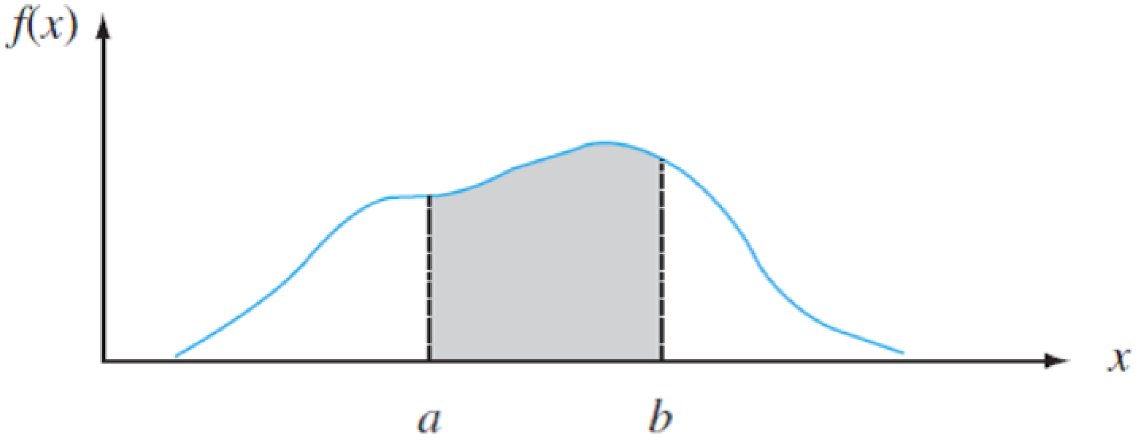
پارامتری

- رفتار داده‌ها شناسایی شده است و اطلاعاتی در زمینه‌ی شکل تابع چگالی توزیع و پارامترهای مورد نیاز برای مشخص‌سازی دقیق داریم.
- فرم بسته‌ای از تابع چگالی کشف شده است
- روابط بین توزیع‌های مختلف اثبات شده است و قضایای موجود در مورد برخی توزیع‌های جهت کاهش درجه سختی مسئله
- امکان داده‌افزایی و شبیه‌سازی و تولید نمونه در صحت سنجی مدل

ناپارامتری

شکل داده‌ها شناخته شده نیست و تعریف مشخصی ندارد یعنی فرم بسته‌ای برای نمونه مشاهده شده یافت نشده است، این موضوع می‌تواند نشات گرفته از خاص بودن مجموعه داده مورد نظر باشد و توزیع‌های پارامتری بیشتر روی پدیده‌های پرتکرارتر با شکل مشخص به وجود آمده‌اند.

تابع چگالی احتمال



1. $f(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)\,dx =$ area under the entire graph of $f(x)$

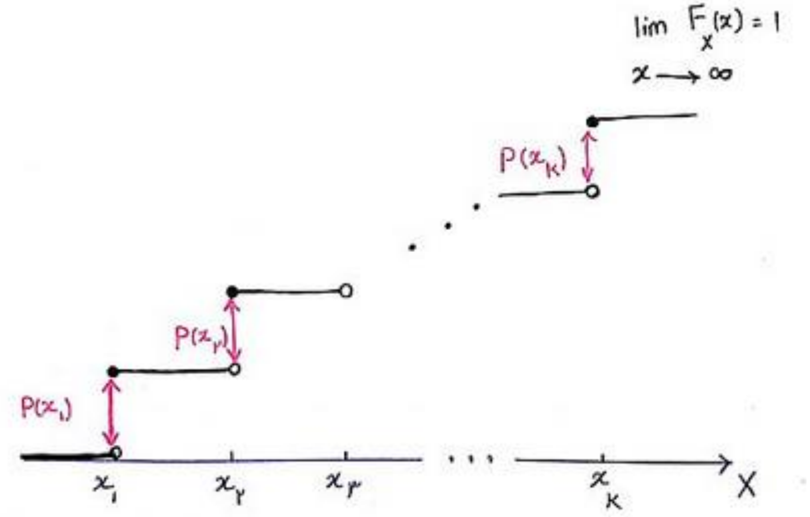
$$= 1$$

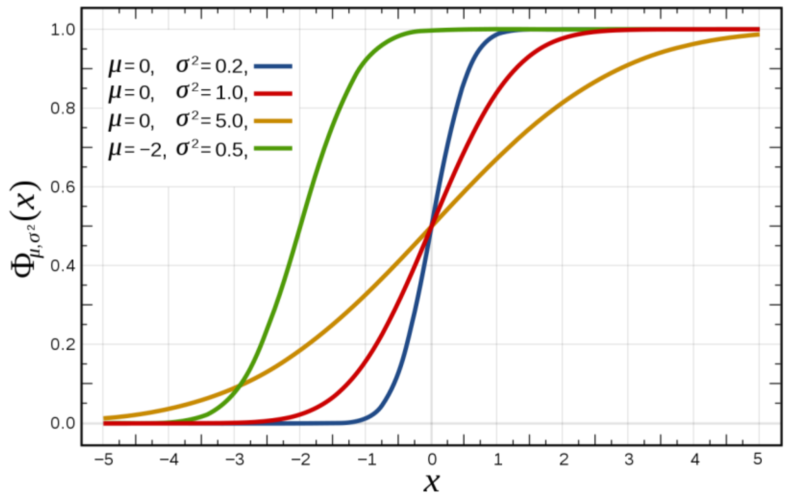
تابع توزیع تجمعی

- تعریف تابع توزیع تجمعی

$$F_X(x) = P(X \leq x) = \sum_{t \leq x} P(t)$$

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$





تابع توزیع تجمعی

- تمام توابع توزیع تجمعی صعودی (ولی نه لزوماً اکیدا صعودی) و از راست پیوسته هستند.

اگر $x_2 <= x_1$ باشد آنگاه:

$$F_X(x_1) \leq F_X(x_2)$$

- بین بازه صفر و یک است

$$0 \leq F_X(x) \leq 1$$

- حد آن در منفی بی‌نهایت صفر و در مثبت بی‌نهایت یک است

$$\lim_{x \to -\infty} F(x) = 0 \qquad \lim_{x \to +\infty} F(x) = 1$$

- $P(X > x) = 1 - F_X(x)$

ارتباط تابع توزیع تجمعی وچگالی(جرم) احتمال

continues

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x)$$

categorical

$$F_X(x) = P(X \le x) = \sum_{t \le x} P(t)$$

$$\Pr(X = x) = F(x_0) - F(x_0-)$$

توزیع برنولی

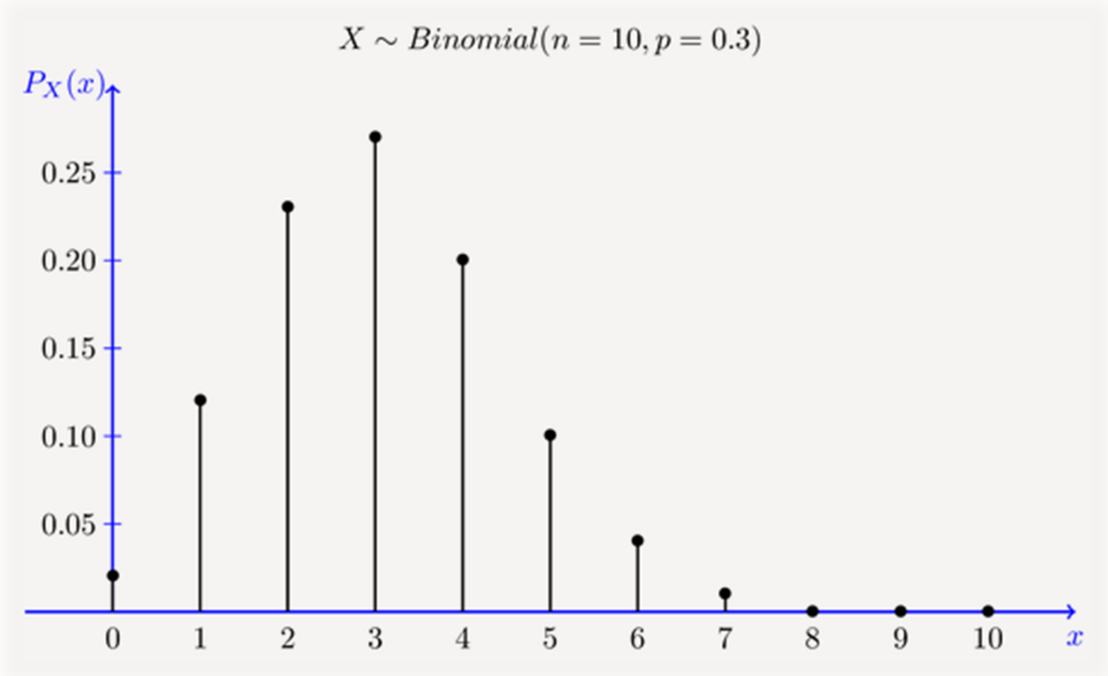
$$X \sim Bernoulli(p)$$

$$P_X(x) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

توزیع دوجمله‌ای

$X \sim Binomial(n, p)$

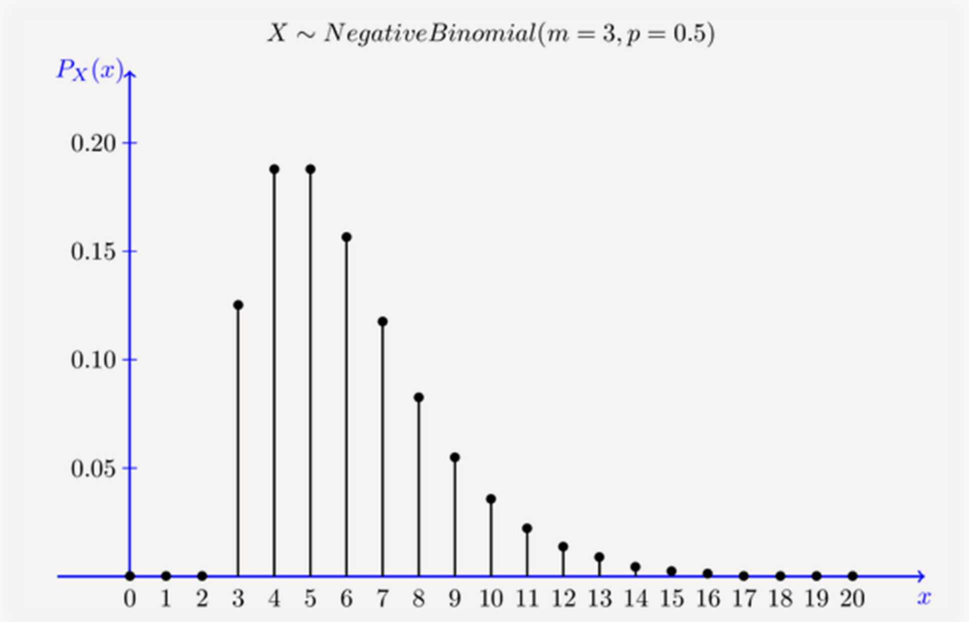
$$P_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } k = 0, 1, 2, \cdots, n \\ 0 & \text{otherwise} \end{cases}$$



توزیع دوجمله‌ای منفی

$X \sim Pascal(m, p)$

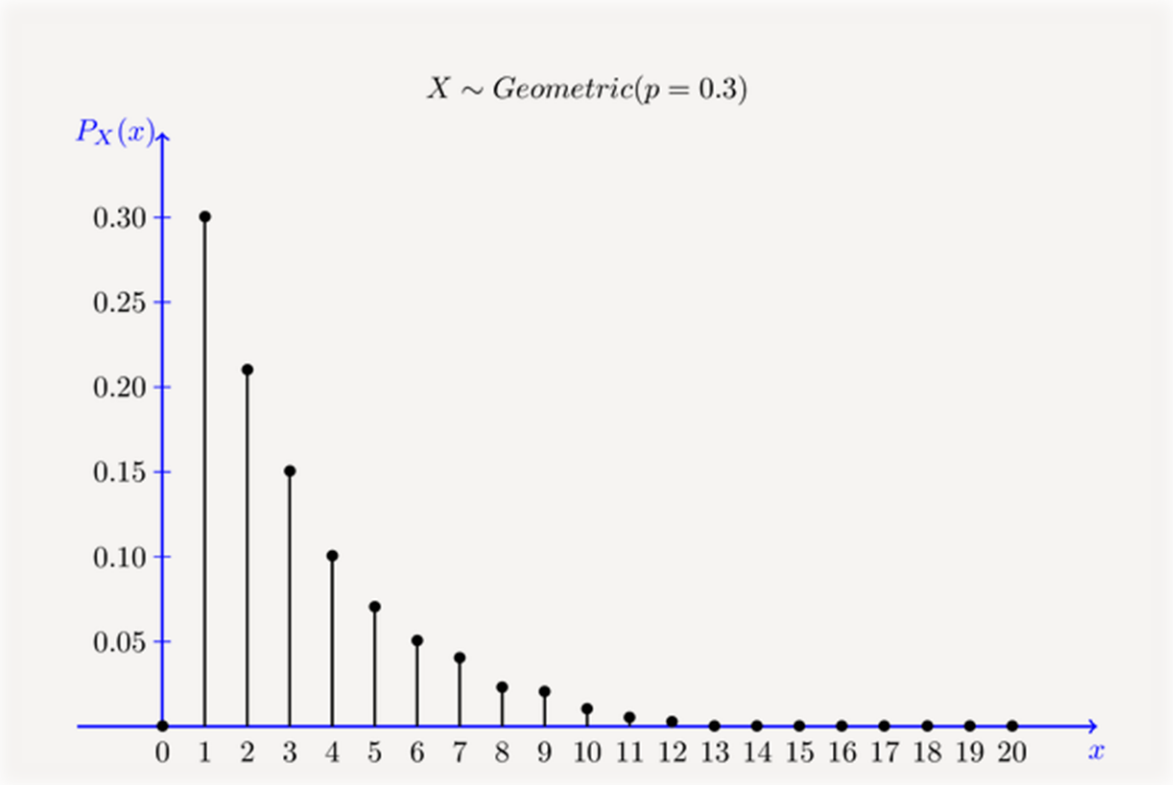
$$P_X(k) = \begin{cases} \binom{k-1}{m-1} p^m (1-p)^{k-m} & \text{for } k = m, m+1, m+2, m+3, \ldots \\ 0 & \text{otherwise} \end{cases}$$



توزیع هندسی

$$X \sim Geometric(p)$$

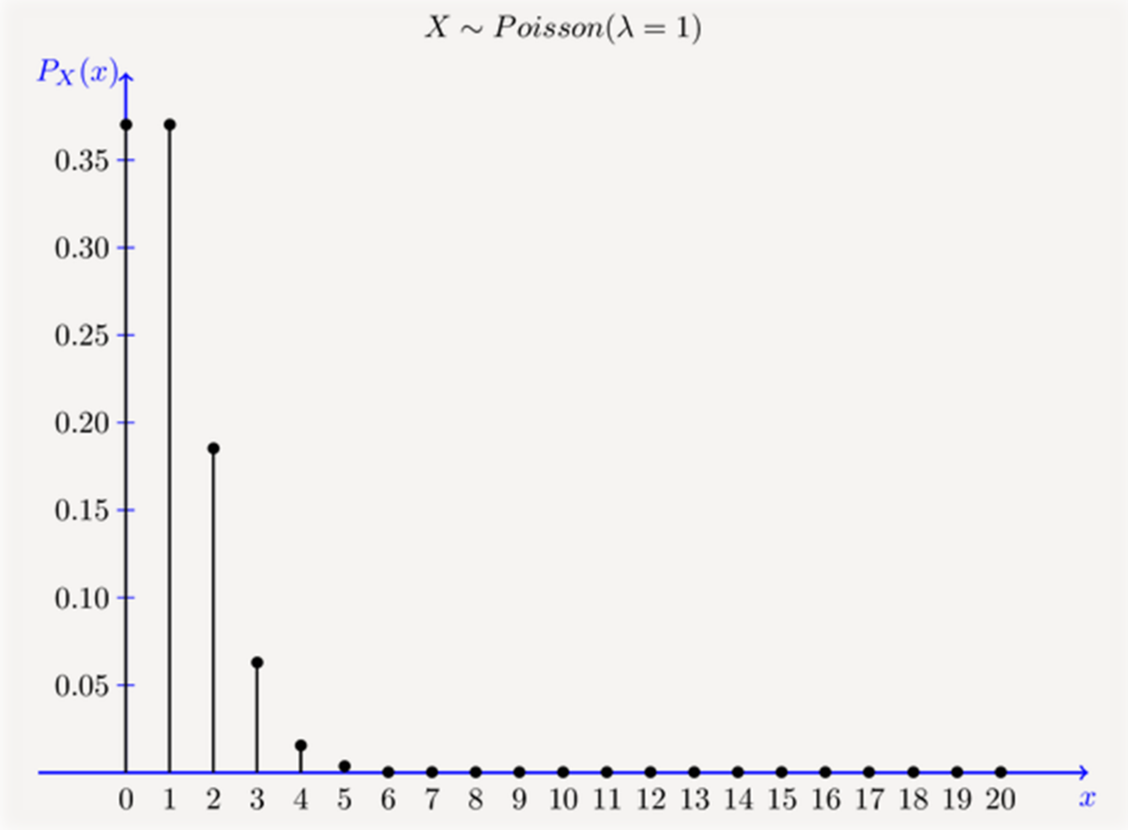
$$P_X(k) = \begin{cases} p(1-p)^{k-1} & \text{for } k = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$



توزیع پواسون

$$X \sim Poisson(\lambda)$$

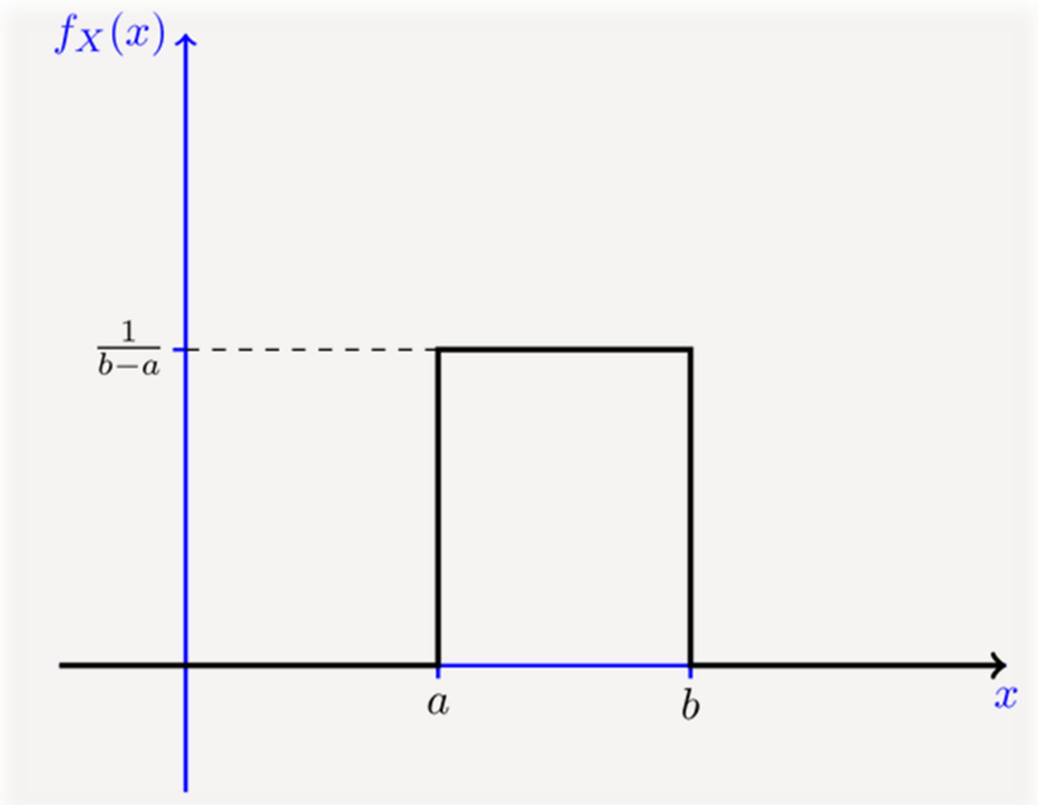
$$P_X(k) = \begin{cases} \frac{e^{-\lambda}\lambda^k}{k!} & \text{for } k \in R_X \\ 0 & \text{otherwise} \end{cases}$$



توزیع یکنواخت پیوسته

$$X \sim Uniform(a,b)$$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x < a \text{ or } x > b \end{cases}$$



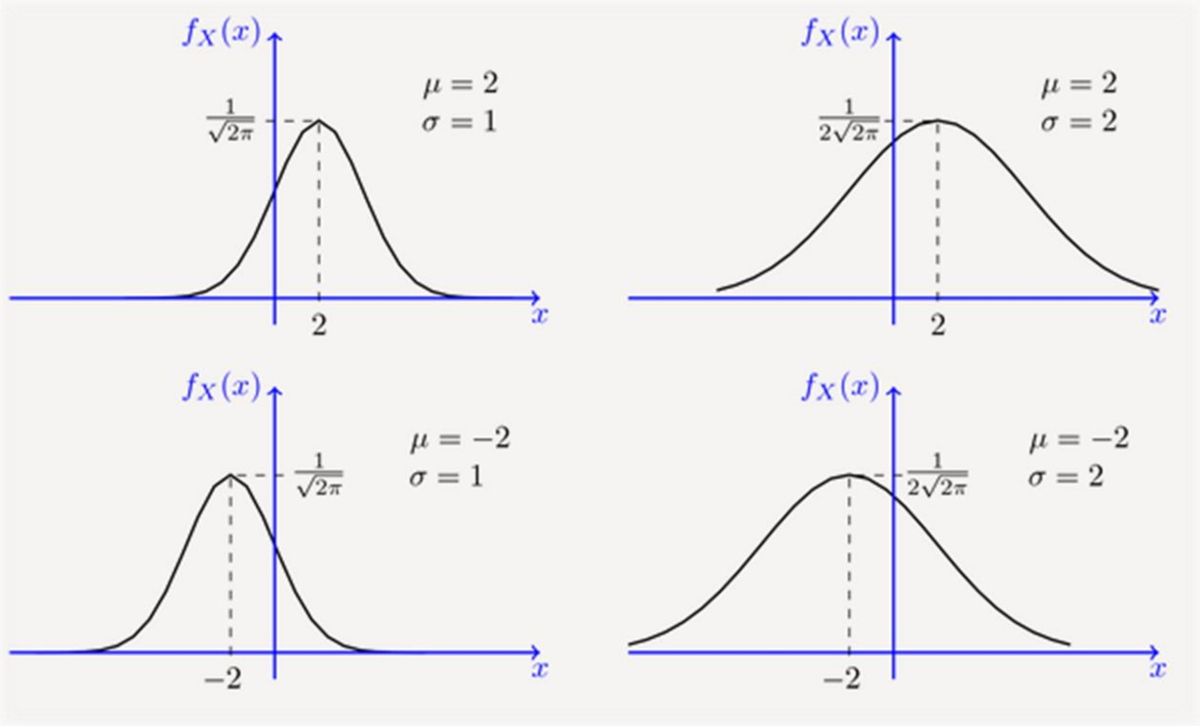
توزیع نرمال

$X \sim N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

$$F_X(x) = P(X \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

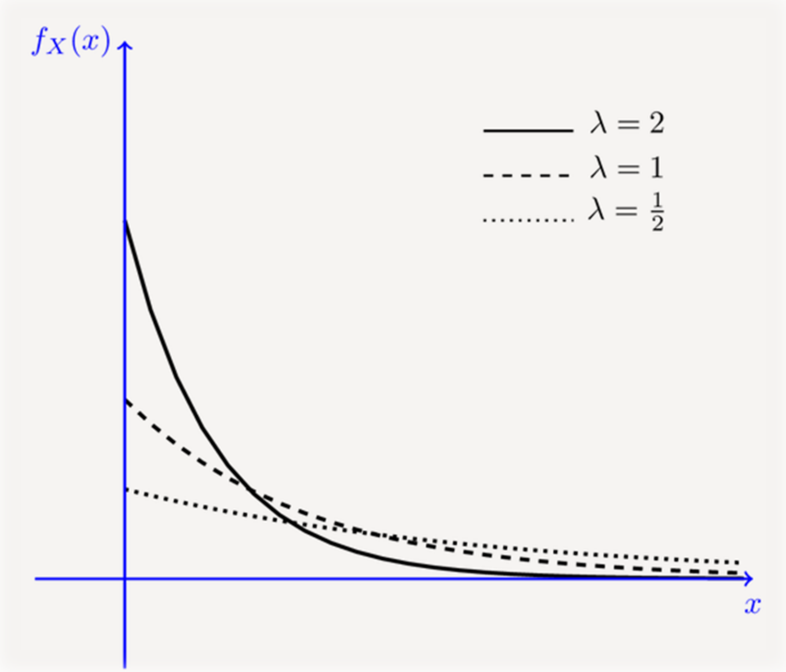
$$P(a < X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$



توزیع نمایی

$X \sim Exponential(\lambda)$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$



توزیع گاما

$X \sim Gamma(\alpha, \lambda)$

$$f_X(x) = \begin{cases} \frac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$