Lecture 3

Primer on Probability and Statistics

Bao-Jun Cai, 3/17/2026

Introduction to Algorithms for Data Science and Physics IMP@Fudan, 2026

Topics of this lecture:

- mean and variance of a distribution $\mathrm{E}_{X\sim p}[X], \mathrm{var}_{X\sim p}[X]$
- positiveness of variance, Jensen's inequality $\mathrm{E}[f(x)] \geq f(\mathrm{E}[x])$
- Bayes' theorem, reduction of variance $\mathrm{var}[w] = \mathrm{E}[\mathrm{var}[w|x]] + \mathrm{var}[\mathrm{E}[w|x]]$
- moment-generating function, binomial distribution $\mathrm{E}[e^{tx}] = \sum e^{tx} p(x) dx$
- central limit theorem: generating Gaussian distribution
- 1D Gaussian, Box-Muller method: uniform->Gaussian

Drawing a fair dice: some basic concepts

$$X = 1, 2, 3, 4, 5, 6$$

$$P_i = P(X = i) = 1/6$$

What will happen if the dice is unfair?

(1) mean/expectation of X, $\mathrm{E}[\cdots] = \overline{\cdots} = \langle \cdots \rangle$

$$\mathrm{E}[X] = \sum_{i=1}^{6} P_i i = \frac{1}{6} \sum_{i=1}^{6} i = \frac{1+2+3+4+5+6}{6} = \frac{7}{2}$$

(2) variance characterizes the deviation of X with respect to $\mathrm{E}[X]$

$$\mathrm{var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \sum_{i=1}^{6} \frac{1}{6} \left(i - \frac{7}{2}\right)^2 = \frac{35}{12}$$

(3) artificially assume that the dice has 3.47,3.48,3.49,3.51,3.52,3.53

$$\mathrm{E}[X] = \frac{7}{2}, \ \mathrm{var}[X] = \frac{2 \cdot (0.01^2 + 0.02^2 + 0.03^2)}{6} = \frac{7}{15000}$$

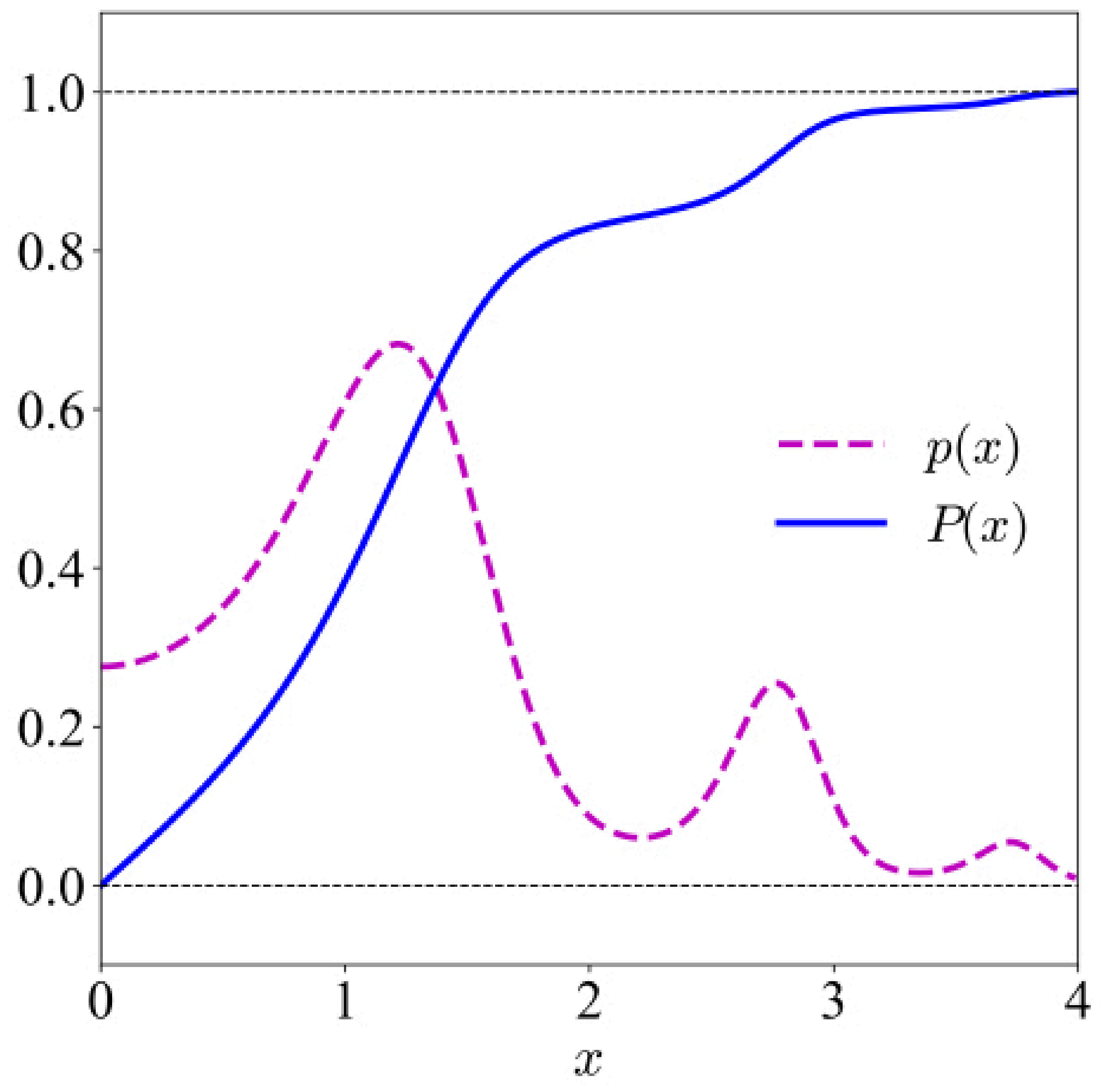
Example: Uniform distribution Unif[a,b]

probability distribution function (pdf)

$p(x)$ itself could be larger than 1!

$$p(x) = \frac{1}{b-a}, \quad a \le x \le b$$

$$P(x) = \int_{-\infty}^{x} p(x')dx'$$

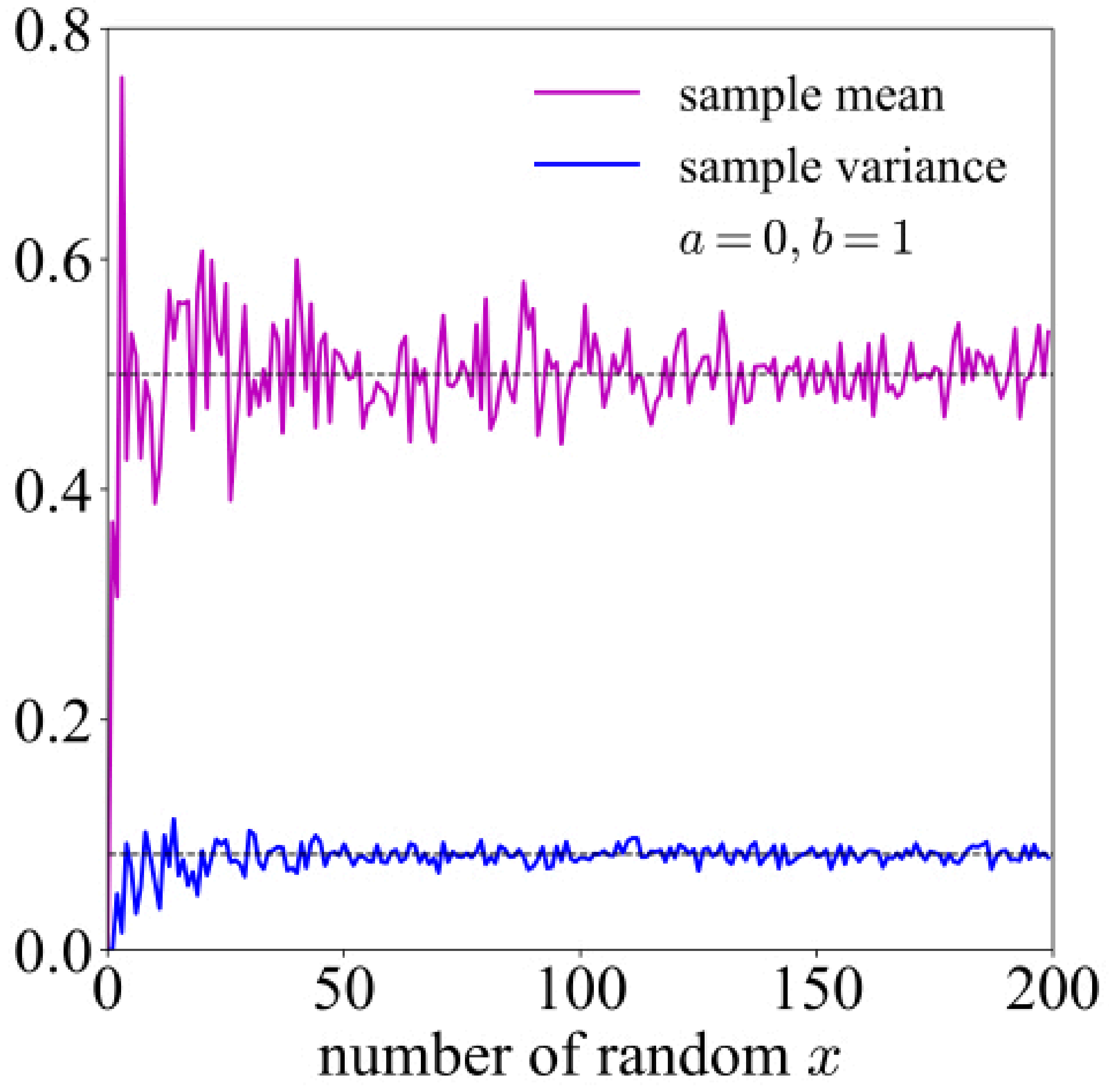


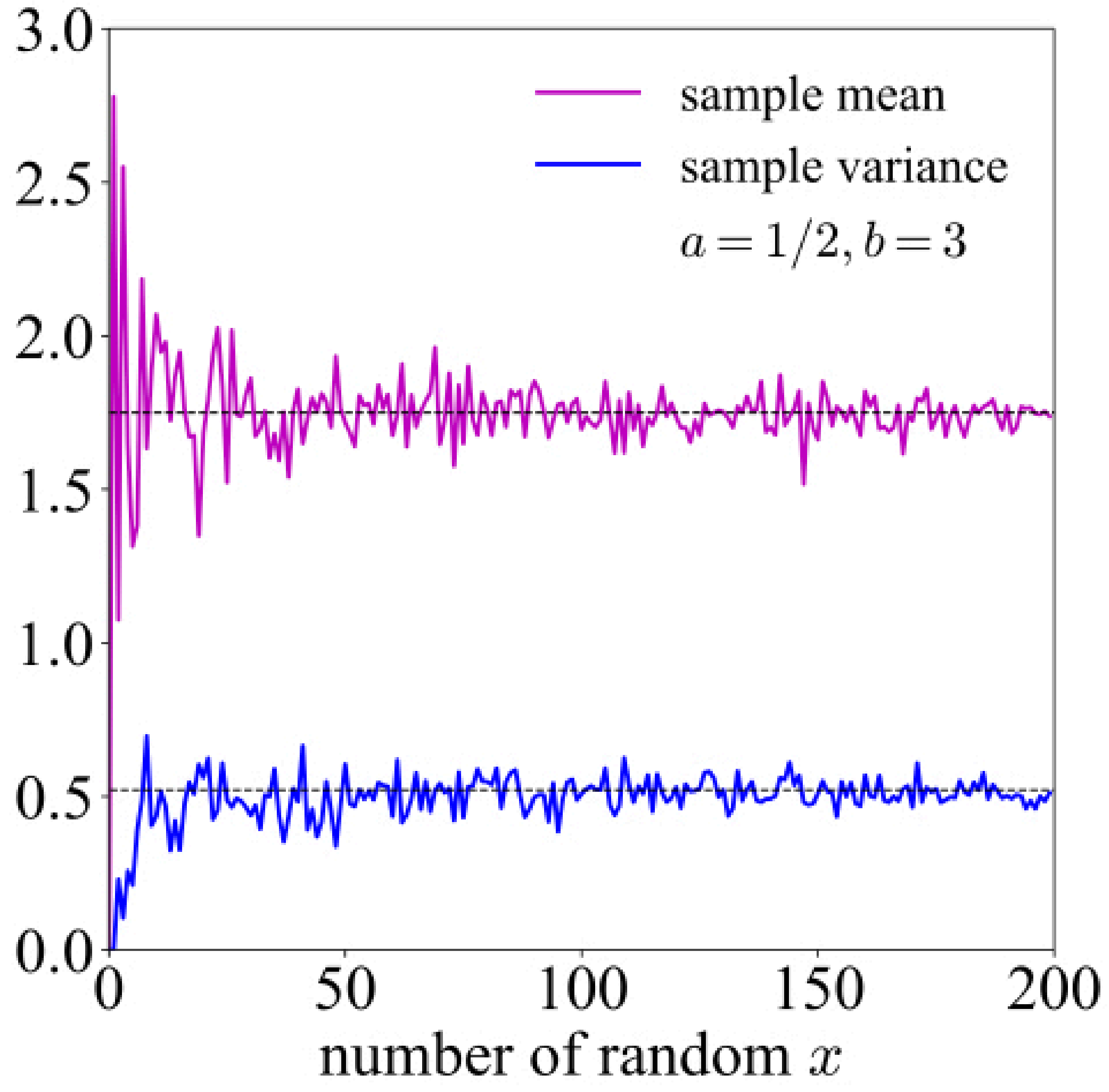
cumulative distribution function (cdf)

$$P(x) = \int_{-\infty}^{x} p(x')dx' = \frac{x-a}{b-a}$$

$$\mathrm{E}[x] \equiv \mathrm{E}[X] = \int_{a}^{b} p(x)xdx = \frac{a+b}{2}$$

Ex.: What is var[x] for Unif[a,b]?





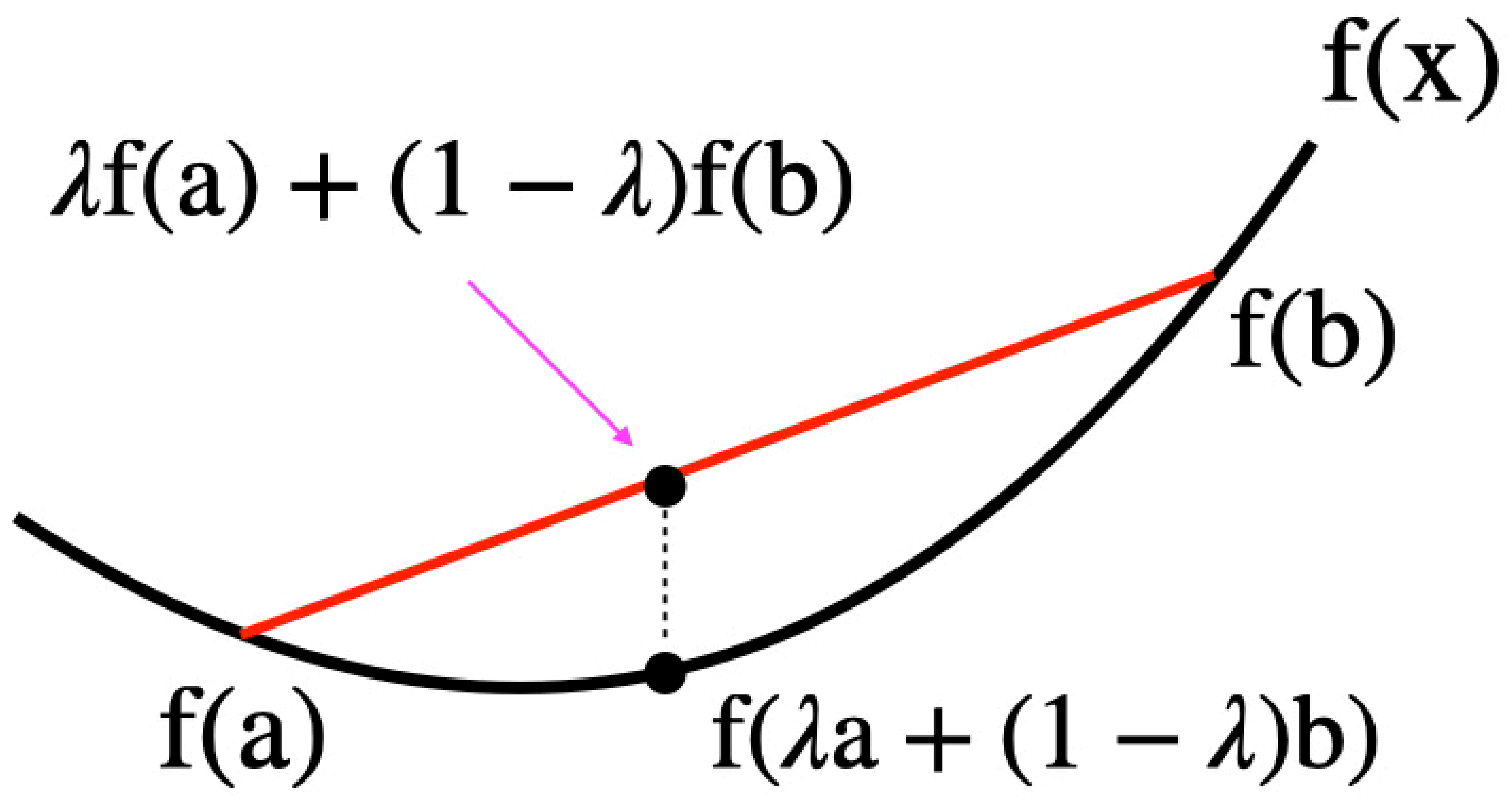
Proof of E[x^2]>E^2[x]

(1) positiveness of variance (definition)

$$
\begin{aligned}
\mathrm{var}[\boldsymbol{x}] &= \mathrm{E}\left[(\boldsymbol{x}-\mathrm{E}[\boldsymbol{x}])^2\right] \\
&= \mathrm{E}\left[\boldsymbol{x}^2 - 2\boldsymbol{x}\mathrm{E}[\boldsymbol{x}] + (\mathrm{E}[\boldsymbol{x}])^2\right] \\
&= \mathrm{E}[\boldsymbol{x}^2] - 2\mathrm{E}[\boldsymbol{x}\mathrm{E}[\boldsymbol{x}]] + \mathrm{E}\left[(\mathrm{E}[\boldsymbol{x}])^2\right] \\
&= \mathrm{E}[\boldsymbol{x}^2] - 2\,(\mathrm{E}[\boldsymbol{x}])^2 + (\mathrm{E}[\boldsymbol{x}])^2 \\
&= \mathrm{E}[\boldsymbol{x}^2] - (\mathrm{E}[\boldsymbol{x}])^2 \equiv \mathrm{E}[\boldsymbol{x}^2] - \mathrm{E}^2[\boldsymbol{x}]
\end{aligned}
$$

Ex.: What can you learn from the function f(x)=-log(x)?

(2) Jensen's inequality for convex function


f(x)
λf(a) + (1 − λ)f(b)
f(b)
f(a)
f(λa + (1 − λ)b)

$$\lambda f(a) + (1-\lambda)f(b) \geq f(\lambda a + (1-\lambda)b)$$

$$\boxed{\langle f(x)\rangle \leq f(\langle x\rangle) \leftrightarrow \mathrm{E}[f(x)] \geq f[\mathrm{E}[x]]}$$

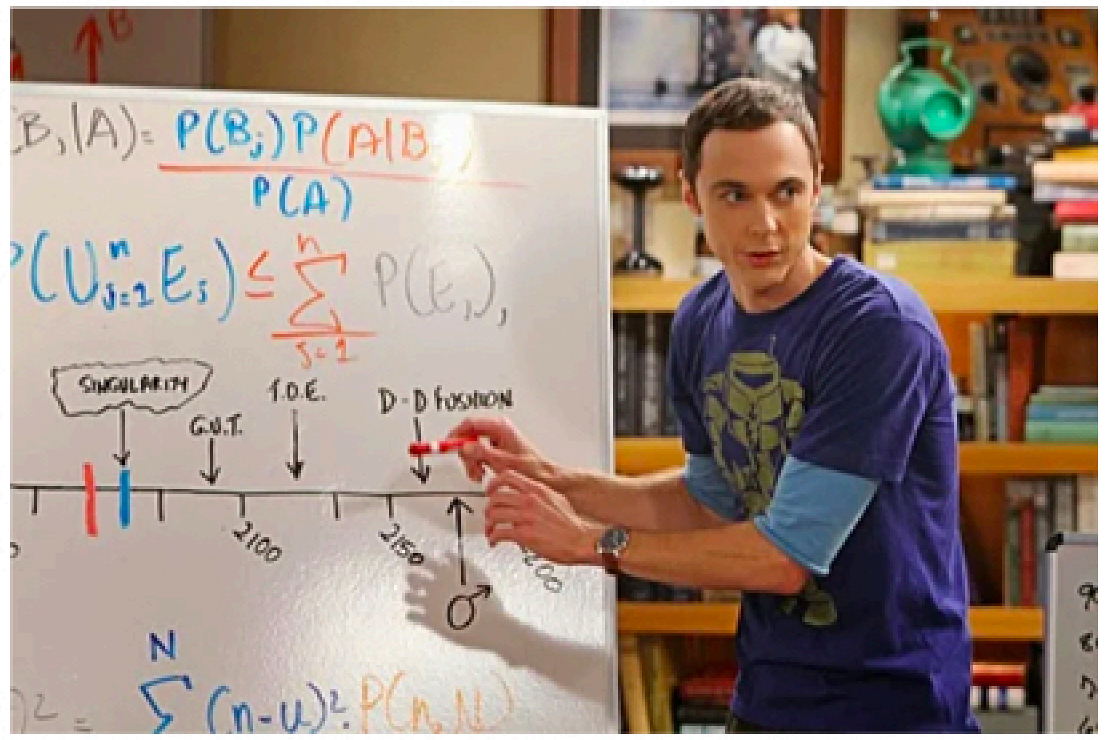
$$f(x) = x^2 \text{ is convex} \rightarrow \langle x^2\rangle > \langle x\rangle^2$$

Joint probability, Bayes' theorem

probability of *A* and *B*

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

$$\xrightarrow{A \perp B} \quad P(A,B) = P(A)P(B)$$



Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example: probability of dice=2

(1) $P(2) = 1/6$

(2) $P(\text{even}) = 1/2,\ P(2|\text{even}) = 1/3$

polynomial fitting: $\vec{w} \leftarrow$ noisy (data)

$$f_{\vec{w}}(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_n x^n$$

A: parameter(s) to be learned/estimated

B: data (experiment/observation)

P(A): knowledge on A before observing data

P(A|B): knowledge on A after observing data

P(B): independent of parameter(s) A

Variance reduction from data: principle of learning

$$\mathrm{E}[w] = \mathrm{E}[\mathrm{E}[w|x]],\ \mathrm{var}[w] = \mathrm{E}[\mathrm{var}[w|x]] + \mathrm{var}[\mathrm{E}[w|x]]$$

variance of parameter w

variance of parameter w after data generation

positive definite

$$\mathrm{var}[x] = \mathrm{E}[x^2] - \mathrm{E}^2[x]$$

proof:

joint distribution for w and x

$$\mathrm{E}[w] = \iint wp(w,x)dwdx = \iint wp(w|x)p(x)dwdx = \int \mathrm{E}[w|x]p(x)dx = \mathrm{E}[\mathrm{E}[w|x]]$$

$$\begin{aligned}\mathrm{E}[\mathrm{var}[w|x]] + \mathrm{var}[\mathrm{E}[w|x]] =& \mathrm{E}\left[\mathrm{E}[w^2|x] - (\mathrm{E}[w|x])^2\right] + \mathrm{E}\left[(\mathrm{E}[w|x])^2\right] - (\mathrm{E}\left[\mathrm{E}[w|x]\right])^2 \\ =& \mathrm{E}[w^2] - \mathrm{E}\left[(\mathrm{E}[w|x])^2\right] + \mathrm{E}\left[(\mathrm{E}[w|x])^2\right] - (\mathrm{E}[w])^2 \\ =& \mathrm{E}[w^2] - (\mathrm{E}[w])^2 \\ =& \mathrm{var}[w].\end{aligned}$$

w: parameter

x: data

variance of the parameter to be estimated as the data is generated is eventually reduced

Quiz 1: 3/17/2026

Quiz 1.1

Input: array A of n integers, and an integer t.

Output: Whether or not A contains the element t.

—

```
for $i = 1$ to $n$ do
    if $A[i] = t$ then
        return TRUE
return FALSE
```

—

What is the running time of the algorithm?

(a) $\mathcal{O}(1)$ (b) $\mathcal{O}(\log n)$ (c) $\mathcal{O}(n)$ (d) $\mathcal{O}(n^2)$

Quiz 1.2

Let $T(n) = n^2/3 + 2n$. Which are true for $T(n) \sim$?

(a) $\mathcal{O}(n)$ (b) $\Omega(n)$ (c) $\Omega(n^2)$ (d) $\mathcal{O}(n^3)$

Quiz 1.3

What are the accuracy of 5-point algorithms for 2nd and 3rd derivative of $f(x)$ with step h?

(a) $\mathcal{O}(h^3)$ and $\mathcal{O}(h^2)$

(b) $\mathcal{O}(h^4)$ and $\mathcal{O}(h^2)$

(c) $\mathcal{O}(h^4)$ and $\mathcal{O}(h^4)$

(d) $\mathcal{O}(h^3)$ and $\mathcal{O}(h^3)$

Quiz 1.4

In a Monte Carlo integration, suppose the sampling surface is located at about 90% of the radius of a d-dimensional sphere. If approximately 10% of the sampled points fall inside the inner volume, what is the minimum dimension d required?

Moment-generating function (MGF)

$$e^{at} \approx 1 + at + \frac{1}{2}a^2t^2 + \frac{1}{6}a^3t^3 + \cdots$$

central moment

$$\nu_k = \mathrm{E}[(x - \mathrm{E}[x])^k]$$

moment

$$\mu_k = \mathrm{E}[x^k]$$

MGF

$$\mathcal{M}_x(t) = \mathrm{E}[e^{tx}] = \int e^{tx} p(x) dx$$

$$e^{tx} \approx 1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \cdots$$

$$\mathcal{M}_x(t) \approx 1 + t\mu_1 + t^2\frac{\mu_2}{2!} + t^3\frac{\mu_3}{3!} + \cdots$$

$$\boxed{\mathcal{M}_x^{(k)}(0) = \mu_k}$$

example: uniform Unif[a,b]

$$\mathcal{M}_x(t) = \int_a^b \frac{e^{tx}}{b-a} dx = \frac{e^{bt} - e^{at}}{t(b-a)}$$

assume that *t* is near zero:

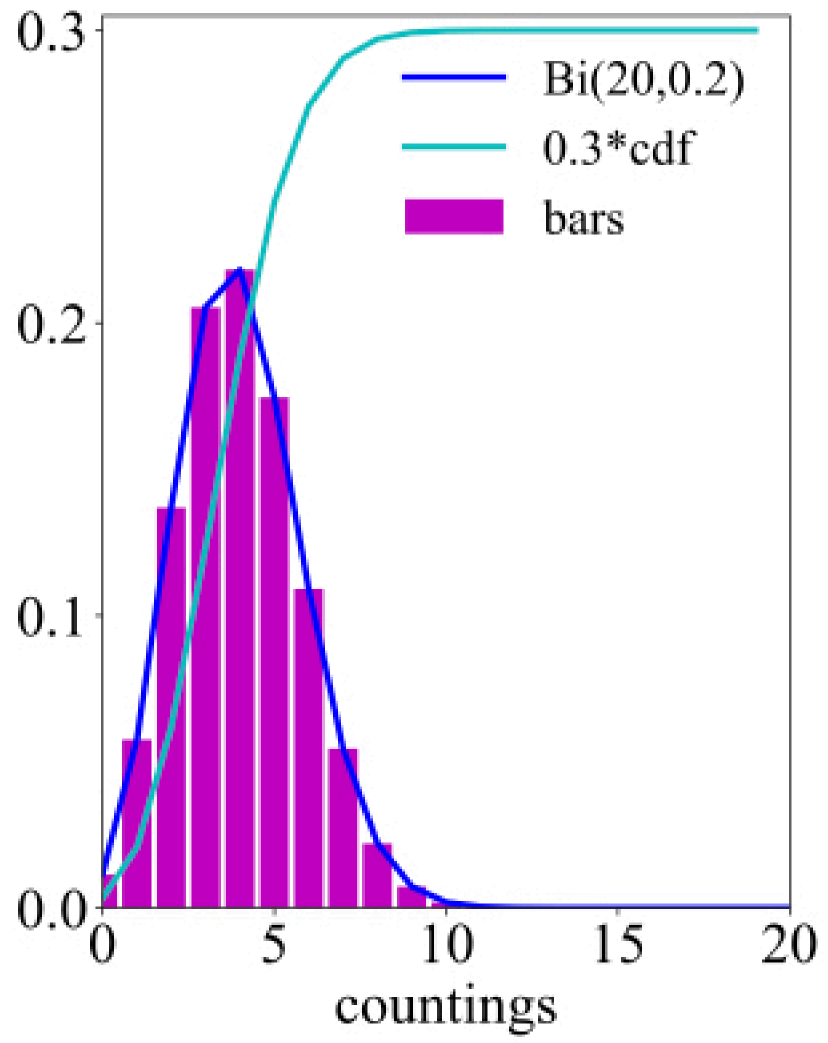
$$\mathcal{M}_x(t) \approx 1 + \frac{b+a}{2}t + \frac{b^2+ab+a^2}{6}t^2 + \frac{b^3+ab^2+a^2b+a^3}{24}t^3 + \cdots$$

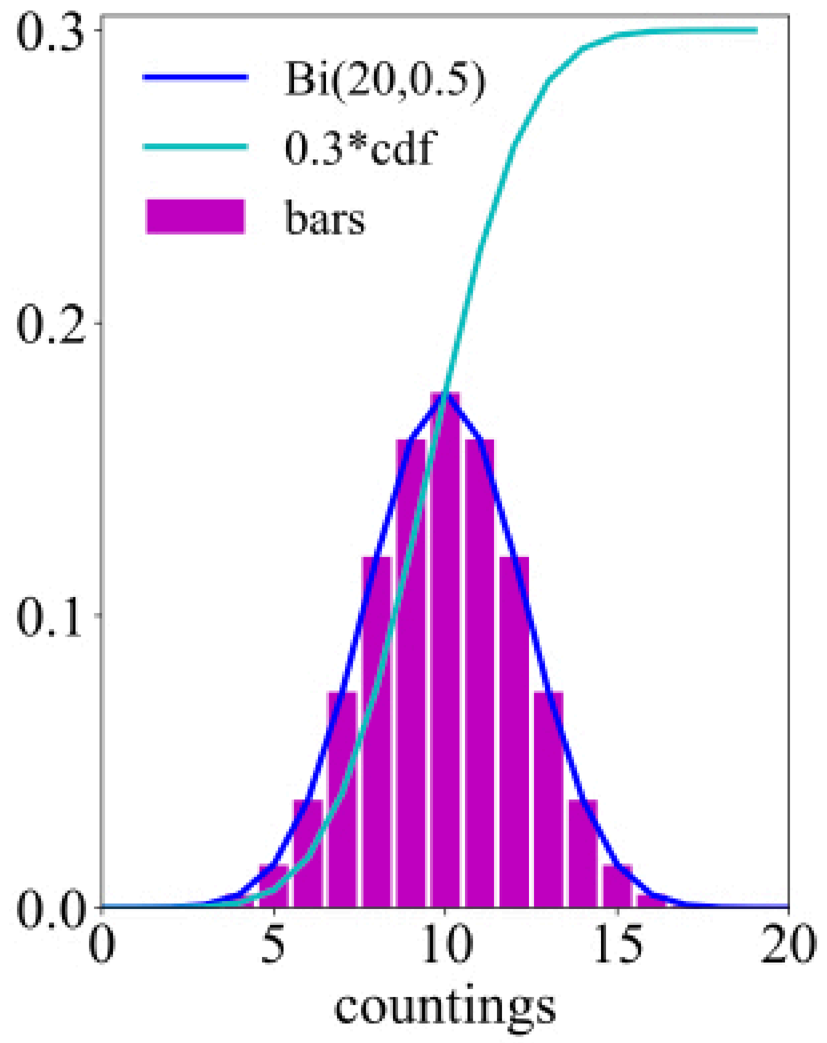
$$\mu_1 = \frac{b+a}{2}, \; \mu_2 = \frac{b^2+ab+a^2}{3}, \cdots$$

Coin: predict the next outcome

p: probability of success (Washington)

q=1-p: probability of failure (Texas map)





$$\mathcal{M}_x(t) = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x q^{n-x} = \sum_{x=0}^{n}\binom{n}{x}\left(pe^t\right)^x q^{n-x} = \left(pe^t + q\right)^n$$

$$\frac{\mathcal{M}_x(t)}{dt} = npe^t\left(pe^t + q\right)^{n-1} \rightarrow \mathrm{E}[x] = \mu_1 = np$$

$$\frac{d^2\mathcal{M}_x(t)}{dt^2} = npe^t\left[\left(pe^t + q\right)^{n-1} + pe^t(n-1)\left(pe^t + q\right)^{n-2}\right]$$

$$\rightarrow \mathrm{var}[x] = \mu_2 - \mu_1^2 = np(1-p)$$

n independent drawings with x heads

$$P(x) = \binom{n}{x}p^x(1-p)^{n-x}$$

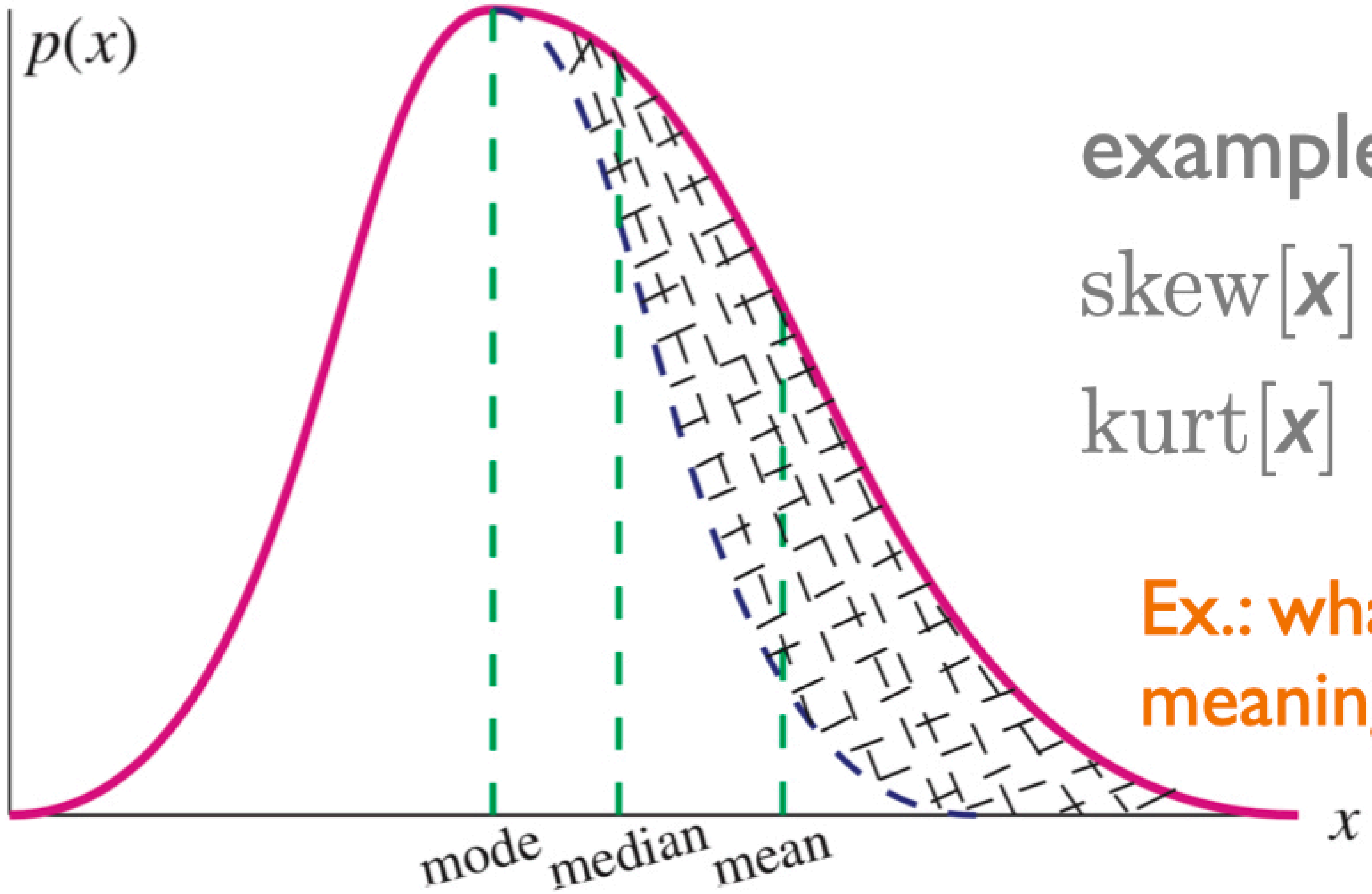
binomial coefficient: $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$

binomial distribution has the maximum variance at p=0.5, if p=0.5 it is difficult to predict the next outcome whether it would be successful or failing, either p=0 or p=1 is deterministic

Geometrical meaning of higher order moments

$$\boxed{\mathrm{skew}[x] \equiv \mathrm{E}\left[\left(\frac{x-\mu}{\sigma}\right)^3\right] = \frac{\nu_3}{\nu_2^{3/2}},\ \mathrm{kurt}[x] \equiv \mathrm{E}\left[\left(\frac{x-\mu}{\sigma}\right)^4\right] - 3 = \frac{\nu_4}{\nu_2^2} - 3}$$

$$\text{skewness} \approx \frac{3(\text{mean-median})}{\text{standard deviation}}$$

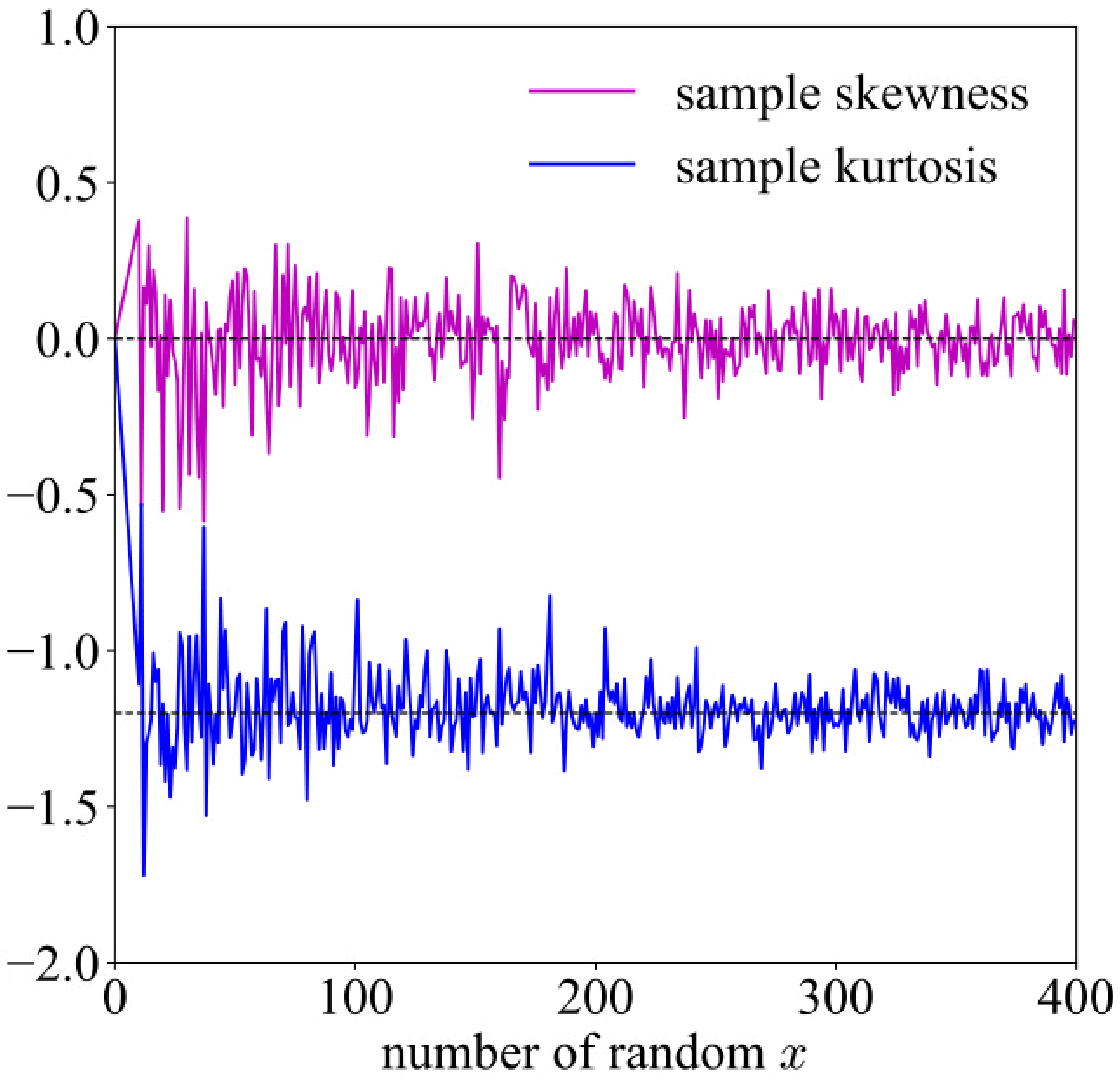


example: Uniform

$\mathrm{skew}[x] = 0$

$\mathrm{kurt}[x] = -6/5$

Ex.: what's the geometrical meaning of kurt[x]?



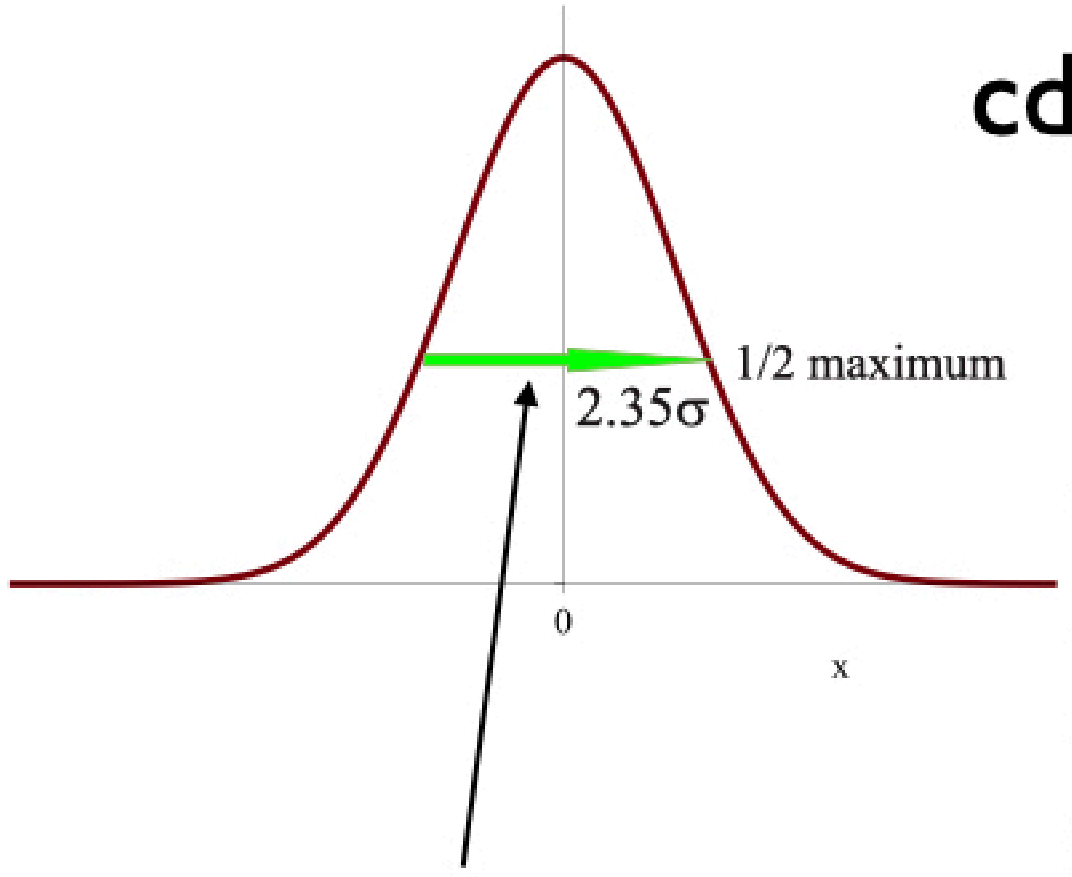
1D Gaussian

$$\mathcal{N}(x|\mu,\sigma^2) \sim \mathcal{N}(\mu,\sigma^2)$$

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

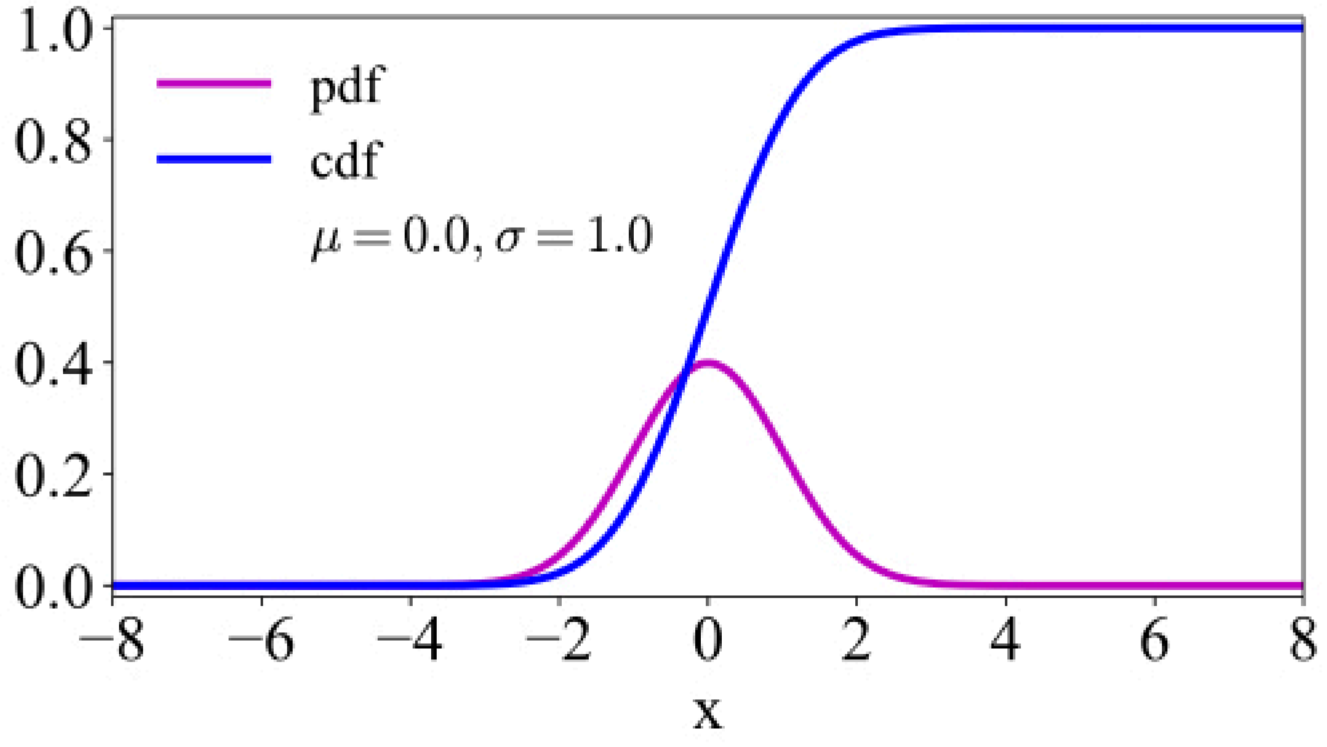
Ex.: Show that p(x) is normalized.

$\mathcal{N}(0,1)$: standard normal distribution



half-width

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



$$\mathcal{M}_x(t) = \int_{-\infty}^{+\infty} e^{tx} p(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} + tx\right) dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{x^2 - 2(\mu+\sigma^2 t)x + \mu^2}{2\sigma^2}\right) dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{\left(x-(\mu+\sigma^2 t)\right)^2}{2\sigma^2} + \mu t + \frac{1}{2}\sigma^2 t^2\right) dx$$

$$= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{\left(x-(\mu+\sigma^2 t)\right)^2}{2\sigma^2}\right) dx}_{\text{normalized: } 1}$$

$$= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

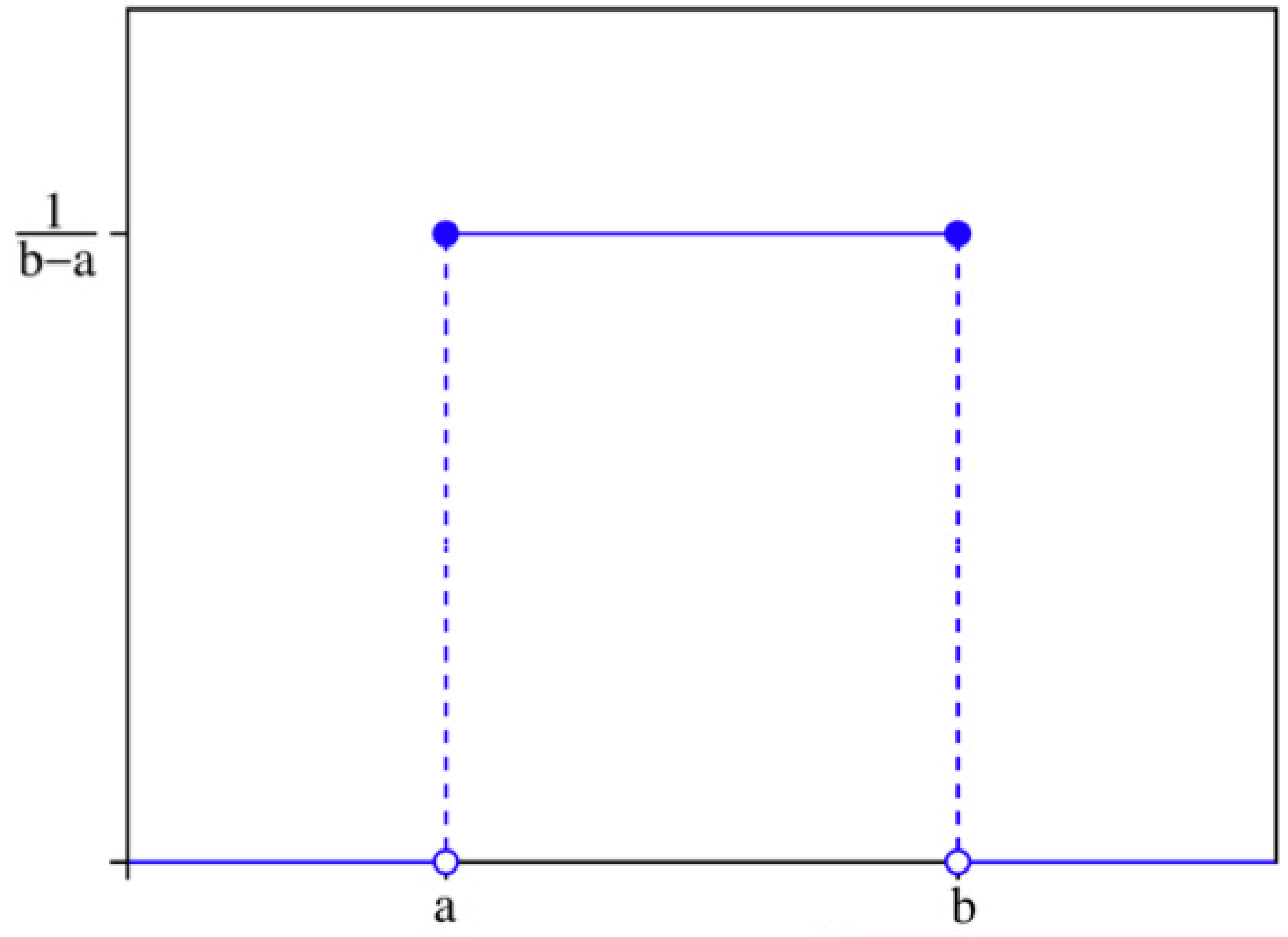
all the information of the distribution is here!

$$\mathcal{M}_x'(0) = ,\ \mathcal{M}_x''(0) = \mu^2 + \sigma^2$$

$$\mathcal{M}_x'''(0) = \mu^3 + 3\mu\sigma^2,\ \mathcal{M}_x''''(0) = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$$

$$\mathrm{skew}[x] = 0,\ \mathrm{kurt}[x] = 0$$

Kurtosis revisited



Unif[a,b]: kurt[x]=-6/5

Exp(λ): kurt[x]=6

positive kurtosis: sharper tail than Gaussian

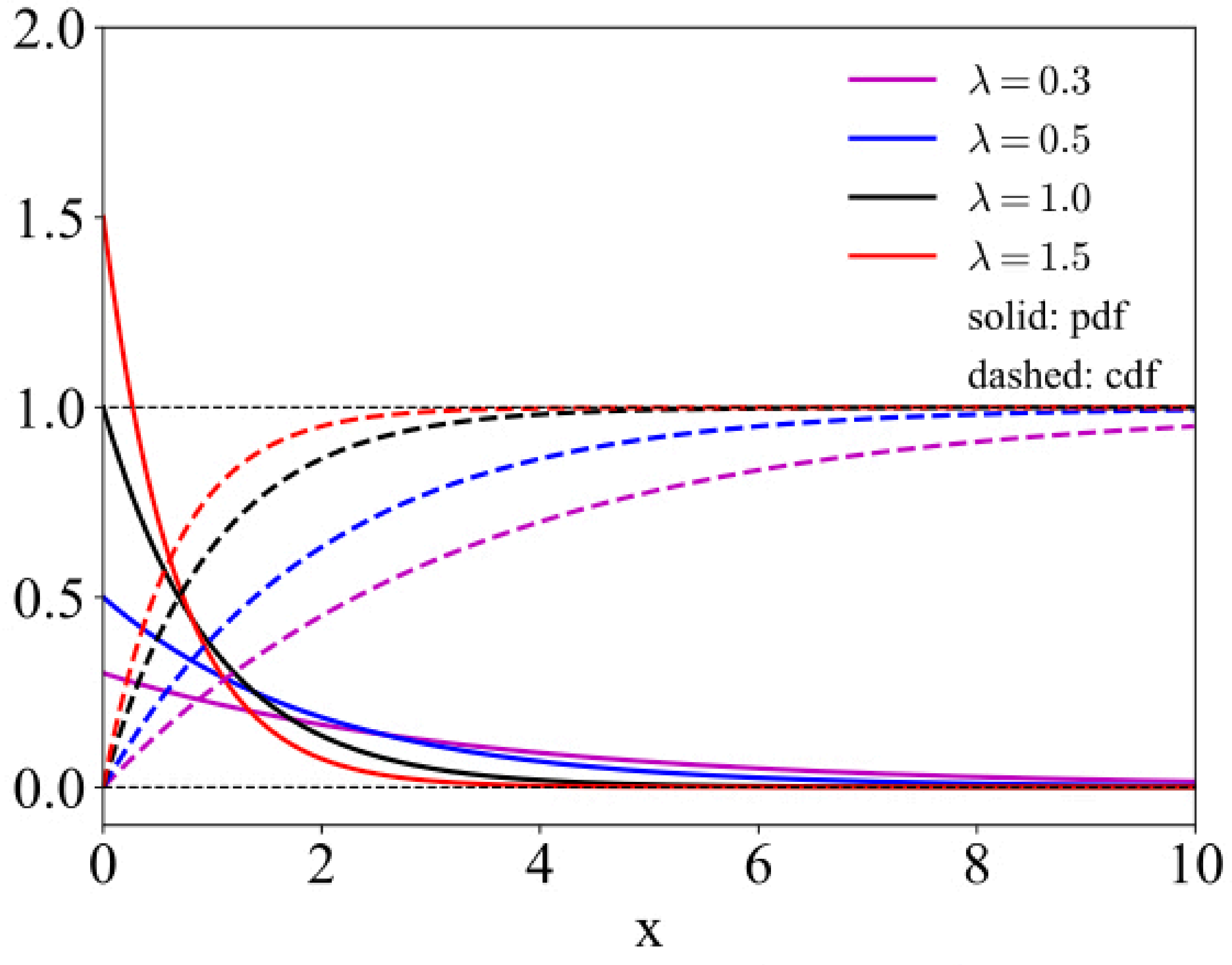
positive skewness: longer tail on the right hand side

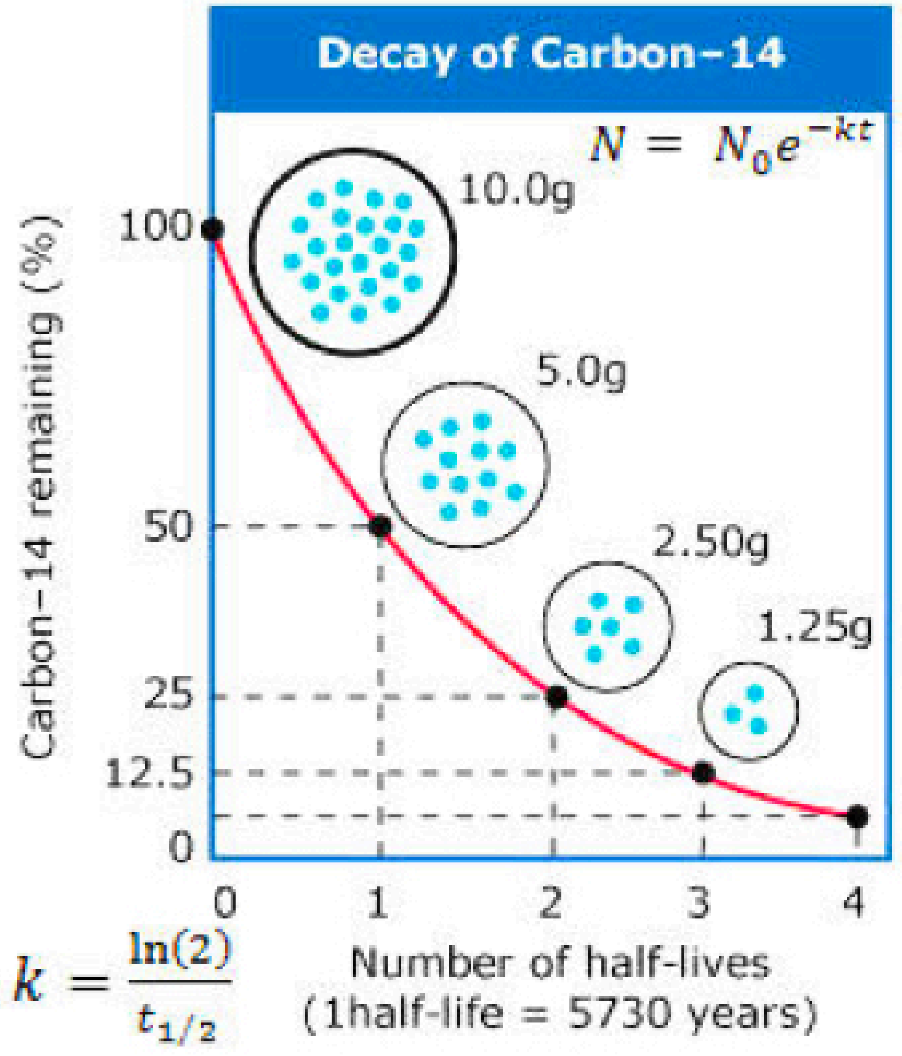
$$\text{Exp}(\lambda) : p_\lambda(x) = \lambda e^{-\lambda x}, x > 0$$

$$\mathcal{M}_x(t) = \int_0^\infty \lambda e^{tx} e^{-\lambda x} dx = \frac{1}{1 - t/\lambda}, \quad t < \lambda$$

$$\approx 1 + \frac{t}{\lambda} + \left(\frac{t}{\lambda}\right)^2 + \left(\frac{t}{\lambda}\right)^3 + \left(\frac{t}{\lambda}\right)^4 + \cdots$$

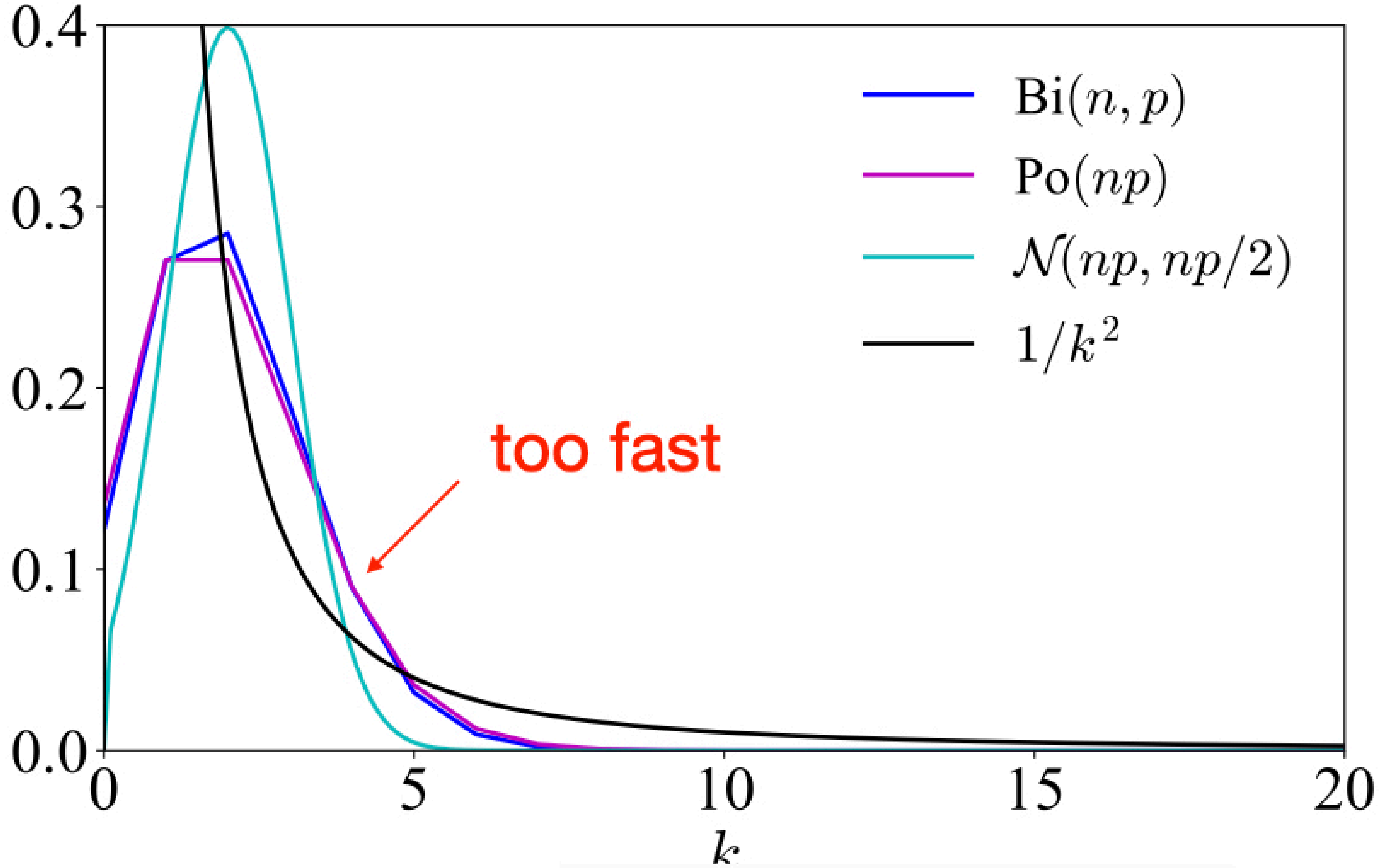
$$\mathcal{M}_x^{(k)}(0) = \frac{k!}{\lambda^k}$$





Ex.: $p_{\text{logn}}(x) = \frac{1}{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$?

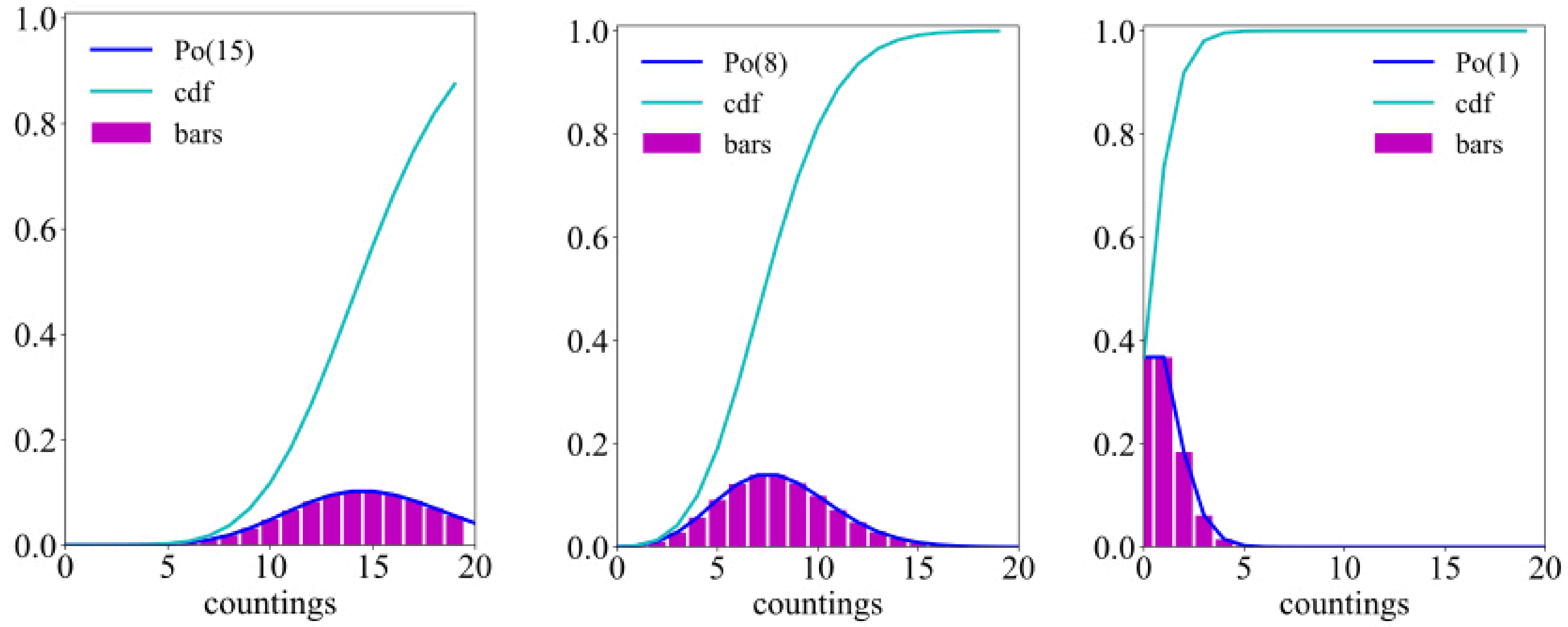
Binomial, power-law, Poisson, Gaussian



if n is large and $\lambda = np$ is fixed:

binomial→Poission

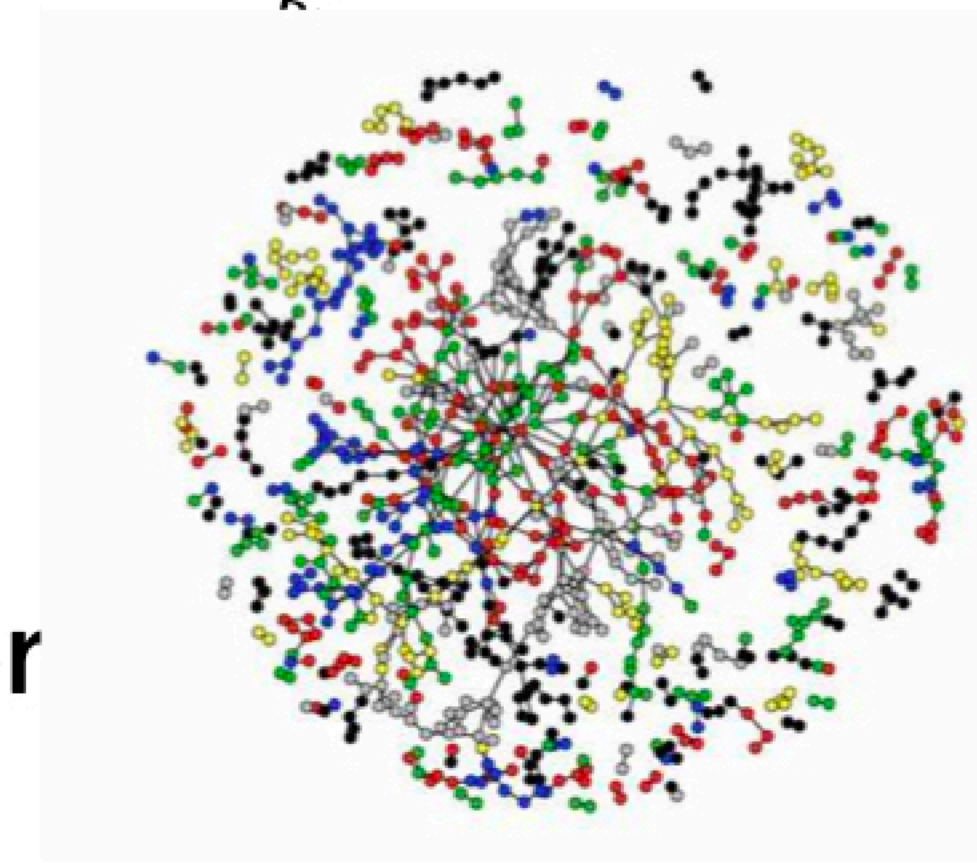
$$\mathrm{Po}(\lambda) : P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

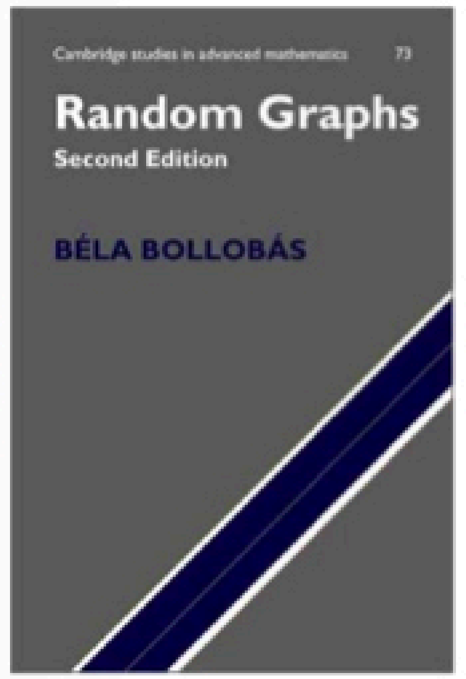


$k^{-\phi}$, $\mathrm{Po}(\lambda)$:

random graph

scale-free networ





Cambridge studies in advanced mathematics 73

Random Graphs

Second Edition

BÉLA BOLLOBÁS

$$\binom{n}{k}p^k(1-p)^{n-k} \approx \frac{1}{\sqrt{\pi np}}\exp\left(-\frac{(np-k)^2}{np}\right)$$

Central limit theorem

$$\boxed{X = \frac{\bar{x}_m - \mu}{\sigma/\sqrt{m}}, \; \bar{x}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}}$$

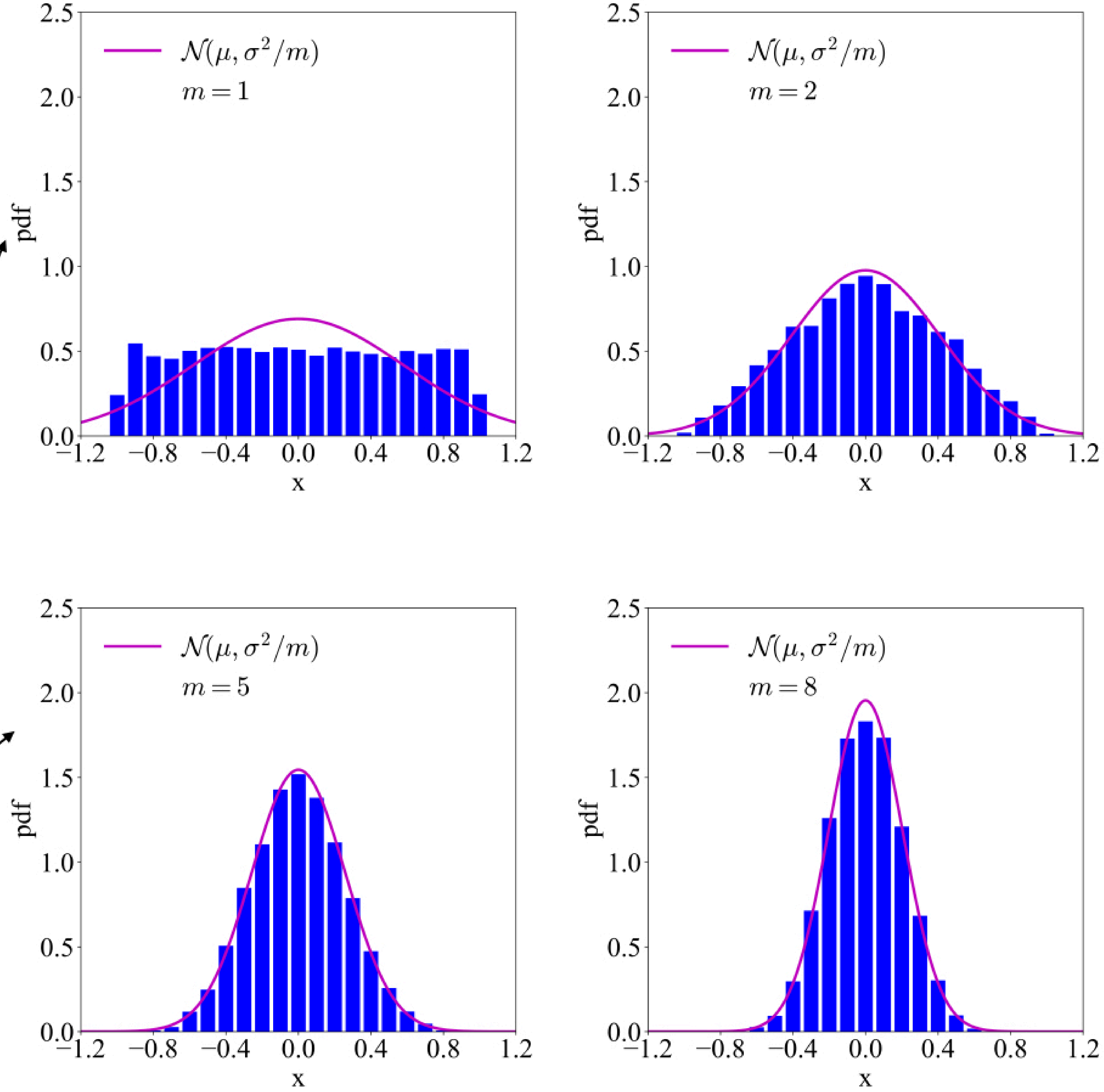
$$\rightarrow X \sim \mathcal{N}(0, 1)$$

indication:

$$\lim_{m\to\infty} \mathcal{M}_x(t) = e^{t^2/2}$$

Ex.: Prove it!

$$x^{(i)} \sim \mathrm{Unif}[-1, 1]$$



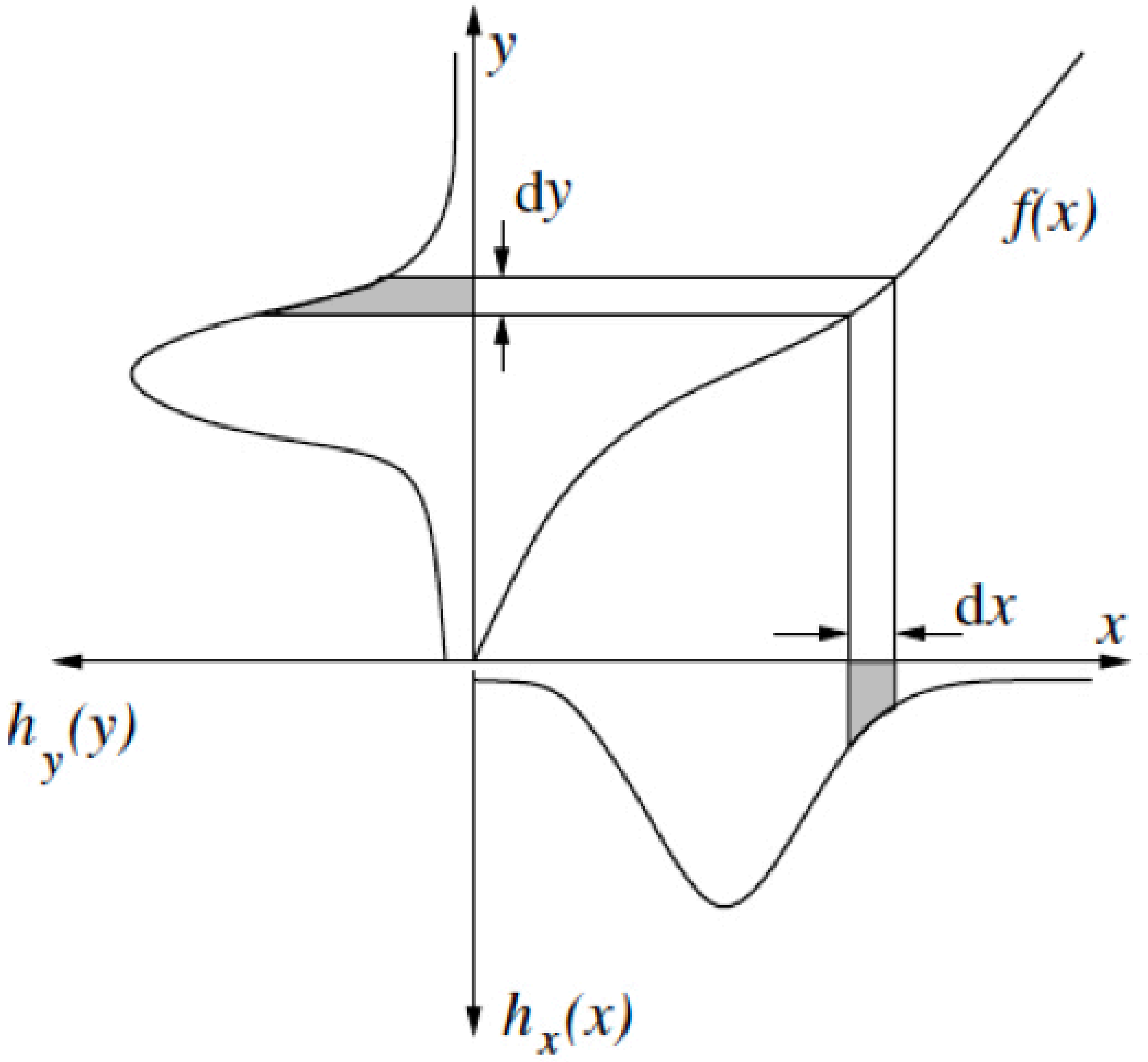
Box-Muller for generating 2D Gaussian

since the very importance of the Gaussian random numbers, algorithm to generate them:

$$y_1 = x_1 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2}, \quad y_2 = x_2 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2}, \quad r^2 = x_1^2 + x_2^2 \le 1$$

$y_1, y_2 \sim \mathcal{N}(0, 1)$

$x_1, x_2 \sim \mathrm{Unif}(0, 1)$



probability conservation:

$p(x)dx = p(y)dy$

$y = f(x)$

$$p(y) = p(x)|\partial x/\partial y| = \frac{p(x)}{f'(x)}$$

proof:

$$p(y_1, y_2) = p(x_1, x_2)\left|\frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right| = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y_1^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y_2^2}{2}\right)$$

$$\frac{\partial(x, y)}{\partial(x', y')} = \begin{vmatrix} \partial x/\partial x' & \partial x/\partial y' \\ \partial y/\partial x' & \partial y/\partial y' \end{vmatrix} = \frac{\partial x}{\partial x'}\frac{\partial y}{\partial y'} - \frac{\partial x}{\partial y'}\frac{\partial y}{\partial x'}$$

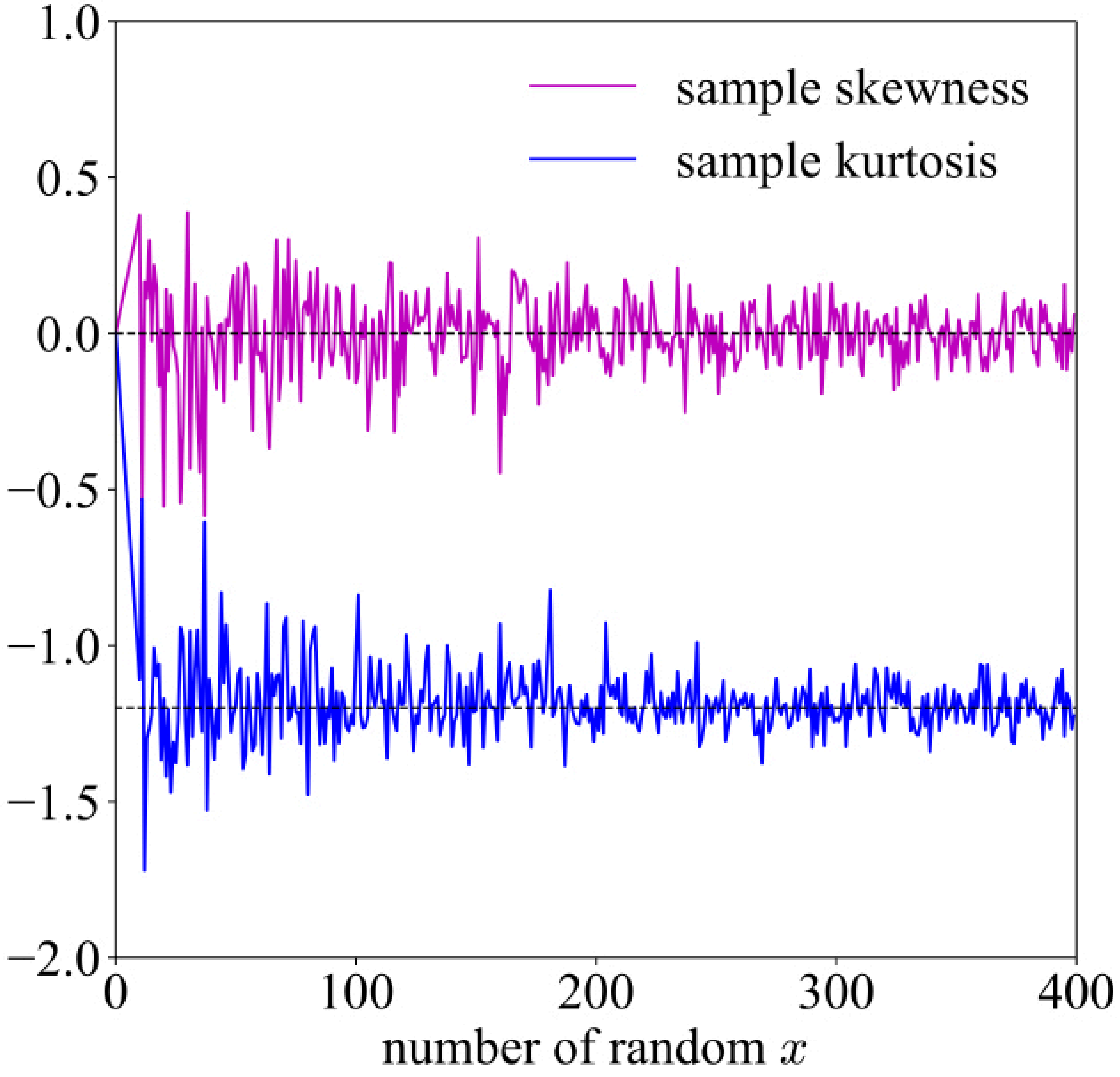
Estimator: 101

$$\widehat{\mu}_1 = \frac{1}{m}\sum_{i=1}^{m} x^{(i)},\ \widehat{\mu}_2 = \frac{1}{m}\sum_{i=1}^{m} x^{(i),2},\ \widehat{\mu}_3 = \frac{1}{m}\sum_{i=1}^{m} x^{(i),3},\ \widehat{\mu}_4 = \frac{1}{m}\sum_{i=1}^{m} x^{(i),4}$$

$x^{(i)} \sim p(x)$: unknown but to be estimated

how good are they?

$$\widehat{\mathrm{skew}}[x] = \frac{\widehat{\mu}_3 - 3\widehat{\mu}_2\widehat{\mu}_1 + 2\widehat{\mu}_1^3}{(\widehat{\mu}_2 - \widehat{\mu}_1^2)^{3/2}},\ \widehat{\mathrm{kurt}}[x] = \frac{\widehat{\mu}_4 - 4\widehat{\mu}_3\widehat{\mu}_1 + 6\widehat{\mu}_2\widehat{\mu}_1^2 - 3\widehat{\mu}_1^4}{(\widehat{\mu}_2 - \widehat{\mu}_1^2)^2} - 3$$



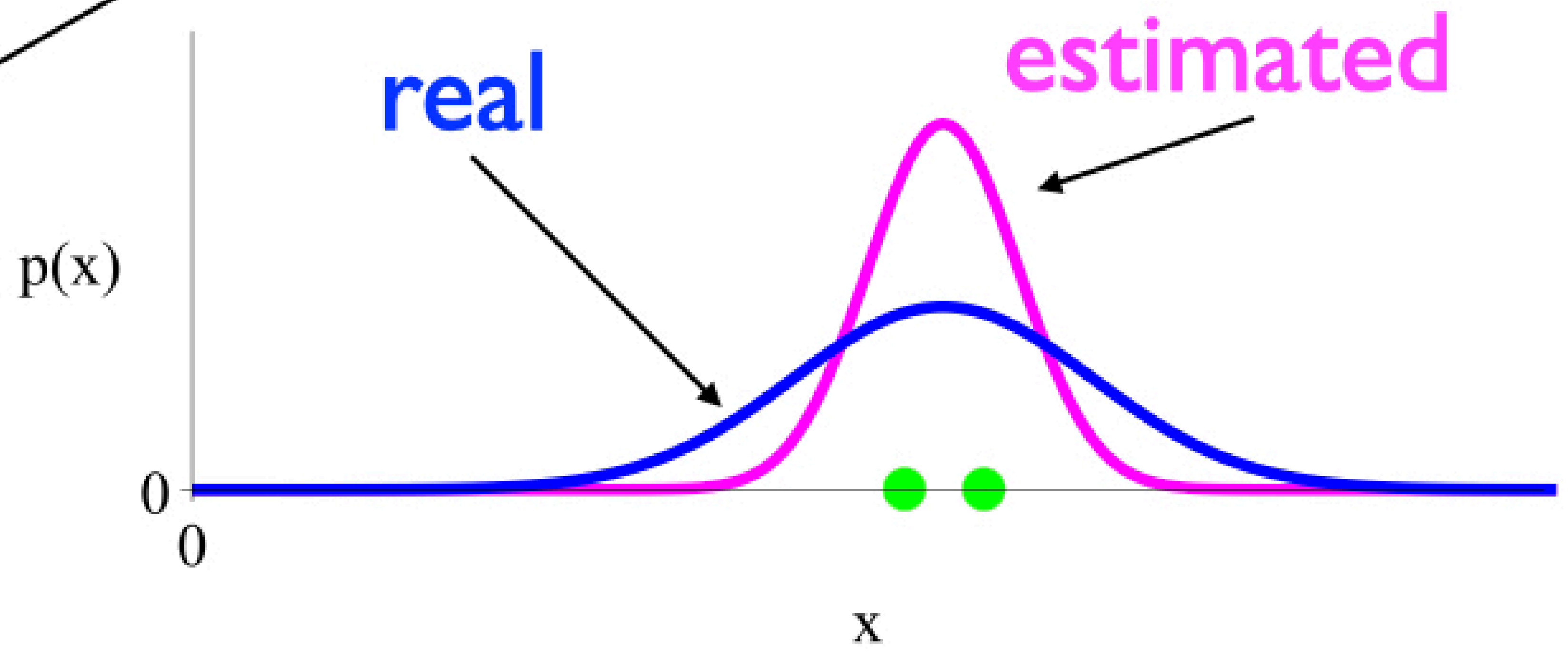
simple example: m samples $x^{(i)} \sim \mathcal{N}(\mu, \sigma^2)$, two estimators:

$$\widehat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}, \widehat{\sigma^2} = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \widehat{\mu}\right)^2 \to \mathrm{E}[\widehat{\mu}] = \mu, \mathrm{E}[\widehat{\sigma^2}] = \left(1 - \frac{1}{m}\right)\sigma^2$$

Ex.: Prove these relations.

BLUE, UMVU, …

$\widehat{\mu}$ is unbiased and $\widehat{\sigma^2}$ is biased


real
estimated
p(x)
x