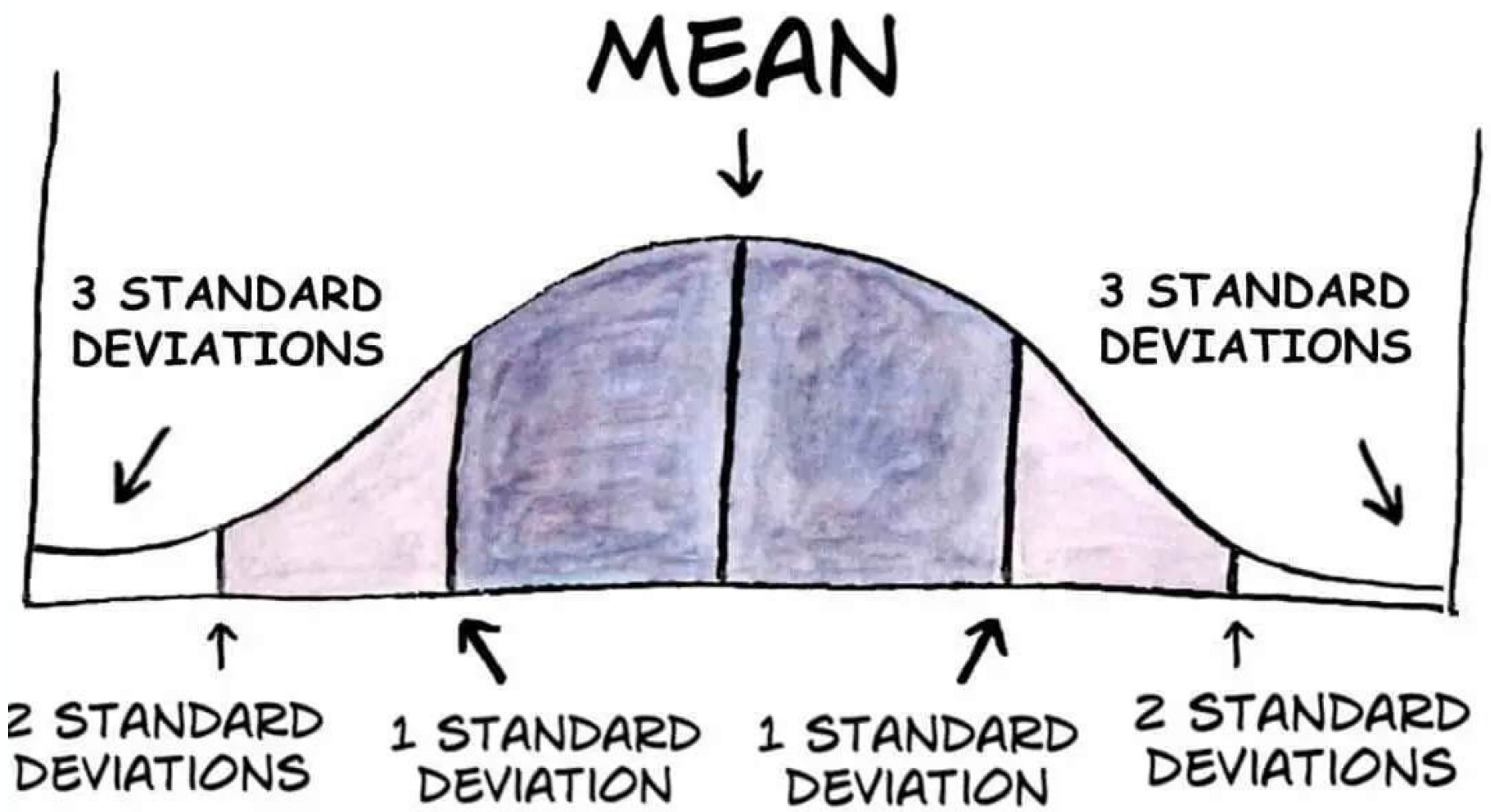
Foundations of Data Science

Moment and Deviation

尹一通，刘明谋 Nanjing University, 2024 Fall

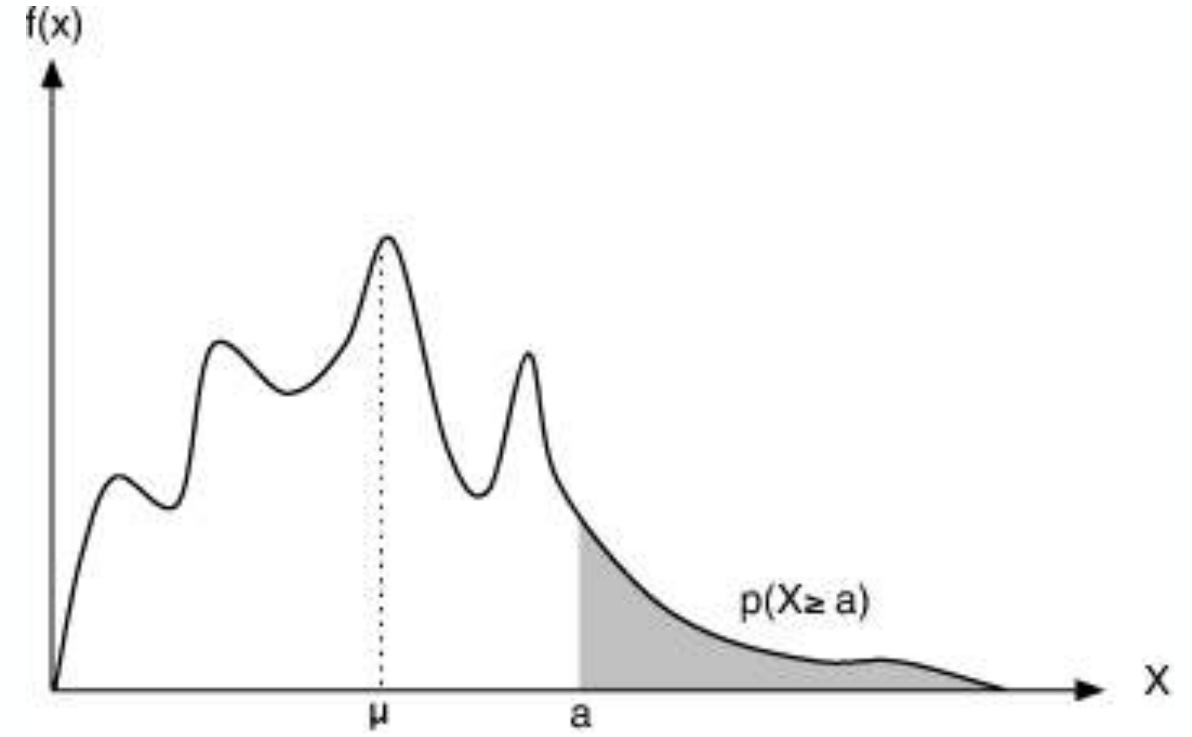
Moments and Deviations

$$\Pr[\,|X - \mathbb{E}[X]|\, > a] = ?$$
$$= \Pr[X < \mathbb{E}[X] - a] + \Pr[X > \mathbb{E}[X] + a]$$
$$= F(\mathbb{E}[X] - a) + (1 - F(\mathbb{E}[X] + a))$$



Markov’s Inequality

(马尔可夫不等式, the first Chebyshev inequality)



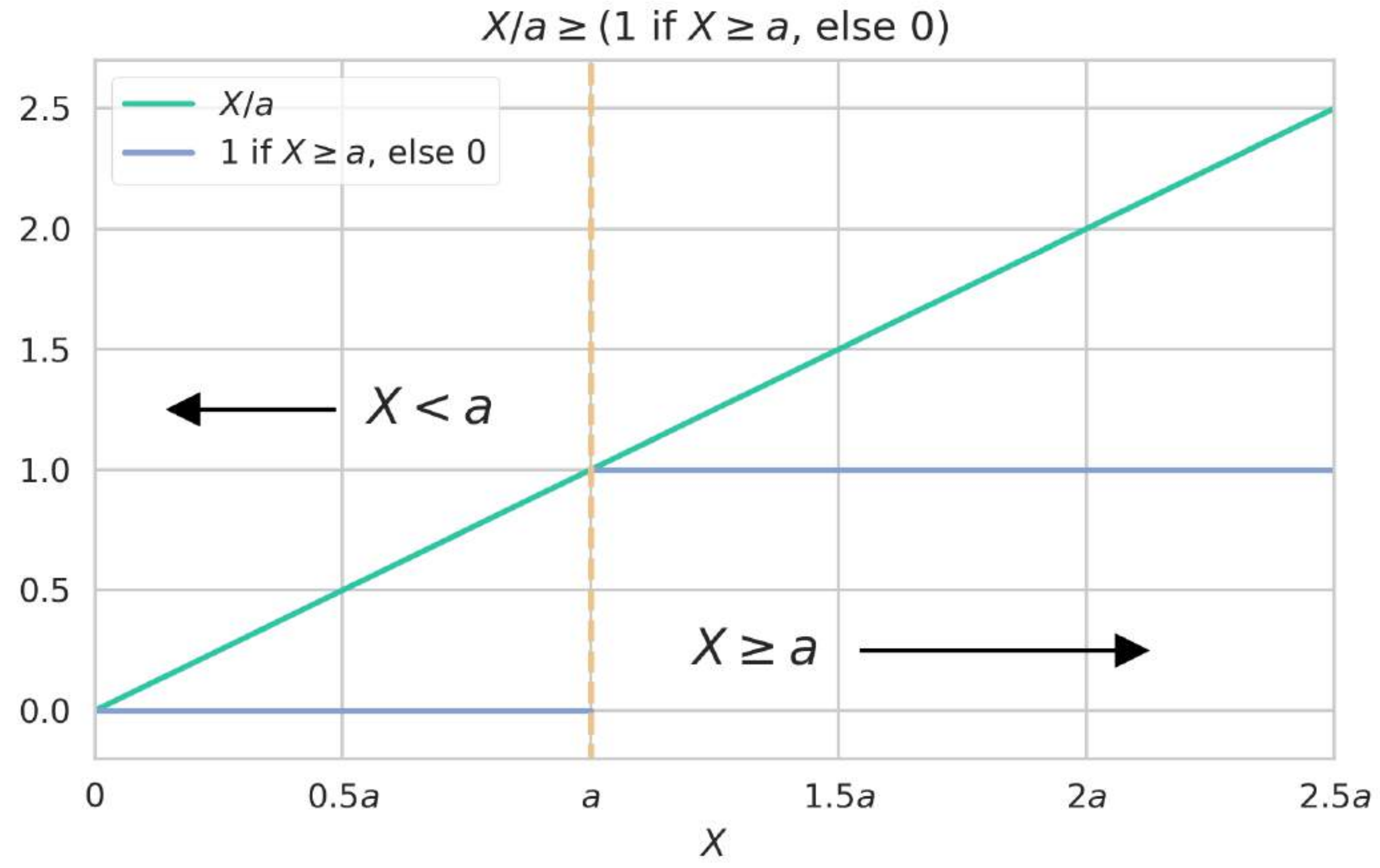
- <u>Markov’s inequality</u>: Let X be a *nonnegative-valued* random variable. Then,

$$\text{for any } a > 0, \qquad \Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

- **Proof** (by indicator): Let $I = I(X \geq a)$. Since $X \geq 0$ and $a > 0$, we have

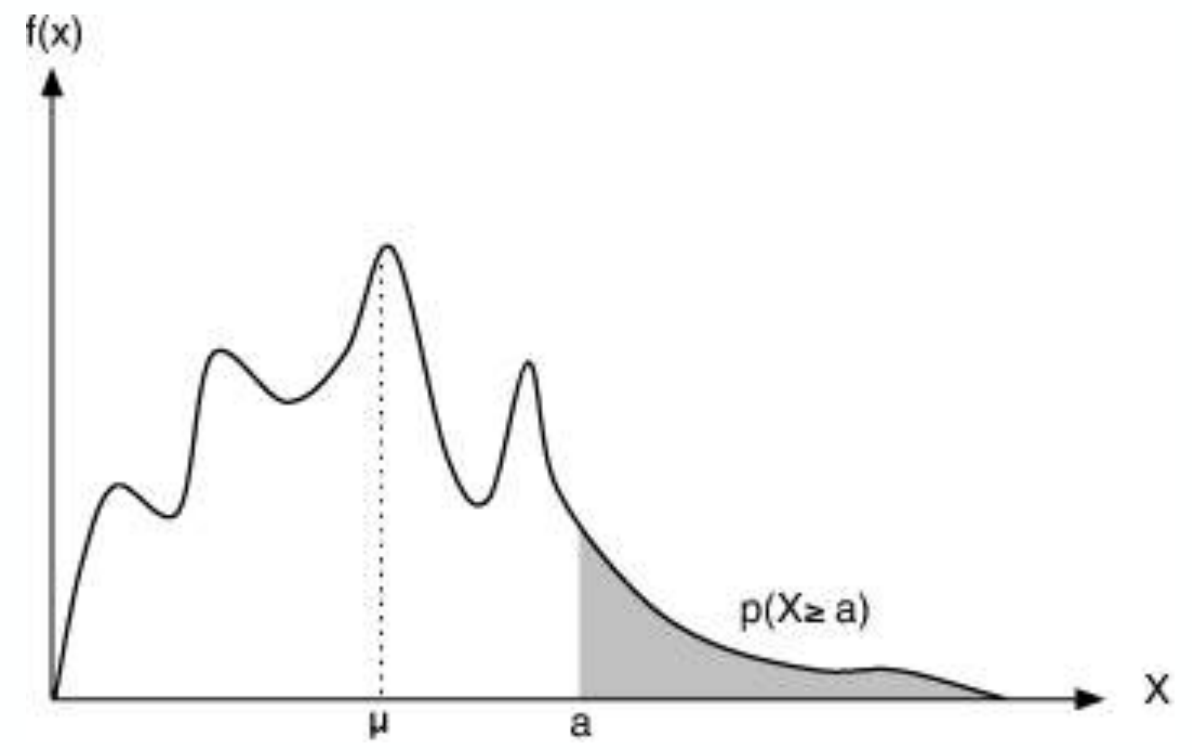
$$I = I(X \geq a) \leq \left\lfloor \frac{X}{a} \right\rfloor \leq \frac{X}{a}.$$

Therefore, $\Pr(X \geq a) = \mathbb{E}[I] \leq \mathbb{E}\left[\frac{X}{a}\right] = \frac{\mathbb{E}[X]}{a}$



Markov's Inequality

(马尔可夫不等式)



- <u>Markov's inequality</u>: Let X be a *nonnegative-valued* random variable. Then,

$$\text{for any } a > 0, \qquad \Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

- **Proof** (by total expectation):

($X \geq a$ is possible) (X is nonnegative)

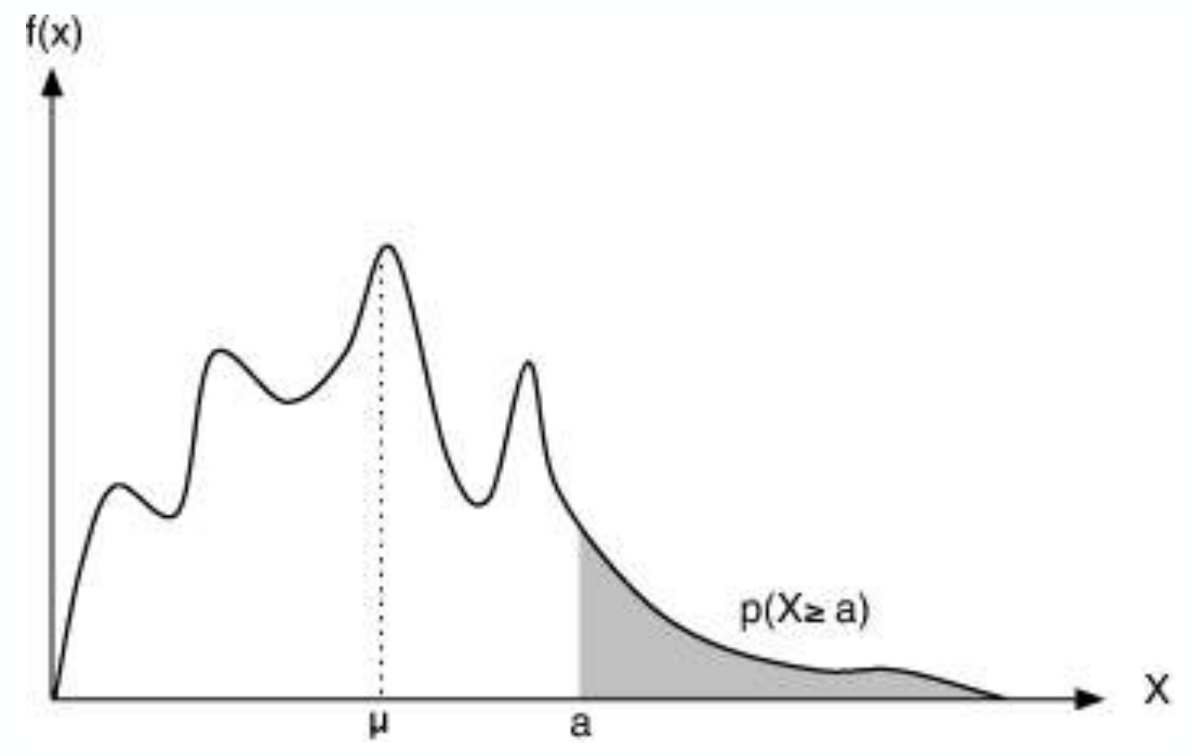
$$\mathbb{E}[X] = \mathbb{E}[X \mid X \geq a] \cdot \Pr(X \geq a) + \mathbb{E}[X \mid X < a] \cdot \Pr(X < a)$$

$$\geq a \cdot \Pr(X \geq a) + 0 \cdot \Pr(X < a) \qquad = a \cdot \Pr(X \geq a)$$

$$\Longrightarrow \quad \Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Markov's Inequality

(马尔可夫不等式)



- <u>Markov's inequality</u>: Let X be a *nonnegative-valued* random variable. Then,

$$\text{for any } a > 0, \qquad \Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

- **Corollary**: for any $c > 1, \qquad \Pr(X \geq c\mathbb{E}[X]) \leq 1/c$

- **Tight in the worst case:** $\forall c > 1, \forall \mu \in \mathbb{R}$, $\exists$ nonnegative X with $\mathbb{E}[X] = \mu$, such that $\Pr(X \geq c\mu) = 1/c$

- **Lower tail variant** (sometimes called *<u>reverse Markov's inequality</u>*):

  $\Pr(X \leq a) \leq (u - \mathbb{E}[X])/(u - a)$ requires X to have bounded range $X \leq u$

From Las Vegas to Monte Carlo

- 

<u>Monte Carlo algorithm</u>: randomized algorithms that are correct by chance

- 

<u>Las Vegas algorithm</u>: randomized algorithms that always give correct result upon termination (but may run for a random period of time before termination)

- If there is a Las Vegas algorithm $\mathscr{A}$ with expected running time at most $t(n)$ for any input of size n ($\mathscr{A}$ has worst-case expected time complexity $t(n)$):

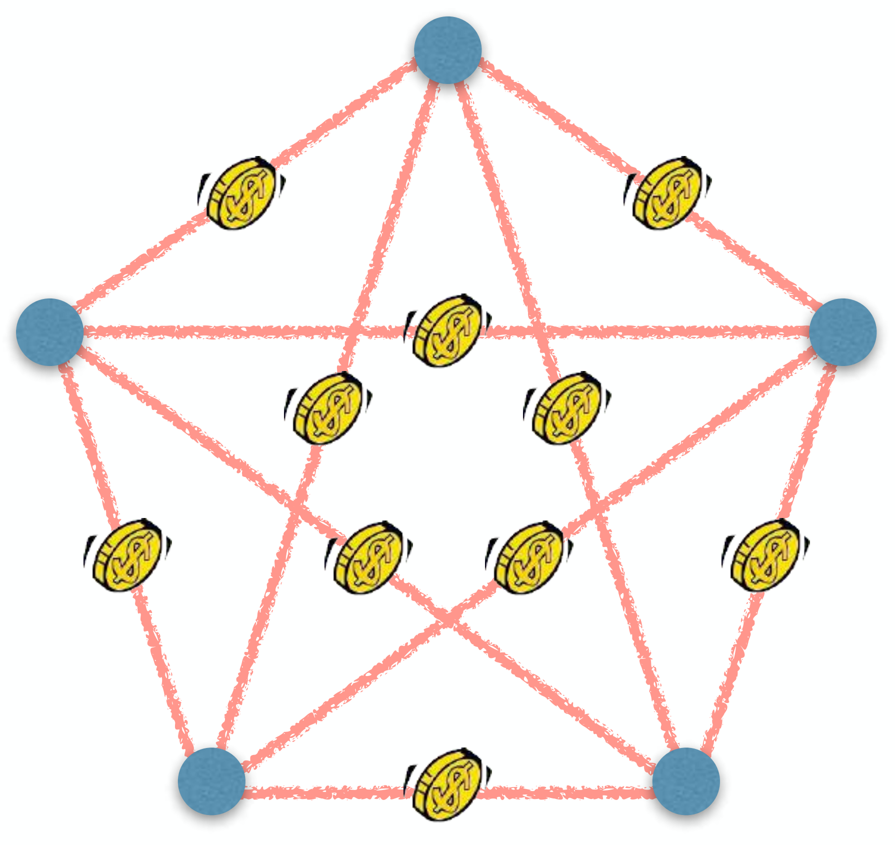
Algorithm $\mathscr{B}$:

simulate algorithm $\mathscr{A}$ up to $\lceil t(n)/\epsilon \rceil$ steps;
if algorithm $\mathscr{A}$ terminates
  return the output of $\mathscr{A}$;
else return an arbitrary answer;

- Algorithm $\mathscr{B}$ is a Monte Carlo algorithm s.t.
  - $\mathscr{B}$ has worst-case running time $\leq \lceil t(n)/\epsilon \rceil$
  - $\mathscr{B}$ is correct with probability at least $1 - \epsilon$
    (by Markov inequality)

Cliques in Random Graph

- $G(n, p)$: between every pair u, v among n vertices, an edge is added i.i.d. with prob. p
- Fix a constant integer $k \geq 3$. Let X be the number of k-**cliques** (K_k) in $G \sim G(n, p)$.
- For every distinct $S \subseteq [n]$ of size $|S| = k$, let $I_S = I(K_S \subseteq G)$. Then:
    - $\mathbb{E}[I_S] = \Pr(K_S \subseteq G) = p^{\binom{k}{2}}$
    - $X = \sum_{S \in \binom{[n]}{k}} I_S$



- **Linearity of expectation:** $\mathbb{E}[X] = \binom{n}{k} p^{\binom{k}{2}} \leq n^k p^{k(k-1)/2} = o(1)$ for $p = o\left(n^{-2/(k-1)}\right)$
- **Markov's inequality:** $\Pr(X \geq 1) \leq \mathbb{E}[X] = o(1) \Longrightarrow \Pr(X = 0) = 1 - o(1)$

  $\Longrightarrow$ If $p = o\left(n^{-2/(k-1)}\right)$, then $G(n, p)$ is K_k-free **a.a.s.** (asymptotically almost surely)

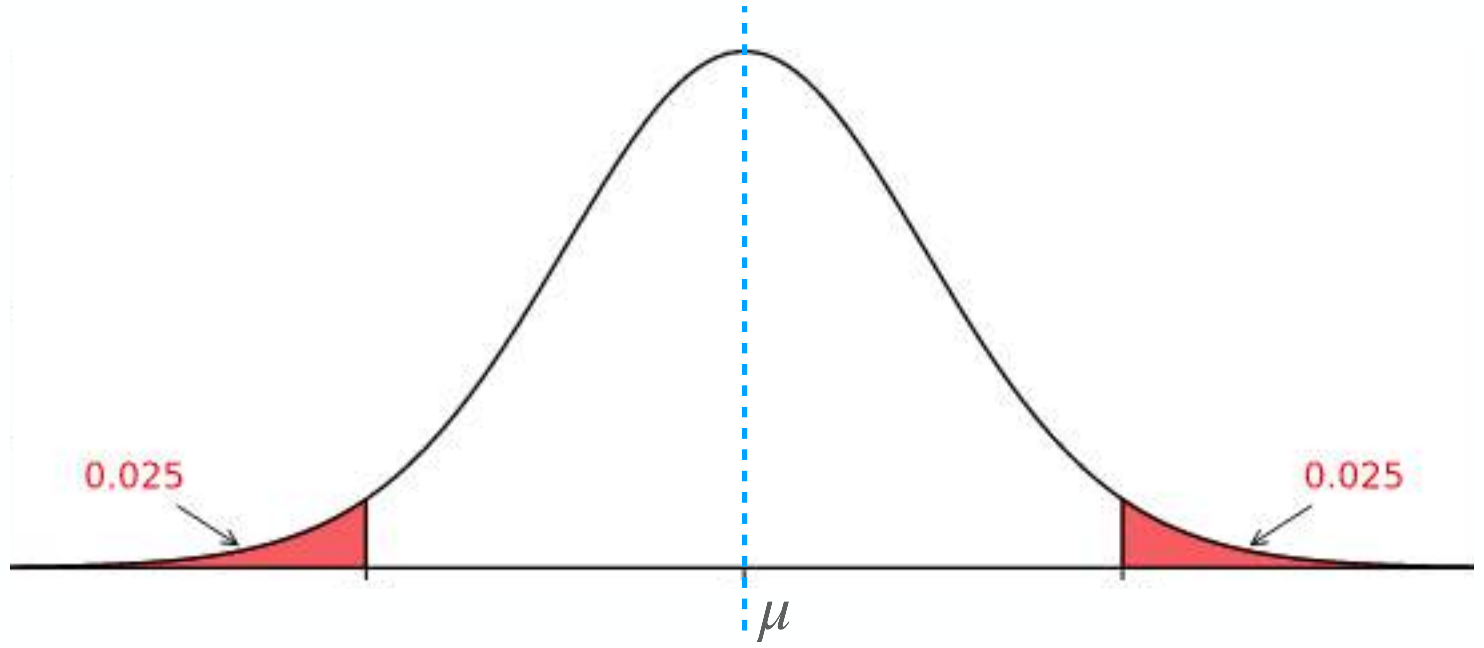
Generalized Markov's Inequality

- Let X be a random variable and $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ a nonnegative-valued function.

$$\text{For any } a > 0, \qquad \Pr(f(X) \geq a) \leq \frac{\mathbb{E}[f(X)]}{a}$$

- **Proof**: Apply the Markov's inequality to the random variable $Y = f(X)$.
- **Applications**: useful if $f(X)$ can "*extract*" useful information about X
  - **Chebyshev's inequality, kth moment method:** $f(X)$ extracts the kth moment
  - **Chernoff-Hoeffding bounds, Bernstein inequalities:** $f(X)$ extracts all moments

Deviation Inequality



- Let X be a random variable with **mean** $\mu = \mathbb{E}[X]$. For $a > 0$

$$\Pr(|X - \mu| \geq a) \leq \text{?}$$

- Applying **Markov's inequality** to $Y = |X - \mu|$ gives us

$$\Pr(|X - \mu| \geq a) \leq \frac{\mathbb{E}[|X - \mu|]}{a}$$

difficult to calculate

- Alternatively, we may apply **Markov's inequality** to $Y = (X - \mu)^2$

$$\Pr(|X - \mu| \geq a) = \Pr((X - \mu)^2 \geq a^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{a^2}$$

Variance

(2nd central moment)

Variance (方差) and Moments (矩)

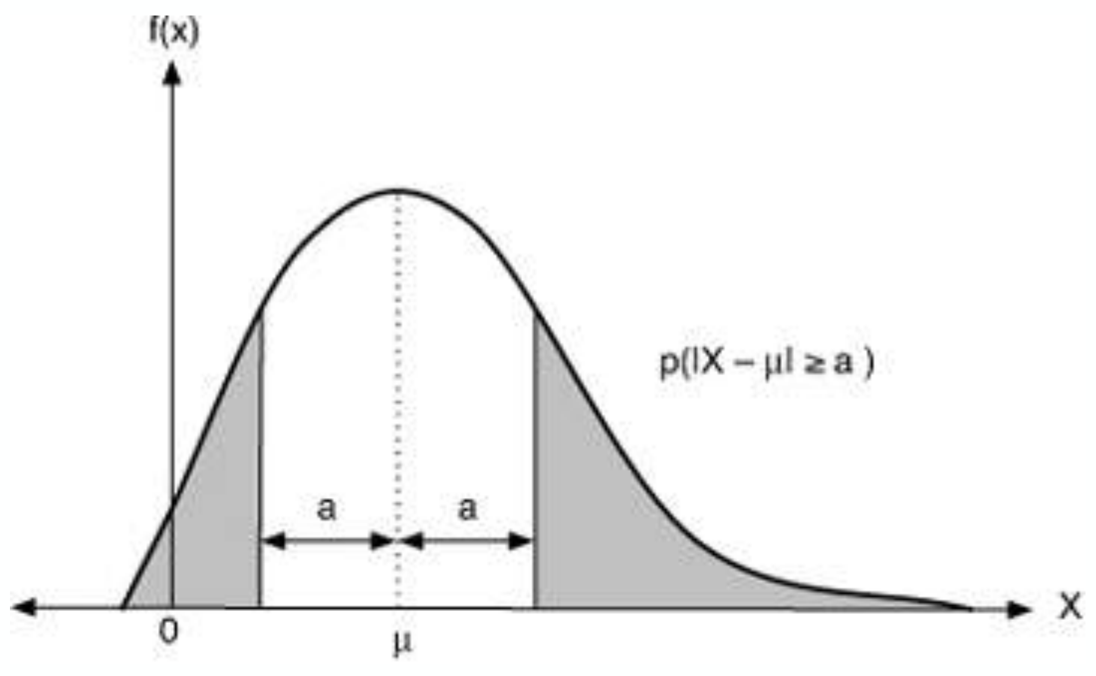
- For integer $k > 0$, the kth moment (k阶矩) of a random variable X is $\mathbb{E}[X^k]$, and the kth central moment (k阶中心矩) of X is $\mathbb{E}[(X - \mathbb{E}[X])^k]$.
- Sometimes, a random variable X is called **centralized** (中心化的) if $\mathbb{E}[X] = 0$. A random variable X can be centralized by $Y = X - \mathbb{E}[X]$.
- The variance (方差) of a random variable X is its 2nd central moment:

$$\mathbf{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

and the **standard deviation** (标准差) of X is $\sigma = \sigma[X] = \sqrt{\mathbf{Var}[X]}$

Chebyshev's Inequality

(切比雪夫不等式, the second Chebyshev inequality)



- <u>Chebyshev's inequality</u>: Let X be a random variable. For any $a > 0$,

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbf{Var}[X]}{a^2}$$

- **Proof**: Apply Markov's inequality to $Y = (X - \mathbb{E}[X])^2$.

- **Corollary**: For standard deviation $\sigma = \sqrt{\mathbf{Var}[X]}$, for any $k \geq 1$,

$$\Pr(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2}$$

Median and Mean

- The **median** (中位数) of random variable X is defined to be any value m s.t.:

$$\Pr(X \le m) \ge 1/2 \quad \text{and} \quad \Pr(X \ge m) \ge 1/2$$

- The expectation $\mu = \mathbb{E}[X]$ is the value that minimizes

$$\mathbb{E}[(X-\mu)^2]$$

- **Proof**: $f(x) = \mathbb{E}[(X-x)^2] = \mathbb{E}[X^2] - 2x\mathbb{E}[X] + x^2$ is convex and has $f'(\mu) = 0$

- The median m is the value that minimizes

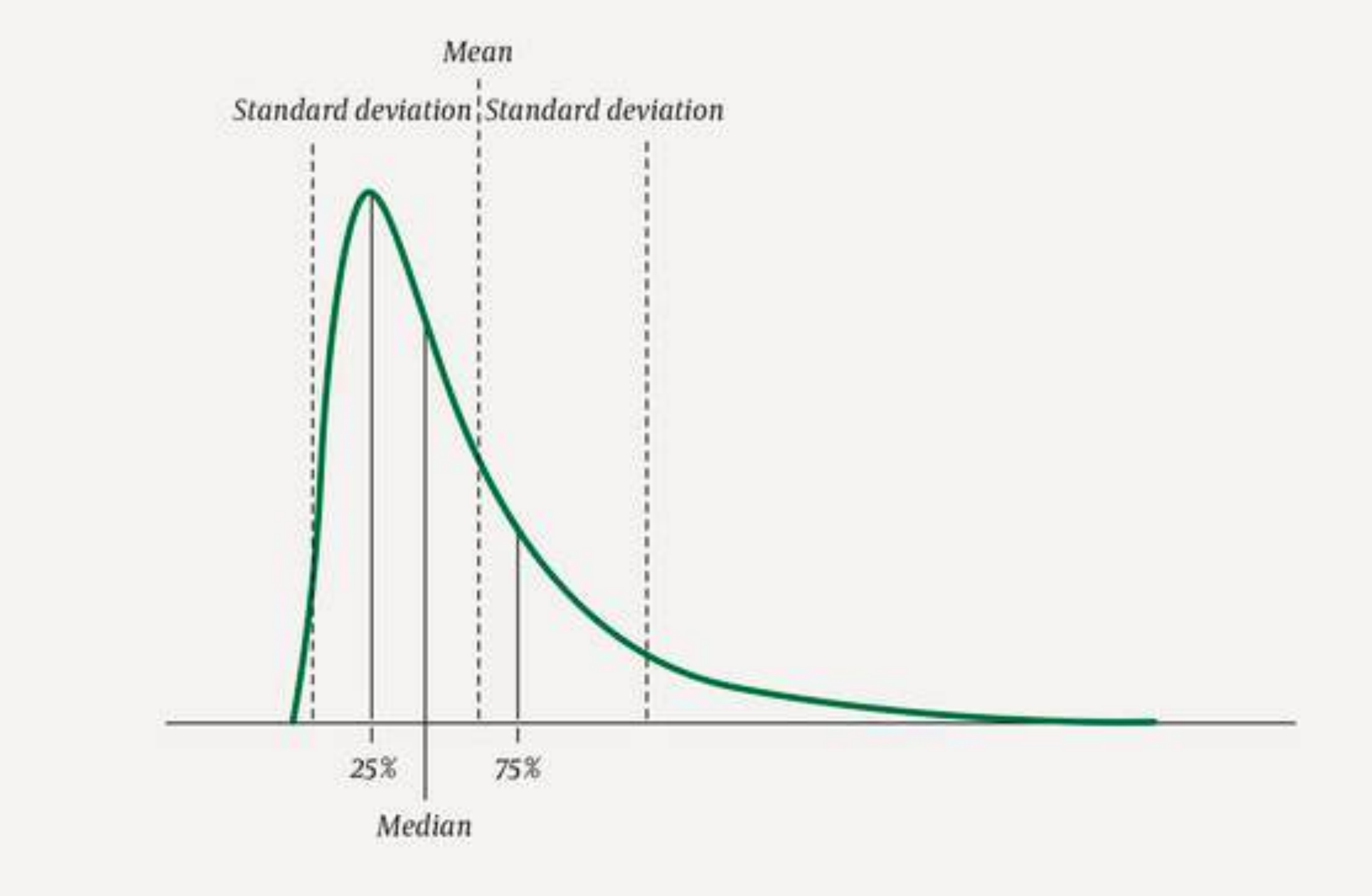
$$\mathbb{E}[\,|X-m|\,]$$

- **Proof**: By symmetry, suppose non-median $y > m$ so that $\Pr(X \ge y) < 1/2$.

$$\mathbb{E}[\,|X-y| - |X-m|\,] = (m-y)\Pr(X \ge y) + \sum_{m<x<y}(m+y-2x)\Pr(X=x) + (y-m)\Pr(X \le m)$$

$$> (m-y)/2 + (y-m)/2 = 0$$

Median and Mean



- If X is a random variable with finite expectation μ, median m, and standard deviation σ, then

$$|\mu - m| \le \sigma$$

- **Proof**: $|\mu - m| = |\mathbb{E}[X] - m| = |\mathbb{E}[X - m]|$

$$\le \mathbb{E}[\,|X - m|\,] \quad \text{(Jensen's inequality)}$$

$$\le \mathbb{E}[\,|X - \mu|\,] \quad \text{(the median } m \text{ minimizes } \mathbb{E}[\,|X - m|\,]\text{)}$$

$$= \mathbb{E}\left[\sqrt{(X-\mu)^2}\right] \le \sqrt{\mathbb{E}\left[(X-\mu)^2\right]} = \sigma \quad \text{(Jensen's inequality)}$$

Variance

Calculation of Variance

$$\mathbf{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- **Proof**: $\mathbf{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$

$$= \mathbb{E}\left[X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2\right]$$

$$= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- X is constant ***a.s.*** $(\Pr(X = \mathbb{E}[X]) = 1) \iff \mathbb{E}[X^2] = \mathbb{E}[X]^2 \iff \mathbf{Var}[X] = 0$

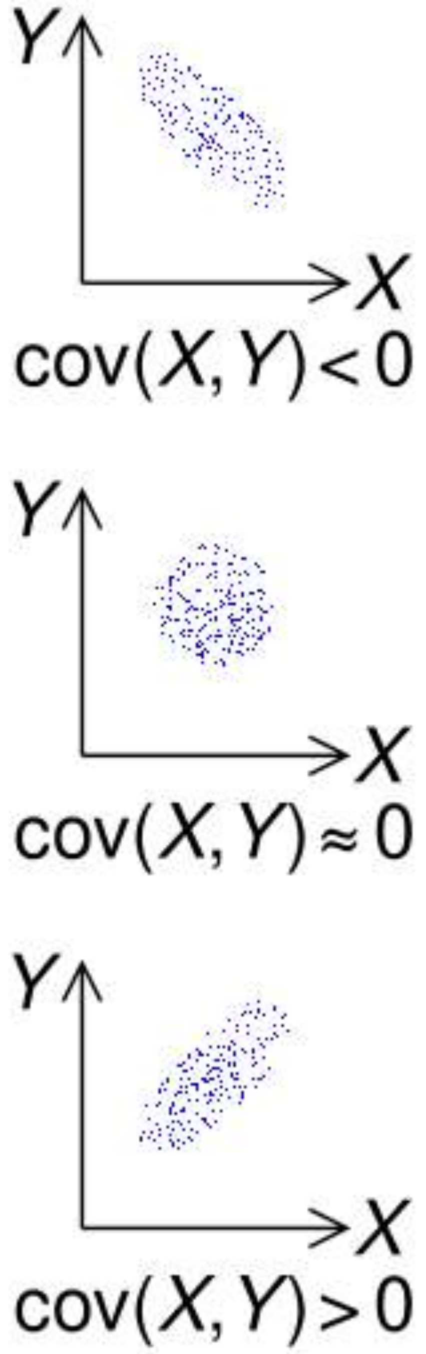
Variance of Linear Function

- For random variables X, Y and real number $a \in \mathbb{R}$:
  - $\mathbf{Var}[a] = 0$
  - $\mathbf{Var}[X + a] = \mathbf{Var}[X]$ (variance is a central moment)
  - $\mathbf{Var}[aX] = a^2\mathbf{Var}[X]$ (variance is quadratic)
  - $\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y])$
- **Proof**: All can be verified through $\mathbf{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Covariance (协方差)

- The covariance (协方差) of two random variables X and Y is

$$\mathbf{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$



- **Properties**: $\mathbf{Var}[X] = \mathbf{Cov}(X,X)$
  - *Symmetric*: $\mathbf{Cov}(X,Y) = \mathbf{Cov}(Y,X)$
  - *Distributive*: $\mathbf{Cov}(X+Y,Z) = \mathbf{Cov}(X,Z) + \mathbf{Cov}(Y,Z)$

$$\mathbf{Cov}(aX,Y) = a\mathbf{Cov}(X,Y)$$

- If X and Y are independent then

$$\mathbf{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$$

Covariance of Independent Variables

- If random variables X and Y are **independent**, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

- If random variables $X_1, X_2, \ldots, X_n$ are **mutually independent**, then

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \mathbb{E}\left[\prod_{i=1}^{n-1} X_i\right] \cdot \mathbb{E}[X_n] = \prod_{i=1}^{n} \mathbb{E}[X_i]$$

Proof: By change of variable (*LOTUS*)

$$\mathbb{E}[XY] = \sum_{x,y} xy \Pr(X = x \cap Y = y) = \sum_{x,y} xy \Pr(X = x) \Pr(Y = y)$$

$$= \left(\sum_{x} x \Pr(X = x)\right)\left(\sum_{y} y \Pr(Y = y)\right) = \mathbb{E}[X]\mathbb{E}[Y]$$

Expectation of Product

- For random variables X and Y:

if X and Y independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

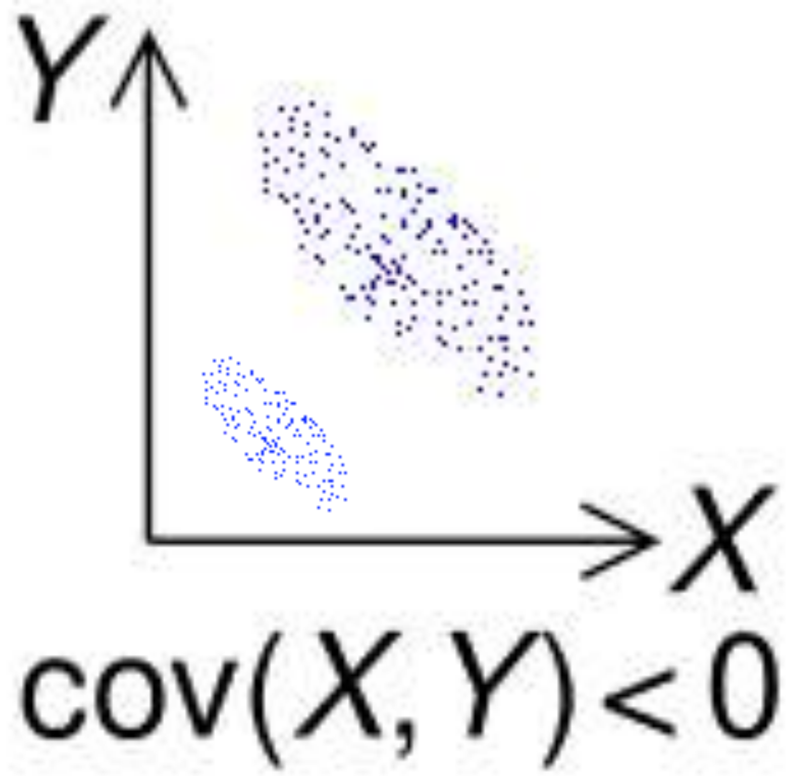
- **(Cauchy-Schwarz)**

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

- **(Hölder)** for any $p, q > 0$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$

$$\mathbb{E}[XY] \leq \mathbb{E}[\,|X|^p\,]^{1/p}\mathbb{E}[\,|Y|^q\,]^{1/q}$$

Correlation (相关性)



- The covariance (协方差) of two random variables X and Y is

$$\mathbf{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- The correlation coefficient (相关系数) of X and Y is

$$\rho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}[X] \cdot \mathbf{Var}[Y]}} \quad \in [-1,1]$$

by Cauchy-Schwarz

- Two random variables X and Y are called uncorrelated if $\mathbf{Cov}(X,Y) = 0$

- X and Y are **uncorrelated** means:
  - $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
  - $\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$

Variance of Sum

- For random variables X, Y:

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X, Y)$$

- For random variables $X_1, X_2, \ldots, X_n$:

$$\mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{Var}[X_i] + \sum_{i \neq j} \mathbf{Cov}(X_i, X_j)$$

- For pairwise independent $X_1, X_2, \ldots, X_n$:

$$\mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{Var}[X_i]$$

Variance of Indicator

- For **Bernoulli random variable** $X \in \{0,1\}$ with parameter p

$$X^2 = X \implies \mathbb{E}[X^2] = \mathbb{E}[X] = p$$

$$\mathbf{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$$

- For the **indicator** random variable $X = I(A)$ of event A:

$$\mathbf{Var}[X] = \Pr(A)(1 - \Pr(A)) = \Pr(A)\Pr(A^c)$$

Variance of Discrete Uniform Distribution

- For integers $a \leq b$, let X be chosen from $[a, b] = \{a, a+1, \ldots, b\}$ **u.a.r.**
- $$\mathbb{E}[X] = \sum_{k=a}^{b} \frac{k}{b-a+1} = \frac{a+b}{2}$$
- $$\mathbb{E}[X^2] = \sum_{k=a}^{b} \frac{k^2}{b-a+1} = \frac{2b^2 + 2ab + 2a^2 + b - a}{6}$$
- $$\mathbf{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(b-a)(b-a+2)}{12}$$

Geometric Distribution (几何分布)

- For geometric random variable $X \sim \text{Geo}(p)$, recall $\mathbb{E}[X] = 1/p$, and

$$\mathbb{E}[X^2] = \sum_{k\geq 1} k^2(1-p)^{k-1}p = (2-p)p^{-2}$$

$$\textbf{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = (2-p)p^{-2} - p^{-2} = (1-p)/p^2$$

- **Total expectation:** $\mathbb{E}[X^2] = \mathbb{E}[X^2 \mid X > 1] \cdot (1-p) + \mathbb{E}[X^2 \mid X = 1] \cdot p$

$$= \mathbb{E}[((X-1)+1)^2 \mid X > 1] \cdot (1-p) + p$$

(memoryless) $$= \mathbb{E}[(X+1)^2] \cdot (1-p) + p$$

$$= (1-p)\mathbb{E}[X^2] + 2(1-p)/p + 1$$

$$\implies \mathbb{E}[X^2] = (2-p)/p^2 \implies \textbf{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = (1-p)/p^2$$

Binomial Distribution (二项分布)

- For binomial random variable $X \sim \text{Bin}(n, p)$, recall $\mathbb{E}[X] = np$, and

$$\mathbf{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} - (np)^2$$

- **Observation**: $X \sim \text{Bin}(n, p)$ can be expressed as $X = X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are i.i.d. Bernoulli random variables with parameter p

- For mutually independent $X_1, \ldots, X_n$:

$$\mathbf{Var}[X] = \sum_{i=1}^{n} \mathbf{Var}[X_i] = np(1-p)$$

Poisson Distribution

- For **Poisson random variable** $X \sim \text{Pois}(\lambda)$, recall $\mathbb{E}[X] = \lambda$, and

$$\mathbb{E}[X^2] = \sum_{k \geq 0} k^2 \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k \geq 1} k \frac{e^{-\lambda}\lambda^k}{(k-1)!}$$

$$= \sum_{k \geq 0} (k+1) \frac{e^{-\lambda}\lambda^{k+1}}{k!} = \lambda \sum_{k \geq 0} (k+1) \frac{e^{-\lambda}\lambda^k}{k!}$$

$$= \lambda \mathbb{E}[X+1] = \lambda(\mathbb{E}[X] + 1) = \lambda(\lambda + 1)$$

$$\mathbf{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda$$

Negative Binomial Distribution (负二项分布)

- For negative binomial random variable X with parameters r, p

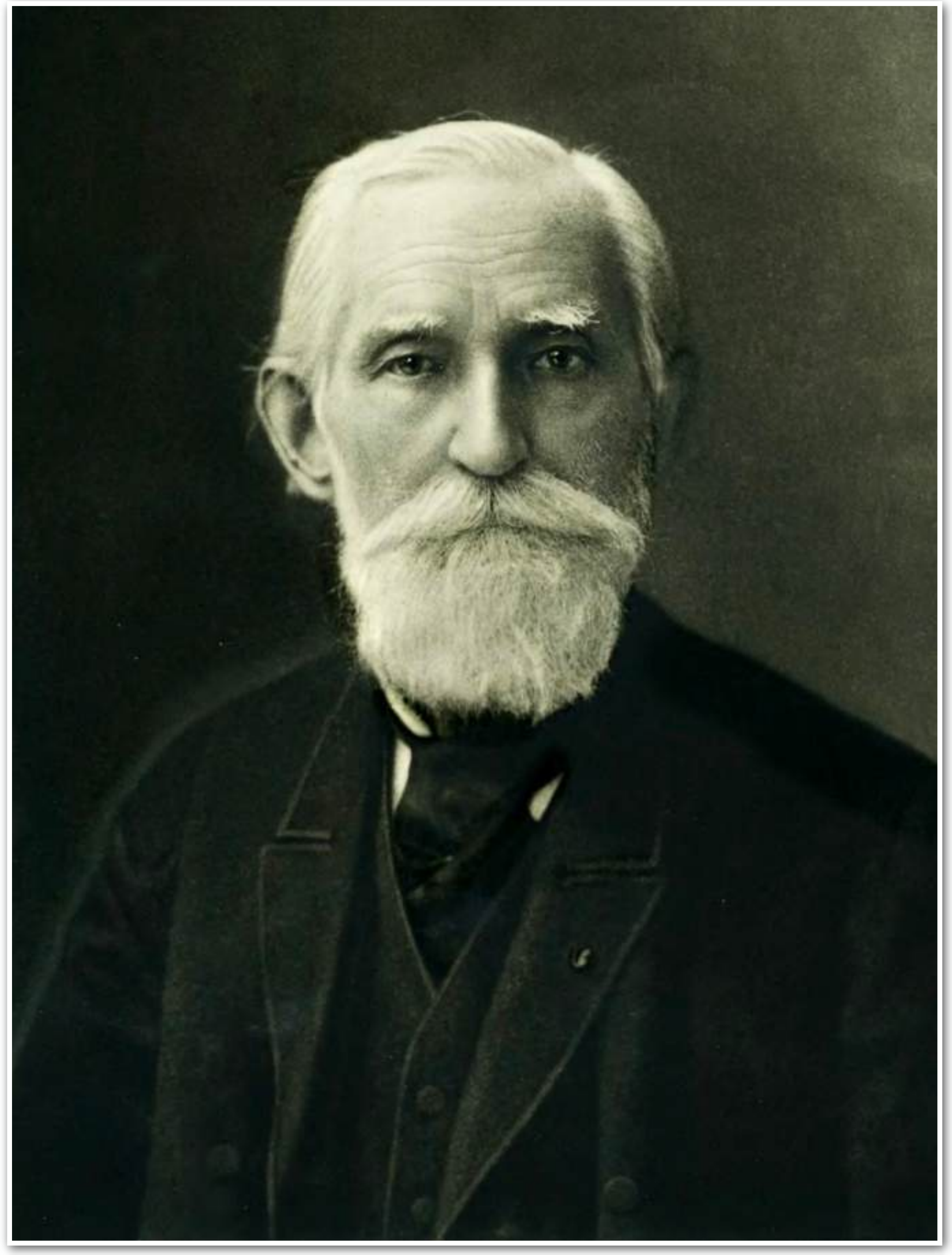
$$\mathbf{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sum_{k\geq 1} k^2 \binom{k+r-1}{k} (1-p)^k p^r - r^2(1-p)^2/p^2$$

- **Observation**: X can be expressed as $X = (X_1 - 1) + \cdots + (X_r - 1)$, where $X_1, \ldots, X_r$ are i.i.d. geometric random variables with parameter p

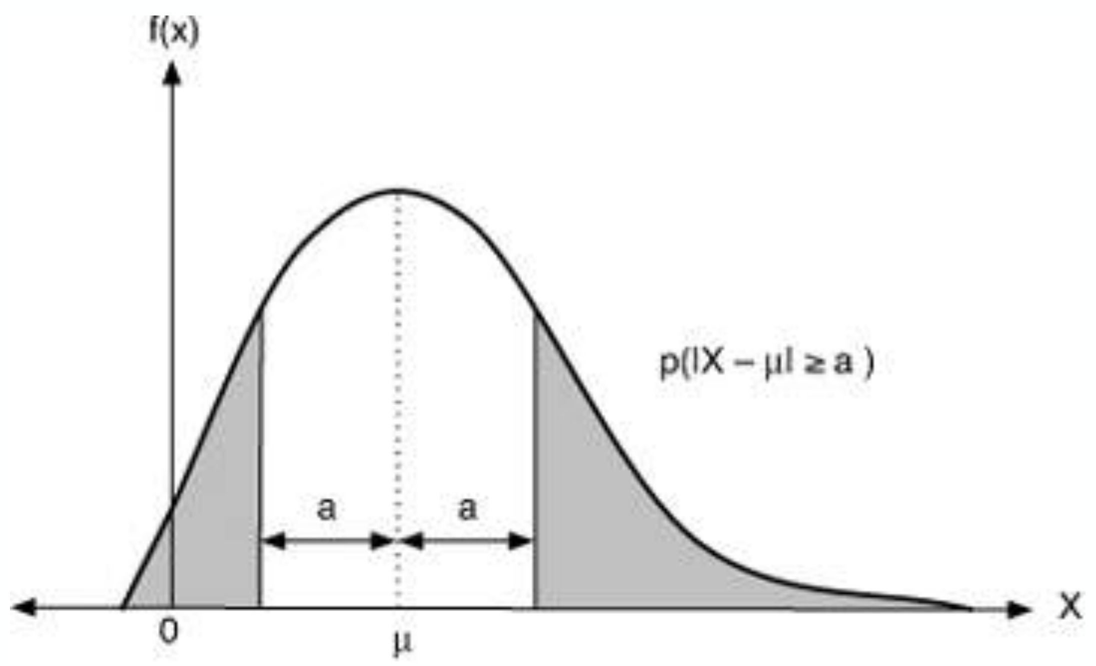
- For mutually independent $X_1, \ldots, X_r$:

$$\mathbf{Var}[X] = \sum_{i=1}^{r} \mathbf{Var}[X_i - 1] = \sum_{i=1}^{r} \mathbf{Var}[X_i] = \frac{r(1-p)}{p^2}$$

Chebyshev (Чебышёв)'s Inequality



Chebyshev's Inequality
(切比雪夫不等式)



- Chebyshev's inequality: Let X be a random variable. For any $a > 0$,

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbf{Var}[X]}{a^2}$$

- **Corollary**: For standard deviation $\sigma = \sqrt{\mathbf{Var}[X]}$, for any $k \geq 1$,

$$\Pr(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2}$$

- **Tight in the worst case:** $\forall k \geq 1$, $\forall \mu \in \mathbb{R}$ and $\forall \sigma > 0$, $\exists X$ with $\mathbb{E}[X] = \mu$ and $\mathbf{Var}[X] = \sigma^2$ such that $\Pr(|X - \mu| \geq k\sigma) = 1/k^2$

Unbiased Estimator (mean trick)

- Let $X_1, \ldots, X_n$ be *i.i.d.* random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbf{Var}[X_i] = \sigma^2$.
- Empirical mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\mathbb{E}[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu \quad \text{and} \quad \mathbf{Var}[\overline{X}] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{Var}[X_i] = \frac{\sigma^2}{n}$$

- **Chebyshev's inequality:**

$$\Pr(|\overline{X} - \mu| \geq \epsilon\mu) \leq \frac{\mathbf{Var}[\overline{X}]}{\epsilon^2\mu^2} = \frac{\sigma^2}{\epsilon^2\mu^2 n} \quad \leq \delta \quad \text{if } n \geq \frac{\sigma^2}{\epsilon^2\mu^2\delta}$$

(one-sided) Error Reduction

- Decision problem $f : \{0,1\}^* \to \{0,1\}$.
- Monte Carlo randomized algorithm $\mathscr{A}$ with ***one-sided*** error:

  for any input x and uniform ***random seed*** $r \in [p]$ for some prime number p
  - $f(x) = 1 \implies \Pr_{r\in[p]}\left(\mathscr{A}(x, r) = 1\right) \geq \epsilon$
  - $f(x) = 0 \implies \mathscr{A}(x, r) = 0$ for all $r \in [p]$
- $\mathscr{A}^k(x, r_1, \ldots, r_k) = \vee_{i=1}^k \mathscr{A}(x, r_i)$: for mutually independent $r_1, \ldots, r_k \in [p]$
  - $f(x) = 1 \implies \Pr\left(\mathscr{A}^k(x, r_1, \ldots, r_k) = 0\right) \leq (1 - \epsilon)^k$

Two-Point Sampling (2-Universal Hashing)

- Let $p > 1$ be a prime number and $[p] = \{0,1,\ldots,p-1\} = \mathbb{Z}_p$.
- Pick $\boldsymbol{a}, \boldsymbol{b} \in [p]$ *u.a.r.* and let $r_i = (\boldsymbol{a} \cdot i + \boldsymbol{b}) \bmod p$ for $i = 1,2,\ldots,p$
  - $r_1, \ldots, r_p \in [p]$ are <u>pairwise independent</u>
  - each r_i is <u>uniformly distributed</u> over $[p]$
- **Proof**: For any $i \neq j$, $\forall c, d \in [p]$, $\Pr(r_i = c \cap r_j = d) = 1/p^2$ because

$$\begin{cases} \boldsymbol{a} \cdot i + \boldsymbol{b} \equiv c \pmod{p} \\ \boldsymbol{a} \cdot j + \boldsymbol{b} \equiv d \pmod{p} \end{cases}$$ has a unique solution $(a, b) \in [p]^2$

$$\Pr(r_i = c) = \Pr(\boldsymbol{a} \cdot i + \boldsymbol{b} \equiv c \pmod{p}) = \frac{1}{p} \sum_{a \in [p]} \Pr(\boldsymbol{b} \equiv c - ai \pmod{p}) = \frac{1}{p}$$

Derandomization with Two-Point Sampling

- $\mathscr{A}$: for any input x and uniform ***random seed*** $r \in [p]$ for prime number p
  - $f(x) = 1 \implies \Pr\left(\mathscr{A}(x, r) = 1\right) \geq \epsilon$
  - $f(x) = 0 \implies \mathscr{A}(x, r) = 0$ for all $r \in [p]$
- $\mathscr{A}^k(x, r_1, \ldots, r_k) = \vee_{i=1}^k \mathscr{A}(x, r_i)$: $k \leq p$ for $r_i = (\boldsymbol{a} \cdot i + \boldsymbol{b}) \bmod p$ with uniform $\boldsymbol{a}, \boldsymbol{b} \in [p]$
  - If $f(x) = 0 \implies \mathscr{A}^k(x, r_1, \ldots, r_k) = \vee_{i=1}^k \mathscr{A}(x, r_i) = 0$
  - If $f(x) = 1 \implies \Pr\left(\mathscr{A}(x, r_i) = 1\right) \geq \epsilon$ because each r_i is uniform over $[p]$
  - Let $X_i = \mathscr{A}(x, r_i)$ and let $X = \sum_{i=1}^k X_i$.
    - $X_1, \ldots, X_k$ are **pairwise independent** Bernoulli random variables with $\Pr(X_i = 1) \geq \epsilon$
    - $\Pr\left(\mathscr{A}^k(x, r_1, \ldots, r_k) = 0\right) = \Pr(X = 0) \leq \Pr\left(|X - \mathbb{E}[X]| \geq \mathbb{E}[X]\right) \leq \dfrac{\mathbf{Var}[X]}{\mathbb{E}[X]^2}$

(Chebyshev's inequality)

Derandomization with Two-Point Sampling

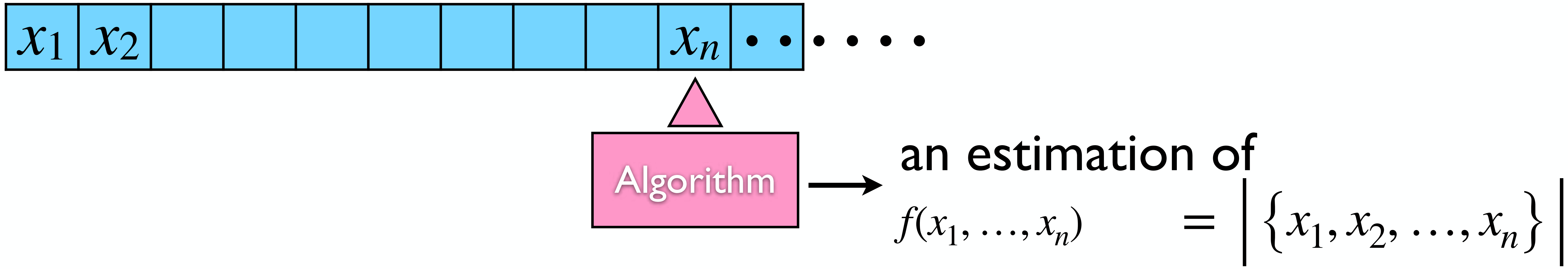
- $\mathscr{A}^k(x, r_1, \ldots, r_k) = \vee_{i=1}^k \mathscr{A}(x, r_i)$: $k \le p$ and $r_i = (\boldsymbol{a} \cdot i + \boldsymbol{b}) \bmod p$ with uniform $\boldsymbol{a}, \boldsymbol{b} \in [p]$
  - If $f(x) = 1 \implies \Pr\left(\mathscr{A}(x, r_i) = 1\right) \ge \epsilon$ because each r_i is uniform over $[p]$
  - Let $X_i = \mathscr{A}(x, r_i)$ and let $X = \sum_{i=1}^k X_i$.
    - $X_1, \ldots, X_k$ are **pairwise independent** Bernoulli random variables with $\Pr(X_i = 1) \ge \epsilon$
    - $\Pr\left(\mathscr{A}^k(x, r_1, \ldots, r_k) = 0\right) = \Pr(X = 0) \le \Pr\left(|X - \mathbb{E}[X]| \ge \mathbb{E}[X]\right) \le \dfrac{\mathbf{Var}[X]}{\mathbb{E}[X]^2} \le \dfrac{1}{\epsilon k}$
      - **Linearity of expectation:** $\mathbb{E}[X] = \sum_{i=1}^k \mathbb{E}[X_i] \ge \epsilon k$
      - **Pairwise independence:** $\mathbf{Var}[X] = \sum_{i=1}^{k} \mathbf{Var}[X_i] \le \sum_{i=1}^{k} \mathbb{E}[X_i^2] = \sum_{i=1}^{k} \mathbb{E}[X_i] = \mathbb{E}[X]$
- Reduce any 1-sided error $1 - \epsilon$ to $1/(\epsilon k)$ with $k \le p$ runs of the algorithm using only **2 random seeds** in total.

Count Distinct Elements

Input: a sequence $x_1, x_2, \ldots, x_n \in U = [N]$

Output: an estimation of $z = \left|\{x_1, x_2, \ldots, x_n\}\right|$

- **Data stream model**: input data item comes one at a time



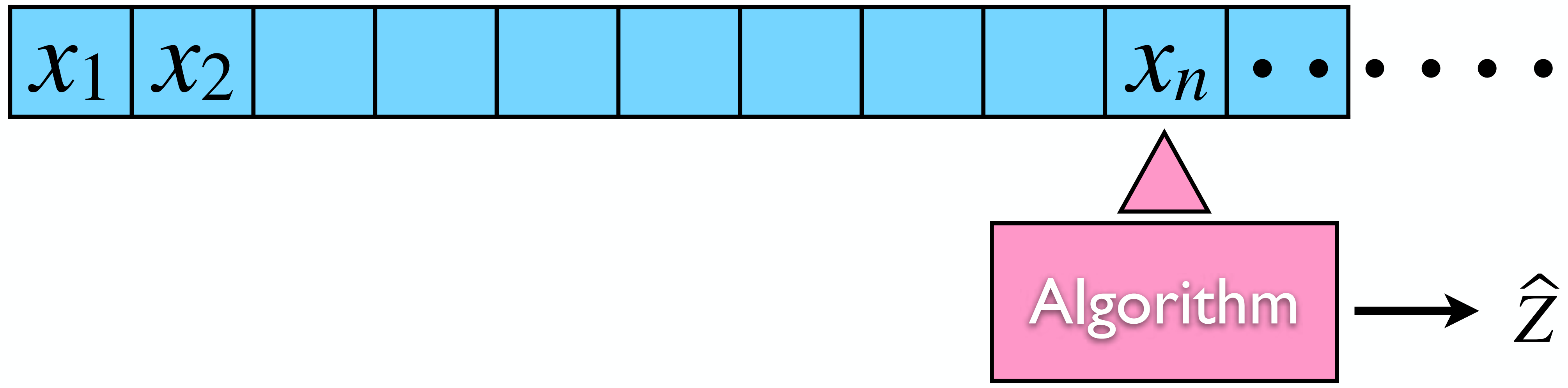
- **Naïve algorithm**: store all distinct data items using $\Omega(z \log N)$ bits
- **Sketch**: (lossy) representation of data using space $\ll z$
- **Lower bound** (Alon-Matias-Szegedy): any deterministic (exact or approx.) algorithm must use $\Omega(N)$ bits of space in the worst case

Count Distinct Elements

Input: a sequence $x_1, x_2, \ldots, x_n \in U = [N]$

Output: an estimation of $z = \left|\{x_1, x_2, \ldots, x_n\}\right|$

- **Data stream model**: input data item comes one at a time



- **(ϵ, δ)-estimator**: randomized variable $\hat{Z}$

$$\Pr\left[(1-\epsilon)z \le \hat{Z} \le (1+\epsilon)z\right] \ge 1-\delta$$

Using only memory equivalent to 5 lines of printed text, you can estimate with a typical accuracy of 5% and in a single pass the total vocabulary of Shakespeare.

——Durand and Flajolet 2003

William Shakespeare

Input: a sequence $x_1, x_2, \ldots, x_n \in U = [N]$

Output: an estimation of $z = \left| \{x_1, x_2, \ldots, x_n\} \right|$

Simple Uniform Hash Assumption (SUHA):

A uniform function is available, whose preprocessing, representation and evaluation are considered to be easy.

- (*idealized*) uniform hash function $h : U \to [0,1]$
  - $x_i = x_j \longrightarrow$ the same hash value $h(x_i) = h(x_j) \in_r [0,1]$
- $\{h(x_1), \ldots, h(x_n)\}$: $z \times$ uniform and independent values in $[0,1]$
  - partition $[0,1]$ into $z+1$ subintervals (with *identically distributed* lengths)

$$\mathbb{E}\left[\min_{1\le i\le n} h(x_i)\right] = \mathbb{E}[\text{length of a subinterval}] = \frac{1}{z+1} \quad \text{(by symmetry)}$$

- estimator: $\widehat{Z} = \dfrac{1}{\min_i h(x_i)} - 1$? Variance is too large!

Input: a sequence $x_1, x_2, \ldots, x_n \in U = [N]$

Output: an estimation of $z = \left|\{x_1, x_2, \ldots, x_n\}\right|$

- (*idealized*) uniform hash function $h : U \to [0,1]$

Min Sketch:

let $Y = \min_{1 \le i \le n} h(x_i)$;

return $\widehat{Z} = \frac{1}{Y} - 1$;

- By symmetry:

$$\mathbb{E}[Y] = \frac{1}{z+1}$$

- Goal:

$$\Pr\left[\widehat{Z} < (1-\epsilon)z \text{ or } \widehat{Z} > (1+\epsilon)z\right] \le \delta$$

assuming $\epsilon \le 1/2$

$$\left|Y - \mathbb{E}[Y]\right| > \frac{\epsilon/2}{z+1} \iff \left|Y - \frac{1}{z+1}\right| > \frac{\epsilon/2}{z+1}$$

Input: a sequence $x_1, x_2, \ldots, x_n \in U = [N]$

Output: an estimation of $z = \left|\{x_1, x_2, \ldots, x_n\}\right|$

- (*idealized*) uniform hash function $h : U \to [0,1]$

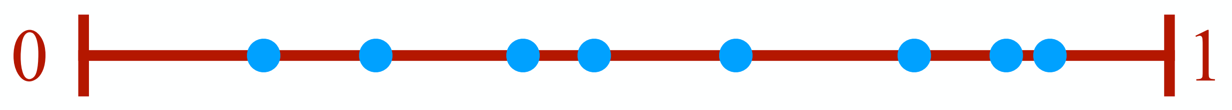
Min Sketch:

let $Y = \min_{1 \le i \le n} h(x_i)$;

return $\widehat{Z} = \frac{1}{Y} - 1$;

- Uniform independent hash values:

$$H_1, \ldots, H_z \in [0,1]$$



- $Y = \min_{1 \le i \le z} H_i$

geometric probability: $\Pr[Y > y] = (1 - y)^z$ ➡ pdf: $p(y) = z(1-y)^{z-1}$

$$\mathbb{E}[Y^2] = \int_0^1 y^2 p(y)\,\mathrm{d}y = \int_0^1 y^2 z(1-y)^{z-1}\,\mathrm{d}y = \frac{2}{(z+1)(z+2)}$$

$$\mathbf{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{z}{(z+1)^2(z+2)} \le \frac{1}{(z+1)^2}$$

Input: a sequence $x_1, x_2, \ldots, x_n \in U = [N]$

Output: an estimation of $z = \left|\{x_1, x_2, \ldots, x_n\}\right|$

- (*idealized*) uniform hash function $h : U \to [0,1]$

Min Sketch:

let $Y = \min_{1 \le i \le n} h(x_i)$;

return $\widehat{Z} = \frac{1}{Y} - 1$;

- By symmetry:

$$\mathbb{E}[Y] = \frac{1}{z+1}$$

- Goal:

$$\Pr\left[\widehat{Z} < (1-\epsilon)z \text{ or } \widehat{Z} > (1+\epsilon)z\right] \le \delta$$

assuming $\epsilon \le 1/2$

(Chebyshev)

$$\mathbf{Var}[Y] \le \frac{1}{(z+1)^2} \Longrightarrow \Pr\left[\left|Y - \mathbb{E}[Y]\right| > \frac{\epsilon/2}{z+1}\right] \le \frac{4}{\epsilon^2}$$

The Mean Trick (for Variance Reduction)

- Variance and covariance:

$$\mathbf{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

$$\mathbf{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

- Useful properties:

$$\mathbf{Var}[X + a] = \mathbf{Var}[X]$$

$$\mathbf{Var}[aX] = a^2\mathbf{Var}[X]$$

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} X_i\right] = \mathbb{E}[X_1]$$

$$\mathbf{Var}\left[\sum_i X_i\right] = \sum_i \mathbf{Var}[X_i] + \sum_{i \neq j} \mathbf{Cov}(X_i, X_j)$$

- For pairwise independent identically distributed X_i's:

$$\mathbf{Var}\left[\frac{1}{k}\sum_{i=1}^{k} X_i\right] = \frac{1}{k^2}\sum_{i=1}^{k} \mathbf{Var}[X_i] = \frac{1}{k}\mathbf{Var}[X_1]$$

Input: a sequence $x_1, x_2, \ldots, x_n \in U = [N]$

Output: an estimation of $z = \left|\{x_1, x_2, \ldots, x_n\}\right|$

- uniform & independent hash functions $h_1, \ldots, h_k : U \to [0,1]$

Min Sketch:

for each $1 \le j \le k$, let $Y_j = \min_{1 \le i \le n} h_j(x_i)$;

return $\widehat{Z} = \frac{1}{\overline{Y}} - 1$ where $\overline{Y} = \frac{1}{k}\sum_{j=1}^{k} Y_j$;

- For every $1 \le j \le k$:

$$\mathbb{E}\left[Y_j\right] = \frac{1}{z+1} \quad \xrightarrow{\text{linearity of expectation}} \quad \mathbb{E}\left[\overline{Y}\right] = \frac{1}{z+1}$$

$$\mathbf{Var}[Y_j] \le \frac{1}{(z+1)^2} \quad \xrightarrow{\text{independence}} \quad \mathbf{Var}\left[\overline{Y}\right] \le \frac{1}{k(z+1)^2}$$

Input: a sequence $x_1, x_2, \ldots, x_n \in U = [N]$

Output: an estimation of $z = \left|\{x_1, x_2, \ldots, x_n\}\right|$

- uniform & independent hash functions $h_1, \ldots, h_k : U \to [0,1]$

Min Sketch:

for each $1 \le j \le k$, let $Y_j = \min_{1 \le i \le n} h_j(x_i)$;

return $\widehat{Z} = \frac{1}{\overline{Y}} - 1$ where $\overline{Y} = \frac{1}{k}\sum_{j=1}^{k} Y_j$;

$$\mathbb{E}\left[\overline{Y}\right] = \frac{1}{z+1}$$

$$\mathbf{Var}\left[\overline{Y}\right] \le \frac{1}{k(z+1)^2}$$

- Goal: $\Pr\left[\widehat{Z} < (1-\epsilon)z \text{ or } \widehat{Z} > (1+\epsilon)z\right] \le \delta$

assuming $\epsilon \le 1/2$

$$\Pr\left[\left|\overline{Y} - \mathbb{E}\left[\overline{Y}\right]\right| > \frac{\epsilon/2}{z+1}\right] \le \frac{4}{k\epsilon^2} \le \delta$$

(Chebyshev)

Set $k = \left\lceil \frac{4}{\epsilon^2 \delta} \right\rceil$

Input: a sequence $x_1, x_2, \ldots, x_n \in U = [N]$

Output: an estimation of $z = \left|\{x_1, x_2, \ldots, x_n\}\right|$

- uniform & independent hash functions $h_1, \ldots, h_k : U \to [0,1]$

Min Sketch: set $k = \lceil 4/(\epsilon^2\delta) \rceil$

for each $1 \le j \le k$, let $Y_j = \min_{1 \le i \le n} h_j(x_i)$;

return $\widehat{Z} = \frac{1}{\overline{Y}} - 1$ where $\overline{Y} = \frac{1}{k}\sum_{j=1}^{k} Y_j$;

$$\Pr\left[(1-\epsilon)z \le \widehat{Z} \le (1+\epsilon)z\right] \ge 1 - \delta$$

- **Space cost**: $k = O\left(\frac{1}{\epsilon^2\delta}\right)$ ***real numbers*** in $[0,1]$
- Storing k ***idealized*** hash functions.

Two-Point Sampling (2-Universal Hashing)

- Let $p > 1$ be a prime number and $[p] = \{0,1,\ldots,p-1\} = \mathbb{Z}_p$.
- Pick $\boldsymbol{a}, \boldsymbol{b} \in [p]$ *u.a.r.* and let $r_i = (\boldsymbol{a} \cdot i + \boldsymbol{b}) \bmod p$ for $i = 1,2,\ldots,p$
  - $r_1, \ldots, r_p \in [p]$ are pairwise independent
  - each r_i is uniformly distributed over $[p]$
- Linear congruential hashing $f : \mathrm{GF}(q) \to \mathrm{GF}(q)$ over finite field $\mathrm{GF}(q)$:
  - Pick $\boldsymbol{a}, \boldsymbol{b} \in \mathrm{GF}(q)$ u.a.r and let $f(x) = \boldsymbol{a} \cdot x + \boldsymbol{b}$ for $x \in \mathrm{GF}(q)$
    - $\{x \in \mathrm{GF}(q)\}$ are pairwise independent
    - each $f(x)$ is uniformly distributed over $\mathrm{GF}(q)$
    - $\mathrm{GF}(2^w)$ exists for any positive integer $w \in \mathbb{Z}^+$

Flajolet-Martin Algorithm

Input: a sequence $x_1, x_2, \ldots, x_n \in [N] \subseteq [2^w]$

Output: an estimation of $z = \left| \{x_1, x_2, \ldots, x_n\} \right|$

- 2-wise independent hash function $h : [2^w] \to [2^w]$
- For $y \in [2^w]$, let $\mathrm{zeros}(y) = \max\{i : 2^i \mid y\}$ denote # of trailing 0's

Flajolet-Martin Algorithm:

let $R = \max_{1 \le i \le n} \mathrm{zeros}(h(x_i))$;

return $\widehat{Z} = 2^R$;

$$\Pr\left[\widehat{Z} < \frac{z}{C} \text{ or } \widehat{Z} > C \cdot z\right] \le \frac{3}{C}$$

Input: a sequence $x_1, x_2, \ldots, x_n \in [N] \subseteq [2^w]$

Output: an estimation of $z = \left| \{x_1, x_2, \ldots, x_n\} \right|$

- 2-wise independent hash function $h : [2^w] \to [2^w]$
- For $y \in [2^w]$, let $\mathrm{zeros}(y) = \max\{i : 2^i \mid y\}$ denote # of trailing 0's

Flajolet-Martin Algorithm:

let $R = \max_{1 \le i \le n} \mathrm{zeros}(h(x_i))$;

return $\widehat{Z} = 2^R$;

Let

$$Y_r = \sum_{\text{distinct } x \in \{x_1, \ldots, x_n\}} I\left[\mathrm{zeros}\left(h(x)\right) \ge r\right]$$

(linearity of expectation)

$$\mathbb{E}[Y_r] = \sum_{\text{distinct } x \in \{x_1, \ldots, x_n\}} \Pr\left[\mathrm{zeros}\left(h(x)\right) \ge r\right] = z2^{-r}$$

(pairwise independence)

$$\mathbf{Var}[Y_r] = \sum_{\text{distinct } x \in \{x_1, \ldots, x_n\}} \mathbf{Var}\left[I[\mathrm{zeros}\left(h(x)\right) \ge r]\right] = z2^{-r}(1 - 2^{-r}) \le z2^{-r}$$

Input: a sequence $x_1, x_2, \ldots, x_n \in [N] \subseteq [2^w]$

Output: an estimation of $z = \left| \{x_1, x_2, \ldots, x_n\} \right|$

- 2-wise independent hash function $h : [2^w] \to [2^w]$
- For $y \in [2^w]$, let $\mathrm{zeros}(y) = \max\{i : 2^i \mid y\}$ denote # of trailing 0's

Flajolet-Martin Algorithm:

let $R = \max_{1 \le i \le n} \mathrm{zeros}(h(x_i))$;

return $\widehat{Z} = 2^R$;

Let

$$Y_r = \sum_{\text{distinct } x \in \{x_1, \ldots, x_n\}} I\left[\mathrm{zeros}\left(h(x)\right) \ge r\right]$$

$$\mathbb{E}[Y_r] = z2^{-r} \qquad \mathbf{Var}[Y_r] \le z2^{-r}$$

(denote $r^* = \lceil \log_2 Cz \rceil$) $\quad \Pr\left[\widehat{Z} > Cz\right] \le \Pr[R \ge r^*]$

(observe $R = \max\{r : Y_r > 0\}$) $\quad \le \Pr[Y_{r^*} > 0] = \Pr[Y_{r^*} \ge 1]$

(Markov's inequality) $\quad \le \mathbb{E}[Y_{r^*}] = z/2^{r^*} \le 1/C$

Input: a sequence $x_1, x_2, \ldots, x_n \in [N] \subseteq [2^w]$

Output: an estimation of $z = \left|\{x_1, x_2, \ldots, x_n\}\right|$

- 2-wise independent hash function $h : [2^w] \to [2^w]$
- For $y \in [2^w]$, let $\text{zeros}(y) = \max\{i : 2^i | y\}$ denote # of trailing 0's

Flajolet-Martin Algorithm:

let $R = \max_{1 \le i \le n} \text{zeros}(h(x_i))$;

return $\widehat{Z} = 2^R$;

Let

$$Y_r = \sum_{\text{distinct } x \in \{x_1, \ldots, x_n\}} I\left[\text{zeros}\left(h(x)\right) \ge r\right]$$

$\mathbb{E}[Y_r] = z2^{-r}$ $\quad$ $\mathbf{Var}[Y_r] \le z2^{-r}$

(denote $r^{**} = \lceil \log_2(z/C) \rceil$) $\quad \Pr\left[\widehat{Z} < z/C\right] \le \Pr[R < r^{**}]$

(observe $R = \max\{r : Y_r > 0\}$) $\quad \le \Pr[Y_{r^{**}} = 0]$

(Chebyshev's inequality) $\quad \le \mathbf{Var}[Y_{r^{**}}]/\mathbb{E}[Y_{r^{**}}]^2 \le 2^{r^{**}}/z$

$\le 2/C$

Input: a sequence $x_1, x_2, \ldots, x_n \in [N] \subseteq [2^w]$

Output: an estimation of $z = \left|\{x_1, x_2, \ldots, x_n\}\right|$

- 2-wise independent hash function $h : [2^w] \to [2^w]$
- For $y \in [2^w]$, let $\text{zeros}(y) = \max\{i : 2^i \mid y\}$ denote # of trailing 0's

Flajolet-Martin Algorithm:

let $R = \max_{1 \le i \le n} \text{zeros}(h(x_i))$;

return $\widehat{Z} = 2^R$;

$$\Pr\left[\widehat{Z} < \frac{z}{C} \text{ or } \widehat{Z} > C \cdot z\right] \le \frac{3}{C}$$

- **Space cost:** $O(\log \log N)$ bits for maintaining R
- $O(\log N)$ bits for storing 2-wise independent hash function

Weierstrass Approximation Theorem
(魏尔施特拉斯逼近定理)

- <u>Weierstrass Approximation Theorem</u>: Let $f : [0,1] \to [0,1]$ be a continuous function. For any $\epsilon > 0$, there exists a polynomial p such that

$$\sup_{x \in [0,1]} |p(x) - f(x)| \le \epsilon$$

- **Proof**: Let integer n be sufficiently large (to be fixed later).

  For $x \in [0,1]$, let $X \sim \frac{1}{n}\text{Bin}(n, x)$. Define polynomial p on $x \in [0,1]$ to be:

$$p(x) = \mathbb{E}\left[f(X)\right] = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) p_X(k) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k}$$

Let $f : [0,1] \to [0,1]$ be continuous. For $x \in [0,1]$, let $X \sim \frac{1}{n}\text{Bin}(n,x)$, and:

$$p(x) = \mathbb{E}\left[f(X)\right] = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} x^k (1-x)^{n-k}$$

$$|p(x) - f(x)| = \left|\mathbb{E}\left[f(X) - f(x)\right]\right| \leq \mathbb{E}\left[\left|f(X) - f(x)\right|\right]$$

(f is continuous on $[0,1] \implies \exists \delta > 0$ s.t. $|f(x) - f(y)| \leq \epsilon/2$ for all $|x - y| \leq \delta$)

$$= \mathbb{E}\left[\left|f(X) - f(x)\right| \mid \left|X - x\right| \leq \delta\right] \cdot \Pr\left(\left|X - x\right| \leq \delta\right)$$

$$+\mathbb{E}\left[\left|f(X) - f(x)\right| \mid \left|X - x\right| > \delta\right] \cdot \Pr\left(\left|X - x\right| > \delta\right)$$

$$\leq \mathbb{E}\left[\epsilon/2\right] + \left|1 - 0\right| \cdot \Pr\left(\left|X - x\right| > \delta\right) \quad \leq \frac{\epsilon}{2} + \frac{x(1-x)}{n\delta^2}$$

(Chebyshev)

$$\leq \frac{\epsilon}{2} + \frac{1}{4n\delta^2} \quad \leq \epsilon$$ if we choose $n \geq \frac{1}{2\epsilon\delta^2}$

Weierstrass Approximation Theorem
(魏尔施特拉斯逼近定理)

- Weierstrass Approximation Theorem: Let $f : [0,1] \to [0,1]$ be a continuous function. For any $\epsilon > 0$, there exists a polynomial p such that

$$\sup_{x \in [0,1]} |p(x) - f(x)| \le \epsilon$$

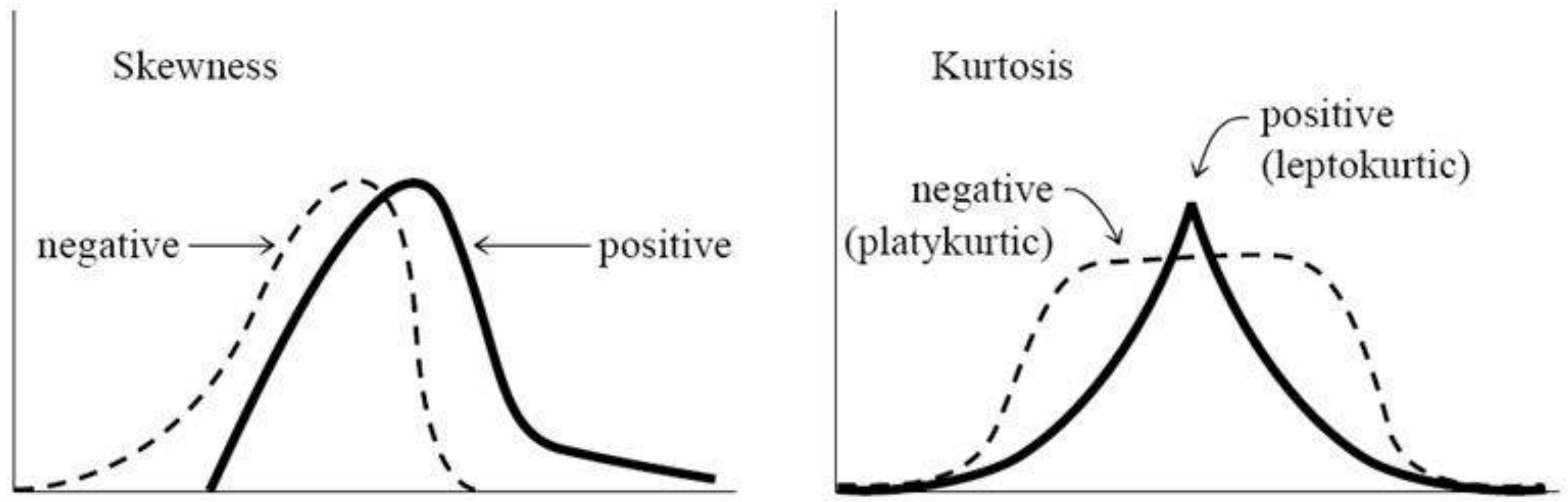
- **Proof**: By continuity, $\exists \delta > 0$ s.t. $|f(x) - f(y)| \le \epsilon/2$ if $|x - y| \le \delta$.

  Let $n \ge 1/(2\epsilon\delta^2)$ be any integer. For $x \in [0,1]$, let $X \sim \frac{1}{n}\text{Bin}(n, x)$, and:

$$p(x) = \mathbb{E}\left[f(X)\right] = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} x^k (1-x)^{n-k}$$

  For any $x \in [0,1]$, it holds that $|p(x) - f(x)| \le \epsilon$.

Higher Moments

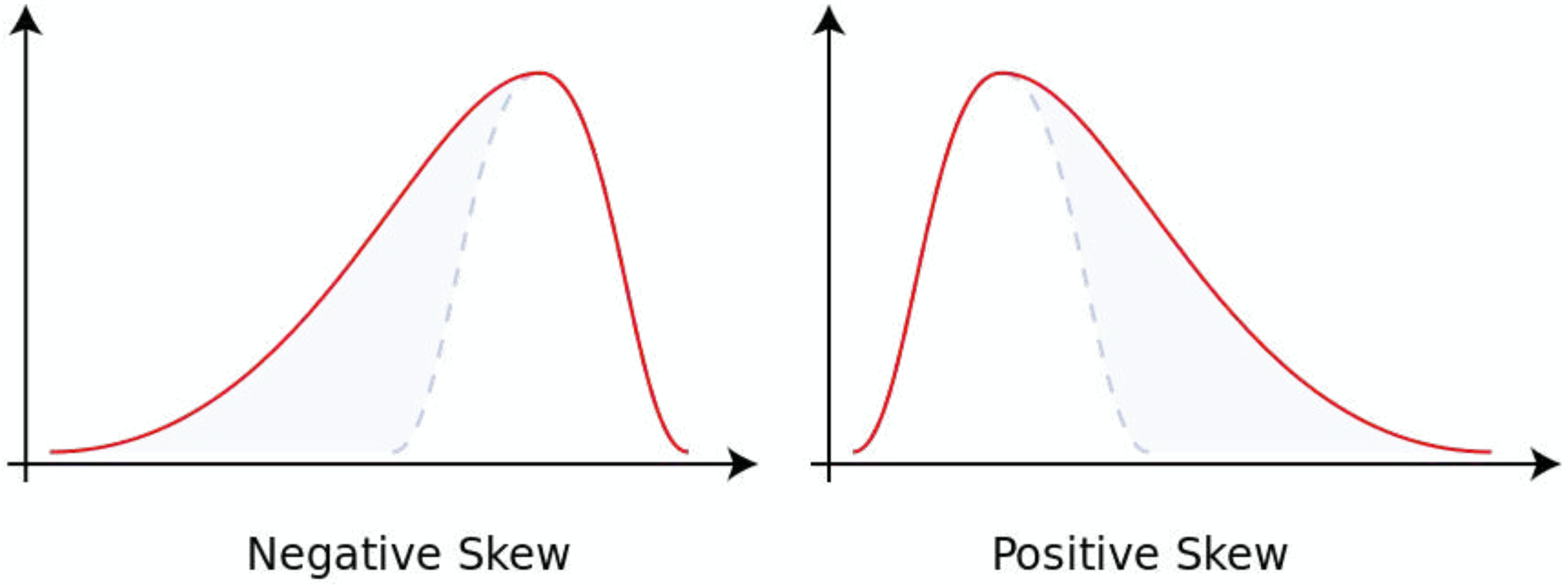


Skewness (偏度)

- The **skewness** (偏度) of a random variable X with expectation $\mu = \mathbb{E}[X]$ and standard deviation $\sigma = \sqrt{\mathbf{Var}[X]}$ is defined by

$$\mathrm{Skew}[X] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3}$$

standardized moment (of degree 3)

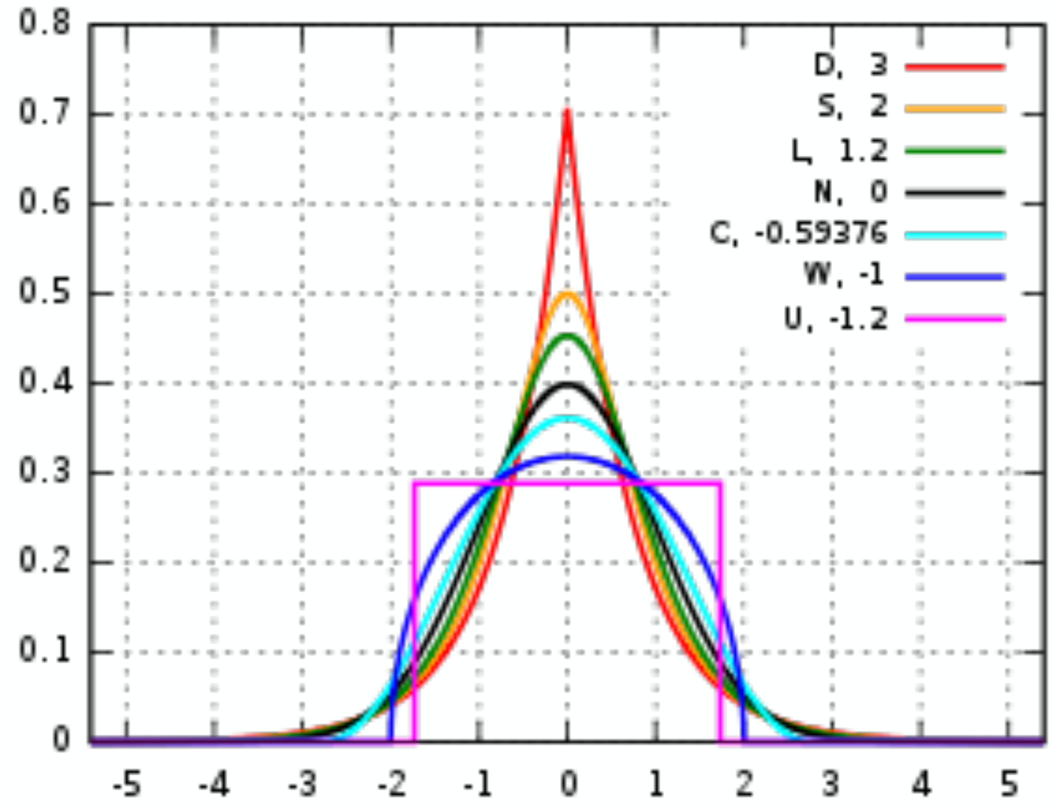


Kurtosis (峰度)

- The **kurtosis** (峰度) of a random variable X with expectation $\mu = \mathbb{E}[X]$ and standard deviation $\sigma = \sqrt{\mathbf{Var}[X]}$ is defined by

$$\mathrm{Kurt}[X] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4}$$

standardized moment (of degree 4)



The kth Moment Method

- Let X be a random variable with $\mathbb{E}[X] = \mu$. For any $C > 1$ and integer $k \geq 1$

$$\Pr\left(|X - \mu| \geq C \cdot \mathbb{E}\left[|X - \mu|^k \right]^{\frac{1}{k}} \right) \leq \frac{1}{C^k}$$

- **Proof**: Apply Markov's inequality to $Z = |X - \mu|^k$.

The Moment Problem

- Do moments $m_k = \mathbb{E}[X^k]$, $\forall k \geq 1$, uniquely identify the distribution of X?
- If X takes values from a **finite** set $\{x_1, \ldots, x_n\}$ with $p_X(x_i) = p_i$ & moments $\{m_i\}$ then solving the Vandermonde system:

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^n & x_2^n & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix}$$

can recover the *pmf* $p_i = p_X(x_i)$

The Moment Problem

- Do moments $m_k = \mathbb{E}[X^k]$, $\forall k \geq 1$, uniquely identify the distribution of X?
  - If $\mathbb{E}[X^k] = \mathbb{E}[Y^k]$ for all $k \geq 1$, are X and Y always identically distributed?
- If X and Y have the same **moment generating function (MGF)**

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{k \geq 0} \frac{t^k \mathbb{E}[X^k]}{k!}$$

  then X and Y are identically distributed.

- The MGF $M_X(t)$ is convergent if the sequence $\mathbb{E}[X^k]$ does not grow too fast.