Generative Models: Fundamentals and Applications

Lecture 1: Basics of Probability and Statistics

Shuigeng Zhou, Yuxi Mi
College of CSAI

September 15, 2025

Outline

- Some basic concepts of probability theory
- Some common distributions
- Transformations of random variables
- Monte Carlo approximation
- Information theory

What is probability?

- Frequentist interpretation
  - Probability represents long run frequencies of events
  - It interprets *probability as the frequency of occurrence of an outcome*
- Bayesian interpretation
  - Probability is used to quantify the uncertainty about something
  - It interprets probability as our believe of the likelihood of a certain outcome, i.e, *probability measures a degree of belief*

Probability of an event

- Let A be an event
  - $P(\mathrm{A})$: the probability that the event A is true
    - $0 \leq P(\mathrm{A}) \leq 1$
    - $P(\mathrm{A}) = 0$ means the event definitely will not happen
    - $P(\mathrm{A}) = 1$ means the event definitely will happen
  - $P(\bar{A})$: the probability that the event not A
    - $P(\bar{A}) = 1 - P(\mathrm{A})$

Probability of a random variable

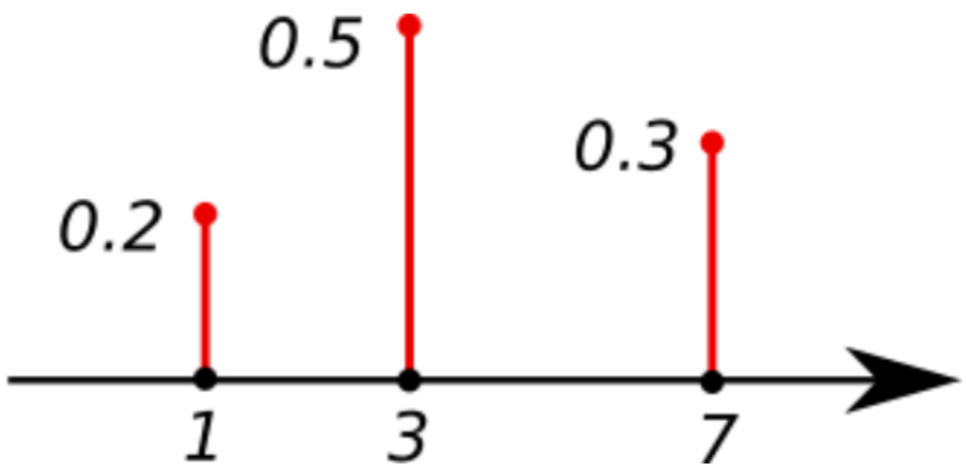
- Discrete random variable X
  - take on any value from a finite or countably infinite set $\mathcal{X}$
- Example: event vs. random variable
  - Event A={明天会下雨}
  - Random variable X={0,1}
    - 1 表示明天会下雨
    - 0 表示明天不会下雨
  - P(X=1) = P(A)

Probability of a random variable

- Example: event vs. random variable
  - Event A={明天会下雨}
  - Random variable X={1,2,3,4}
    - 1 表示明天会下雨
    - 2 表示明天下雪
    - 3 表示明天下冰雹
    - 4 表示明天是晴天
  - P(X=1) = P(A)

Probability of a random variable

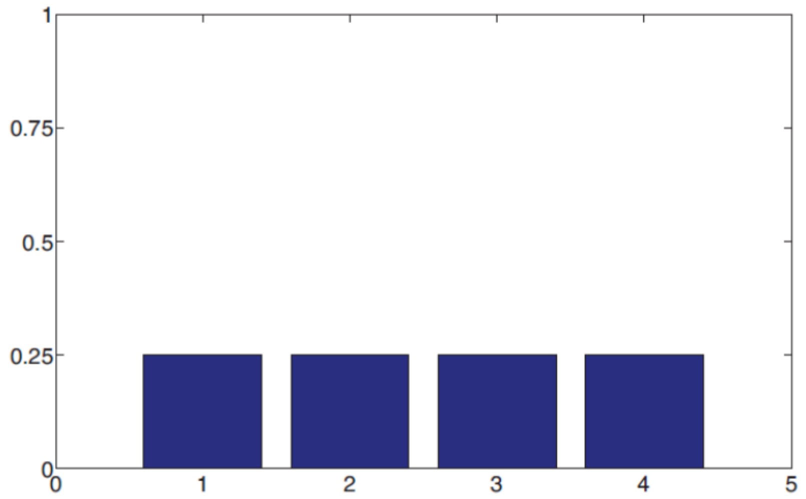
- Probability mass function $P(X{=}x)$ (or $P(x)$)
  - the probability of the event that X = x
  - $0 \le P(x) \le 1$
  - $\sum_{x \in \mathcal{X}} P(x) = 1$



$\boxed{\mathcal{X} = \{1,3,7\}}$

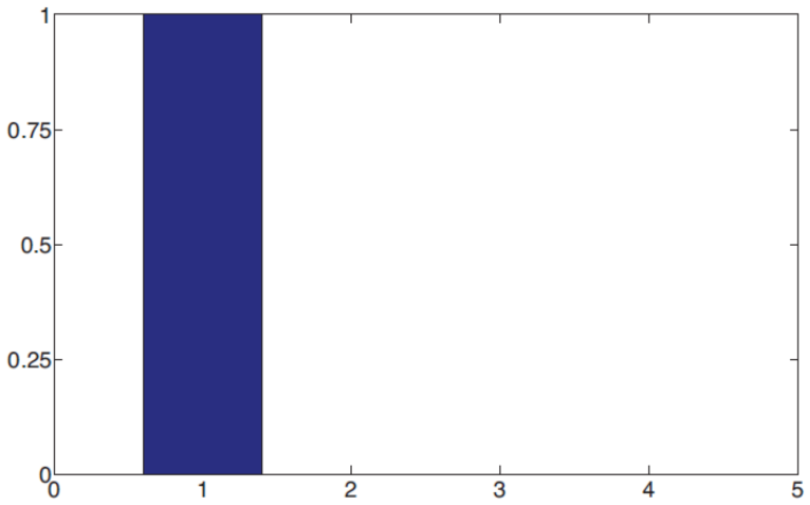
Probability of a random variable

- Special case



(a) 均匀分布

$\boxed{\mathcal{X} = \{1,2,3,4\}}$



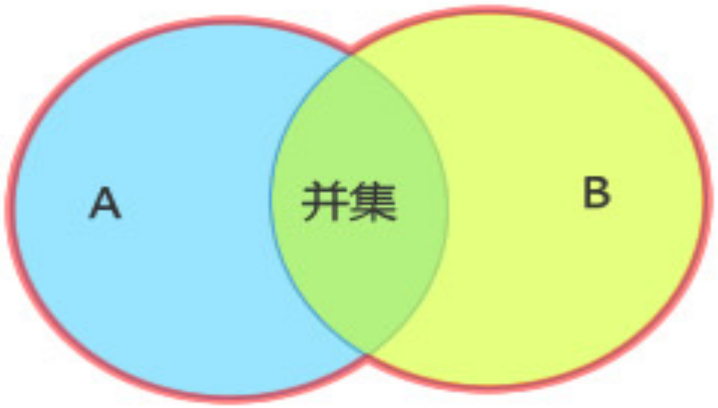
(b) 退化分布

$\boxed{\mathcal{X} = \{1\}}$

Basic rules of probability

- Probability of a union of two events
  - Given two events A and B

$$
\begin{aligned}
P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\
&= P(A) + P(B) \text{ if A and B are mutually exclusive}
\end{aligned}
$$



Basic rules of probability

- Joint probability
  - Given two events A and B

$$\boxed{\begin{aligned} P(A,B) = P(A \cap B) &= P(A)P(B|A) \\ &= P(B)P(A|B) \end{aligned}}$$

  - Marginal distribution (rule of total probability)

$$\boxed{P(A) = \sum_b P(A,B) = \sum_b P(A|B=b)P(B=b)}$$

Basic rules of probability

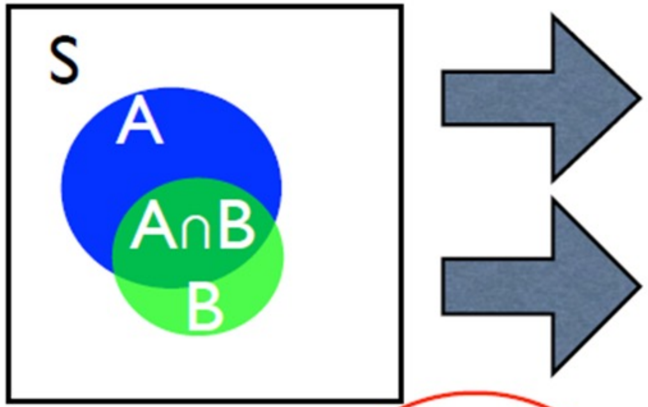
- Joint probability
  - chain rule

$$P(X_1, X_2, \dots X_D) \\ = P(X_1)P(X_2|X_1)P(X_3|X_1,X_2)\dots P(X_D|X_1,X_2,\dots,X_{D-1})$$

- Conditional probability

$$P(A|B) = \frac{P(A,B)}{P(B)}, \qquad \text{if } P(B) > 0$$

Bayes rule/theorem



$$P(A\cap B) = P(B|A)P(A)$$

$$P(A\cap B) = P(A|B)P(B)$$

Posterior, Likelihood, Prior

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

Example: medical diagonsis

- Question： If the nucleic acid test (NAT) is positive, what is the probability you have COVID-19?
  - X = 1 is the event NAT is positive
  - Y = 1 is the event you have COVID-19
  - Sensitivity score: P(X = 1 | Y = 1) = 0.8
  - P(Y = 1) = 0.004
  - P(X = 1 | Y = 0) = 0.1

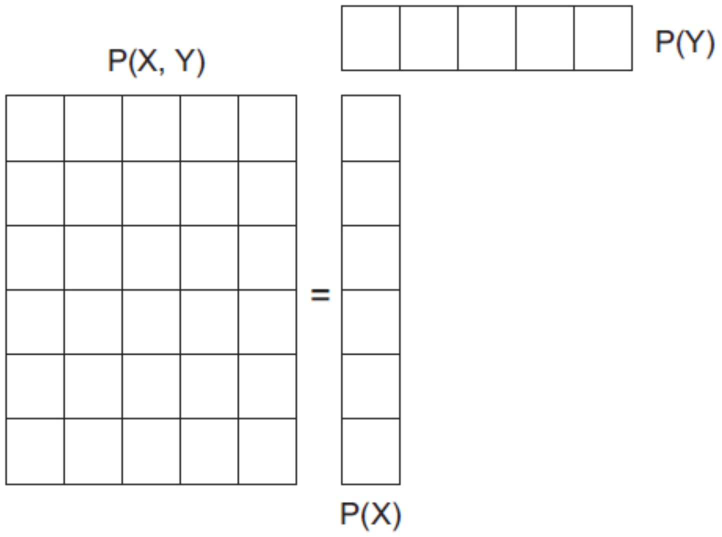
$$\begin{aligned} P(y=1|x=1) &= \frac{P(y=1)P(x=1|y=1)}{P(y=1)P(x=1|y=1)+P(y=0)P(x=1|y=0)} \\ &= \frac{0.004 \times 0.8}{0.004 \times 0.8 + (1-0.004) \times 0.1} = 0.031 \end{aligned}$$

Independence and conditional independence

- Unconditionally independent (marginally independent)

We say X and Y are **unconditionally independent** or **marginally independent**, if we can represent the joint as the product of the two marginals

$$\boxed{X \perp Y \iff P(X,Y) = P(X)P(Y)}$$



Independence and conditional independence

- Conditionally independent (CI)

We say X and Y are conditionally independent (CI) given Z iff the conditional joint can be written as a product of conditional marginals:

$$\boxed{P(X \perp Y|Z) \Leftrightarrow P(X,Y|Z) = P(X|Z)P(Y|Z)}$$

Probability in continuous case

- Continuous random variable X
  - Cumulative distribution function (cdf)

$$\boxed{F(x) = P(X \leq x)}$$

  - Probability density function (pdf)

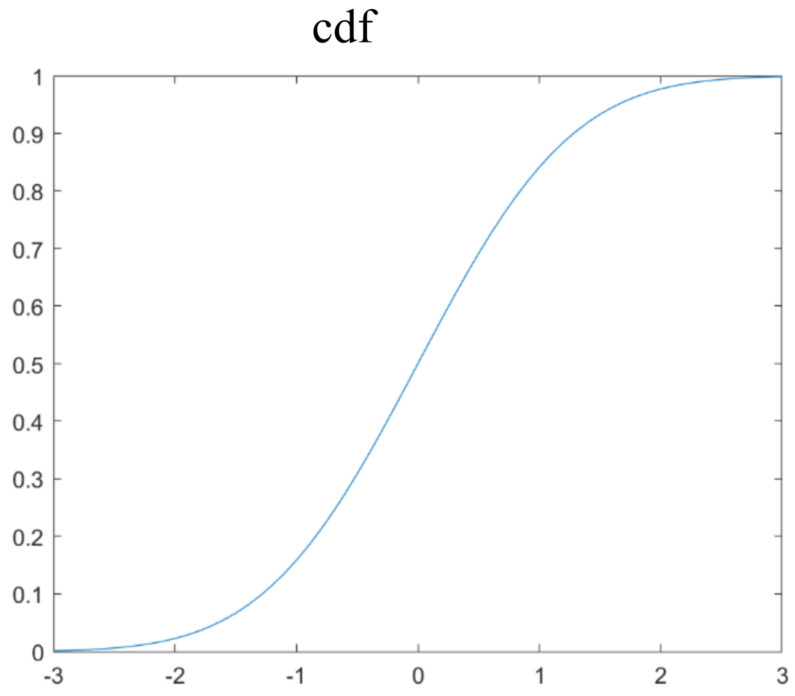
$$\boxed{p(x) = \frac{d}{dx} F(x)}$$

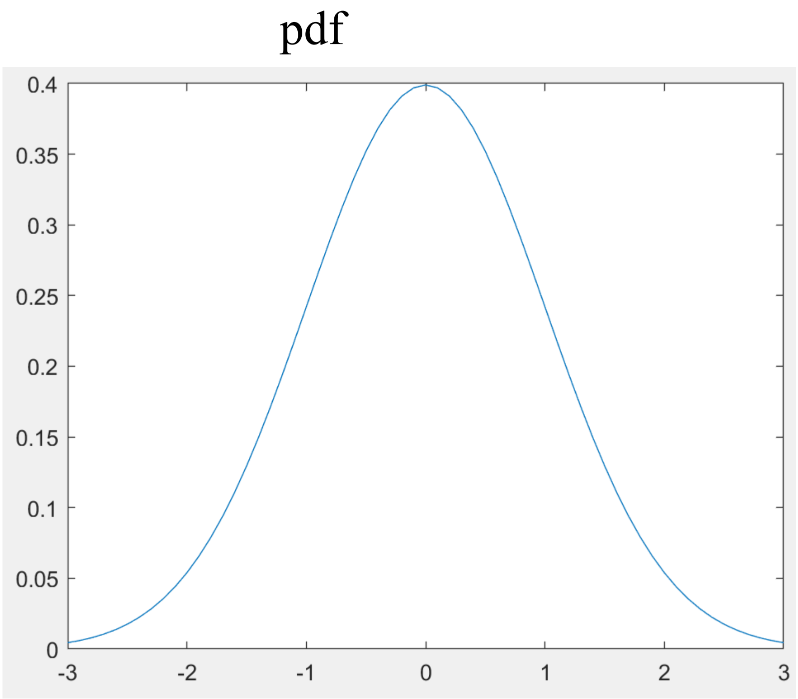
  - Probability in some interval [a, b]

$$\boxed{P(a < X \leq b) = F(b) - F(a) = \int_a^b p(x)\, dx}$$

Probability in continuous case

- Example

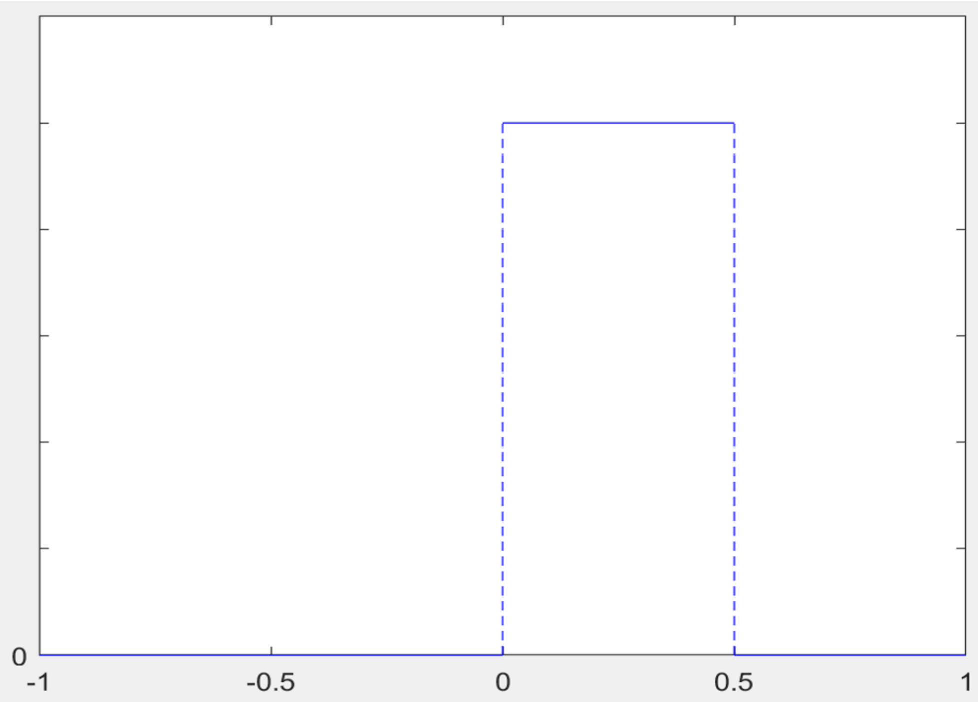




Probability in continuous case

- Example

The pdf p(x)



$$p(x) = \begin{cases} 2, & 0 \le x \le 0.5 \\ 0, & \text{otherwise} \end{cases}$$

Mean and Variance

- Mean (expected value) μ
  - Discrete case

$$\boxed{E[X] \triangleq \sum_{x \in \mathcal{X}} x P(x)}$$

  - Continuous case

$$\boxed{E[X] \triangleq \int_{x \in \mathcal{X}} x p(x)\, dx}$$

Mean and Variance

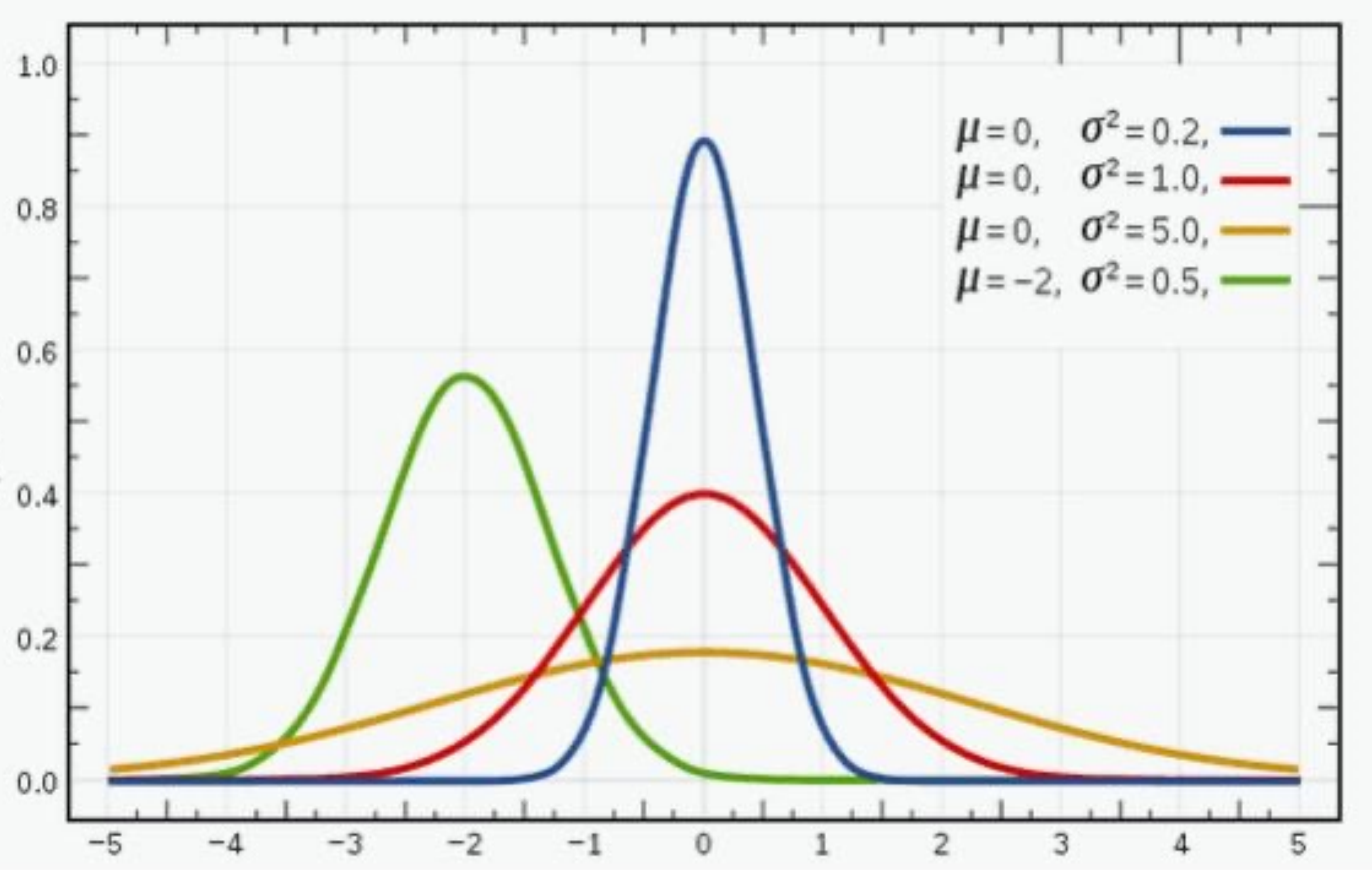
- Variance σ^2

$$\boxed{\operatorname{var}[X] \triangleq E[(X-\mu)^2] = E[X^2] - \mu^2}$$

- Standard deviation

$$\boxed{\operatorname{std}[X] \triangleq \sqrt{\operatorname{var}[X]}}$$

Mean and Variance

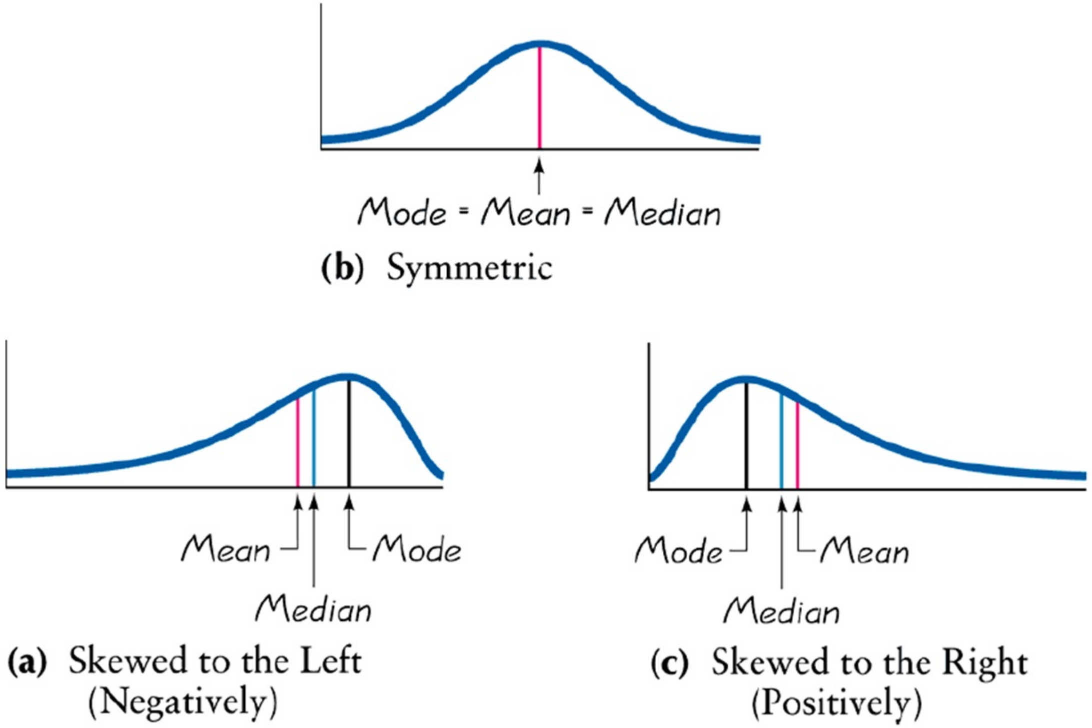


The variance estimation

- Population variance
  - $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 - \mu^2$
  - $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Sample variance
  - Taking *n* samples from the population, estimate the variance
    - ${\sigma_y}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_y)^2, \mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i$
  - Sampling multiple times, computing the expected valued of ${\sigma_y}^2$
    - $E\left({\sigma_y}^2\right) = \frac{n-1}{n}\sigma^2$, so $\sigma^2 = \frac{n}{n-1}E\left({\sigma_y}^2\right)$
  - We take the variance of one time sampling as $E\left({\sigma_y}^2\right)$, the sample variance s^2 is
    - $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mu_y)^2$

Mean, median and mode: measure of central tendency

- Mode: the most frequent number occurring in the data set



Covariance and correlation

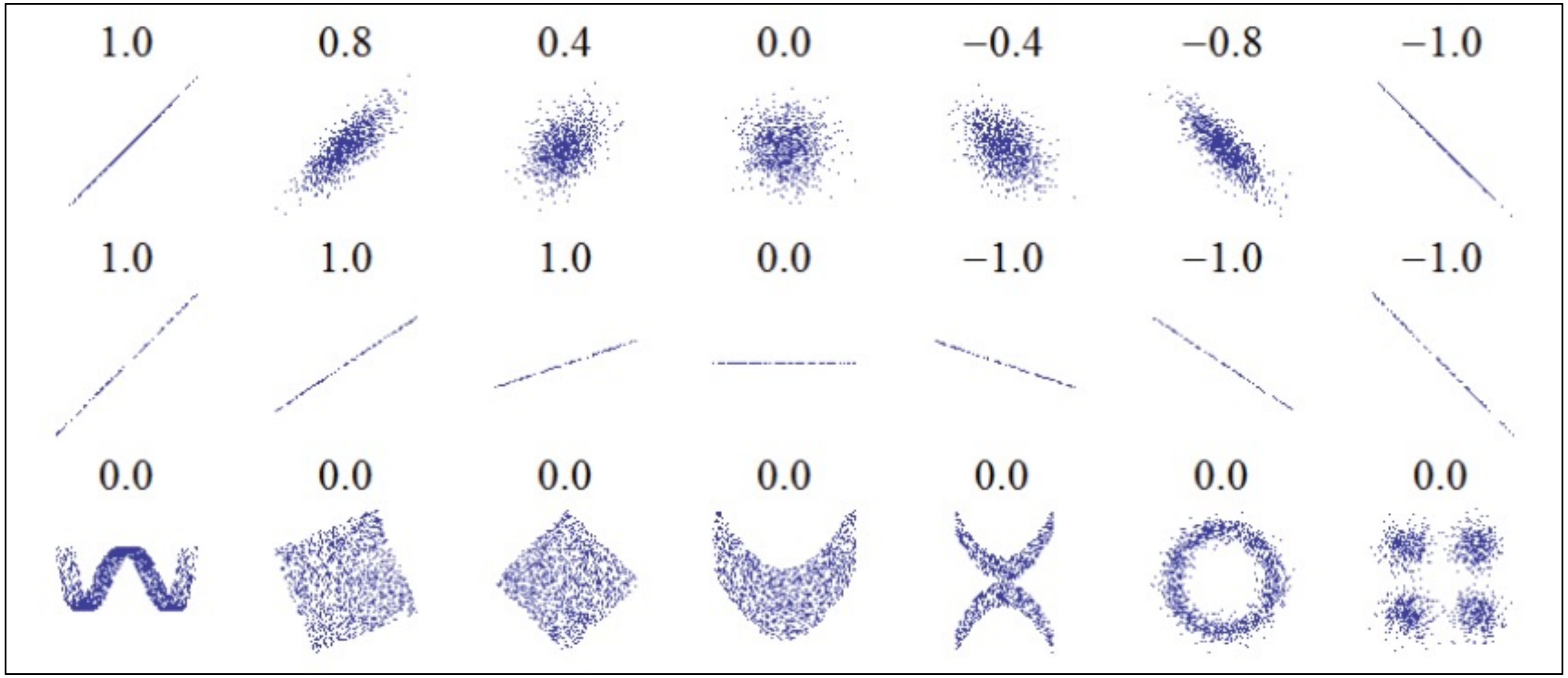
- Covariance
  - Given two random variables X and Y
  - Measure the degree to which X and Y are (linearly) related

$$\boxed{\operatorname{cov}[X,Y] \triangleq \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]}$$

- Correlation coefficient

$$\boxed{\operatorname{corr}[X,Y] \triangleq \frac{\operatorname{cov}[X,Y]}{\sqrt{\operatorname{var}[X]\operatorname{var}[Y]}}}$$

Covariance and correlation



The correlation reflects the noisiness and direction of a linear relationship

Independence vs. correlation

- If two random variables are independent, they are uncorrelated
- If two random variables are uncorrelated, they may be dependent
- If two variables of gaussian distribution are uncorrelated, they are independent

Joint distribution probability

- For $X_1, X_2, \ldots, X_d$
  - More than 1 variable
  - Model stochastic relationships between the variables
    - Denote a d-dimensional random vector $\boldsymbol{x} = (X_1, X_2, \ldots, X_d)$
  - Number of parameters $O(K^d)$
    - K: the number of states for each variable
    - d: the number of variables

Covariance and correlation

- Covariance matrix
  - For a *d*-dimensional random vector $x = (X_1, X_2, \ldots, X_d)$

$$
\begin{aligned}
\operatorname{cov}[\mathbf{x}] &\triangleq \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T\right] \\
&= \begin{pmatrix}
\operatorname{var}[X_1] & \operatorname{cov}[X_1, X_2] & \cdots & \operatorname{cov}[X_1, X_d] \\
\operatorname{cov}[X_2, X_1] & \operatorname{var}[X_2] & \cdots & \operatorname{cov}[X_2, X_d] \\
\vdots & \vdots & \ddots & \vdots \\
\operatorname{cov}[X_d, X_1] & \operatorname{cov}[X_d, X_2] & \cdots & \operatorname{var}[X_d]
\end{pmatrix}
\end{aligned}
$$

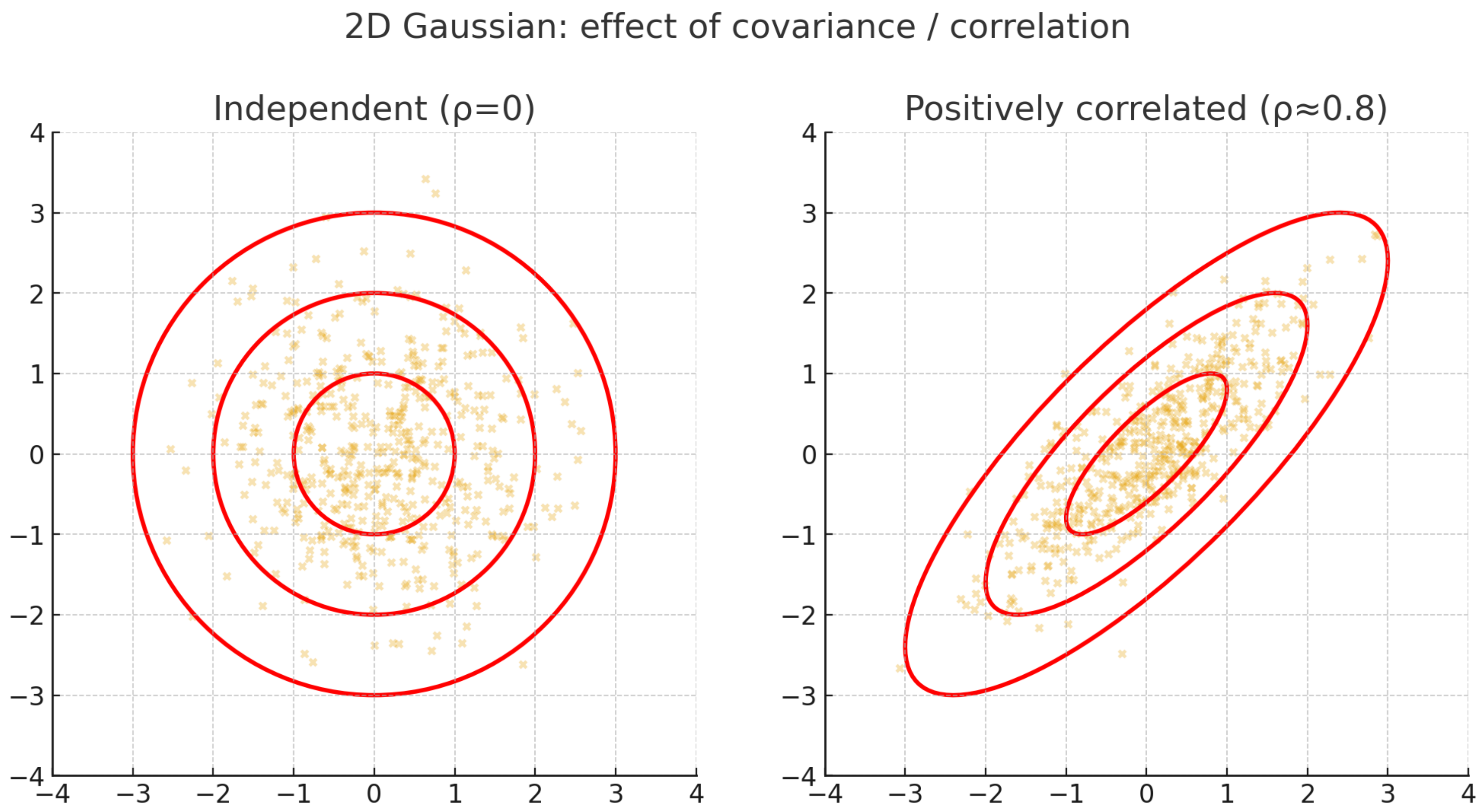
Covariance and correlation

- Correlation matrix
  - For a *d*-dimensional random vector $x = (X_1, X_2, \ldots, X_d)$

$$\mathbf{R} = \begin{pmatrix} \operatorname{corr}[X_1, X_1] & \operatorname{corr}[X_1, X_2] & \cdots & \operatorname{corr}[X_1, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{corr}[X_d, X_1] & \operatorname{corr}[X_d, X_2] & \cdots & \operatorname{corr}[X_d, X_d] \end{pmatrix}$$

  - All diagonal elements are 1, and the others fall in [-1, 1]

Covariance and correlation



Some common distributions

- Empirical distribution
- Binominal/Bernoulli distribution
- Multinominal/Bernoulli distribution
- Uniform distribution
- Gaussian distribution
- The multivariate Gaussian distribution
- Poisson distribution
- Beta distribution
- Dirichlet distribution

Empirical distribution

- Also called empirical measure
  - Given a set of data $\mathcal{D} = \{x_1, x_2, \dots, x_N\}$, the empirical distribution (empirical measure) is defined as

$$\boxed{p_{\text{emp}}(A) \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(A)}$$

where $\delta_x(A)$ is the Dirac measure, defined by

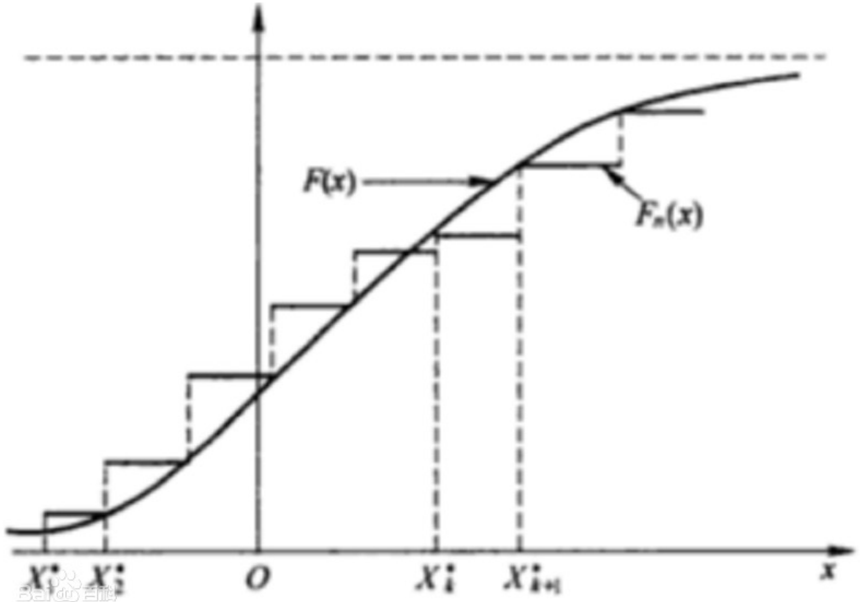
$$\boxed{\delta_x(A) = \begin{cases} 0 & \text{if } x \notin A \\ 1 & \text{if } x \in A \end{cases}}$$

and A is a given value

Empirical distribution

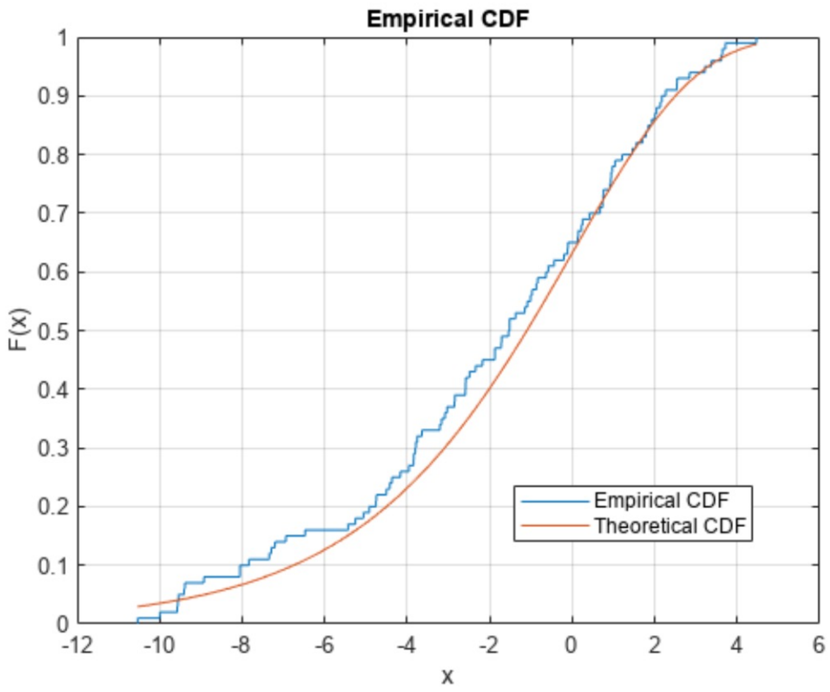
- Empirical CDF (eCDF)
  - Generalized definition: weight

$$F_{\text{emp}}(t) = P_{\text{emp}}((-\infty, t]) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{x_i \le t\}$$



Empirical distribution

- The empirical distribution converges to the true distribution with probability 1



The binomial and Bernoulli distributions

- Binomial distribution: toss a coin n times, the probability of having k heads

$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

mean=$n\theta$, var= $n\theta(1-\theta)$

where

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!}$$

- Bernoulli: a special case of binominal distribution where tossing a coin only once

$$\mathrm{Ber}(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

The multinomial and multinoulli distributions

- Multinomial distribution
  - tossing a die of K-side n times, x=(x1, x2, …, xk) is a vector indicating the appearing time of each side

$$\mathrm{Mu}(\mathbf{x}|n,\boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

where θ_j is the probability that side j shows up, and

$$\binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1! x_2! \cdots x_K!}$$

- Multinoulli: a special case of multinomial distribution with n=1

$$\mathrm{Mu}(\mathbf{x}|1,\boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)}$$

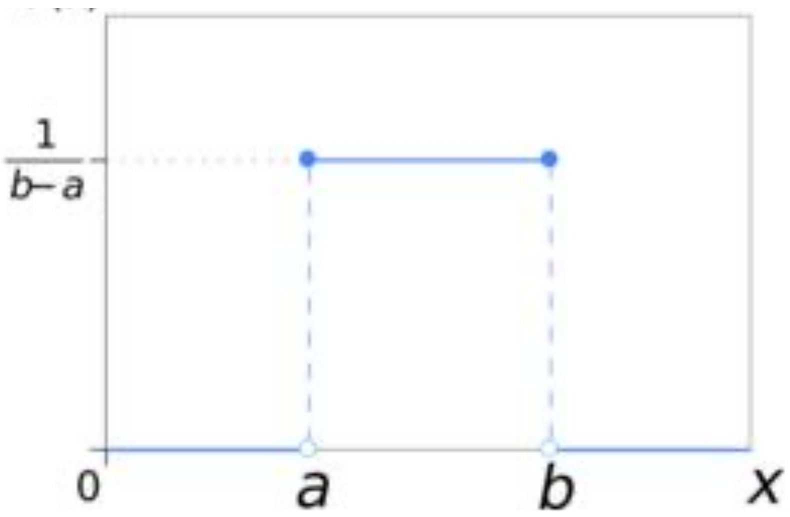
Summary of the multinomial and related distributions

Name	n	K	x
Multinomial	-	-	$\mathbf{x} \in \{0, 1, \ldots, n\}^K, \sum_{k=1}^K x_k = n$
Multinoulli	1	-	$\mathbf{x} \in \{0, 1\}^K, \sum_{k=1}^K x_k = 1$ (1-of-K encoding)
Binomial	-	1	$x \in \{0, 1, \ldots, n\}$
Bernoulli	1	1	$x \in \{0, 1\}$

Uniform distribution

- Uniformly distributed in the interval [a,b]

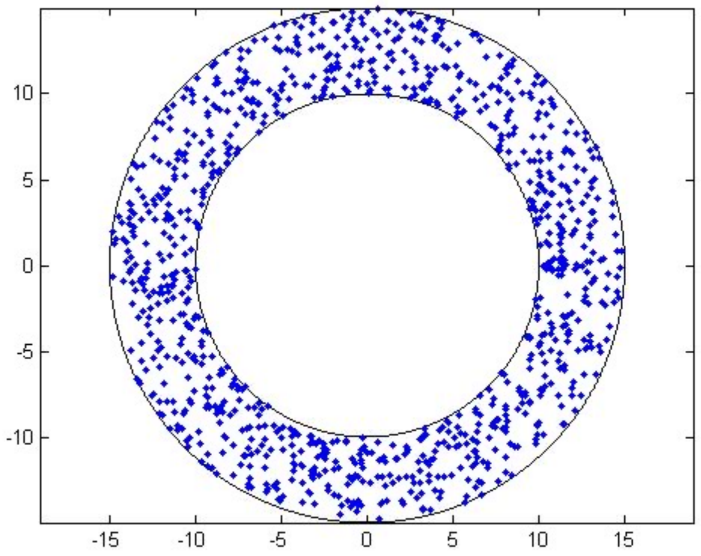
$$\text{Unif}(x|a,b) = \frac{1}{b-a}\mathbb{I}(a \le x \le b)$$



Uniform distribution

- Uniformly distributed in a region R

$$\boxed{\mathrm{Unif}(x|R) = \frac{1}{|R|}\mathbb{I}[x \in R]}$$



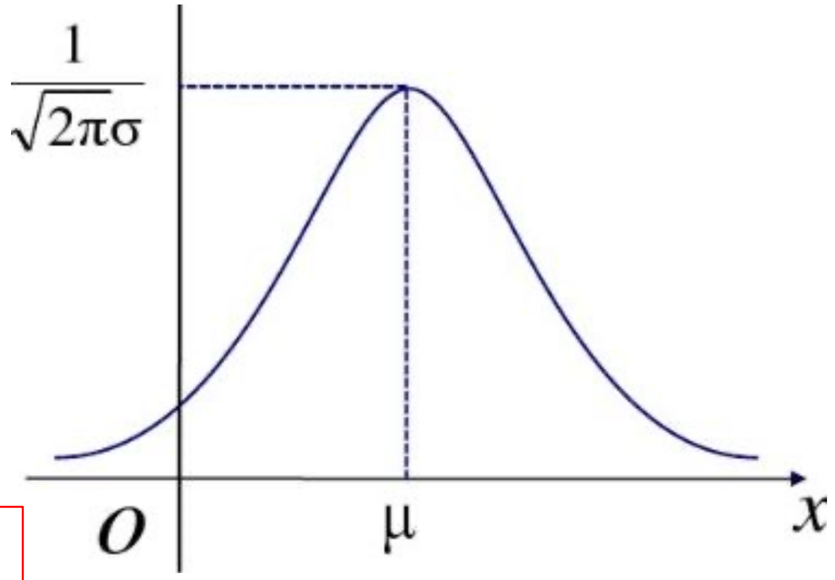
Gaussian distribution

- Also called normal distribution
  - Univariate continuous probability distribution
  - Probability density function

$$\mathcal{N}(x|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

  - Cumulative distribution function

$$\Phi(x;\mu,\sigma^2) \triangleq \int_{-\infty}^{x} \mathcal{N}(z|\mu,\sigma^2)dz$$



Gaussian distribution

- Standard normal distribution
  - Let $y = \frac{x-\mu}{\sigma}$, then $y \sim \mathcal{N}(0,1)$
- Why it is the most widely used distribution?
  - It is simple with only two parameters, and easy to be used
  - Many phenomena in real world have an approximate Gaussian distribution
  - According to the central limit theorem, the sums of independent random variables have an approximate Gaussian distribution

Multivariate Gaussian distribution

- Also called multivariate normal distribution
  - For a D-dimensional random vector $\boldsymbol{x} = (X_1, X_2, \ldots, X_D)$
  - Mean vector $E(\boldsymbol{x}) = \boldsymbol{\mu}$
  - Covariance matrix $\Sigma = \mathrm{cov}(\boldsymbol{x})$
  - Probability density function

$$\boxed{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}$$

The Poisson distribution

We say that $X \in \{0, 1, 2, \ldots\}$ has a **Poisson** distribution with parameter $\lambda > 0$ $X \sim \text{Poi}(\lambda)$, if its pmf is

$$\text{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

cdf $\quad P(x <= k) = e^{-\lambda} \sum_{i=0}^{k} \frac{\lambda^i}{i!}, k = 0, 1, 2..., \lambda > 0$

The Poisson distribution is often used as a **model for counts of rare events** like radioactive decay and traffic accidents

The Poisson distribution

- Considering a binomial distribution

$$P(x=k)=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

$$\lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}=\lim_{n\to\infty}\frac{n!}{k!\,(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}=\frac{\lambda^k}{k!}\lim_{n\to\infty}\frac{n!}{n^k(n-k)!\left(1-\frac{\lambda}{n}\right)^k}\left(1-\frac{\lambda}{n}\right)^n$$

$$=\frac{\lambda^k}{k!}\lim_{n\to\infty}\frac{n!}{(n-k)!\,(n-\lambda)^k}\left(1-\frac{\lambda}{n}\right)^n=\frac{\lambda^k}{k!}\lim_{n\to\infty}\frac{n!}{(n-k)!\,(n-\lambda)^k}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n$$

$$=\frac{\lambda^k}{k!}\lim_{n\to\infty}\frac{n(n-1)(n-2)\ldots(n-k+1)}{(n-\lambda)^k}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n$$

$$\lim_{n\to\infty}\frac{n(n-1)(n-2)\ldots(n-k+1)}{(n-\lambda)^k}=\lim_{n\to\infty}\frac{n^k}{n^k}=1.$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n=e^{-\lambda}.$$

Mean and Variance of Poisson Distribution

- Recall the mean of a binomial distribution $B(n, p) = np$, variance of $B(n, p) = np(1\text{-}p) = \lambda(1\text{-p})$
- Since Poisson distribution is an approximation of binomial distribution **when *n* is approaching infinity, and *p* is extremely** small, then its mean $E(x)=np=\lambda$
- Variance $\lambda(1\text{-}p) \sim \lambda$ when p is very small
- Mean and Variance of Poisson distribution are the same: λ

Student t distribution

- Gaussian distribution is sensitive to outliers.
  - A **more robust** distribution is Student *t* distribution

$$\mathcal{T}(x|\mu,\sigma^2,\nu) \propto \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\left(\frac{\nu+1}{2}\right)}$$

where μ is the mean, $\sigma^2 > 0$ is the scale parameter, and $\nu > 0$

$$\text{mean} = \mu, \text{mode} = \mu, \ \text{var} = \frac{\nu\sigma^2}{(\nu-2)}$$

  - When v=1, it is known as Cauchy or Lorentz distribution, which has a heavy tail
  - When v>>5, it approaches to Gaussian distribution

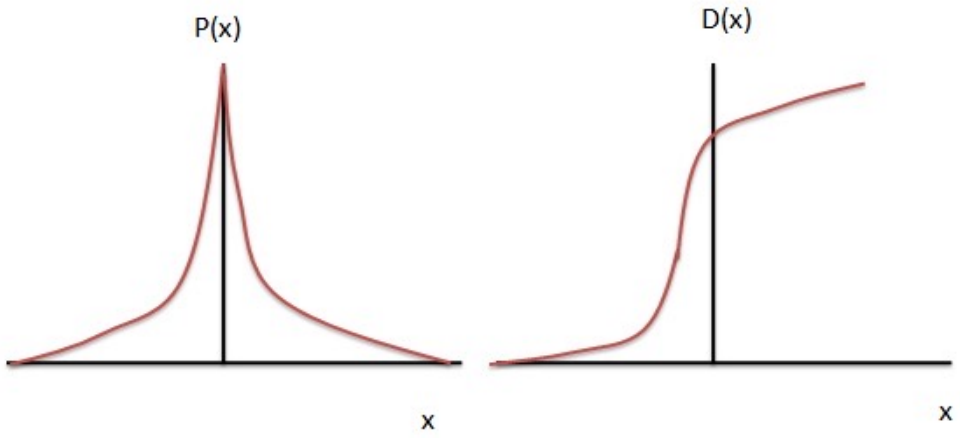
The Laplace distribution

- Also called double sided exponential distribution

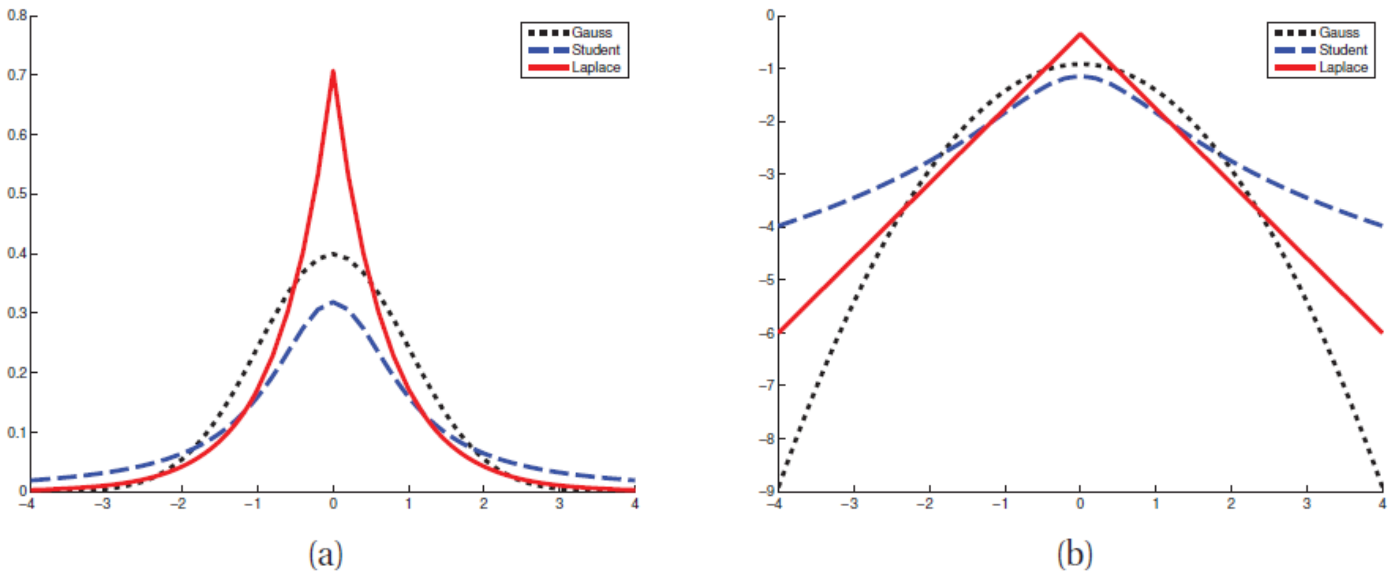
$$\mathrm{Lap}(x|\mu, b) \triangleq \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

Here μ is a location parameter and $b > 0$ is a scale parameter.

$$\text{mean} = \mu, \ \text{mode} = \mu, \ \text{var} = 2b^2$$

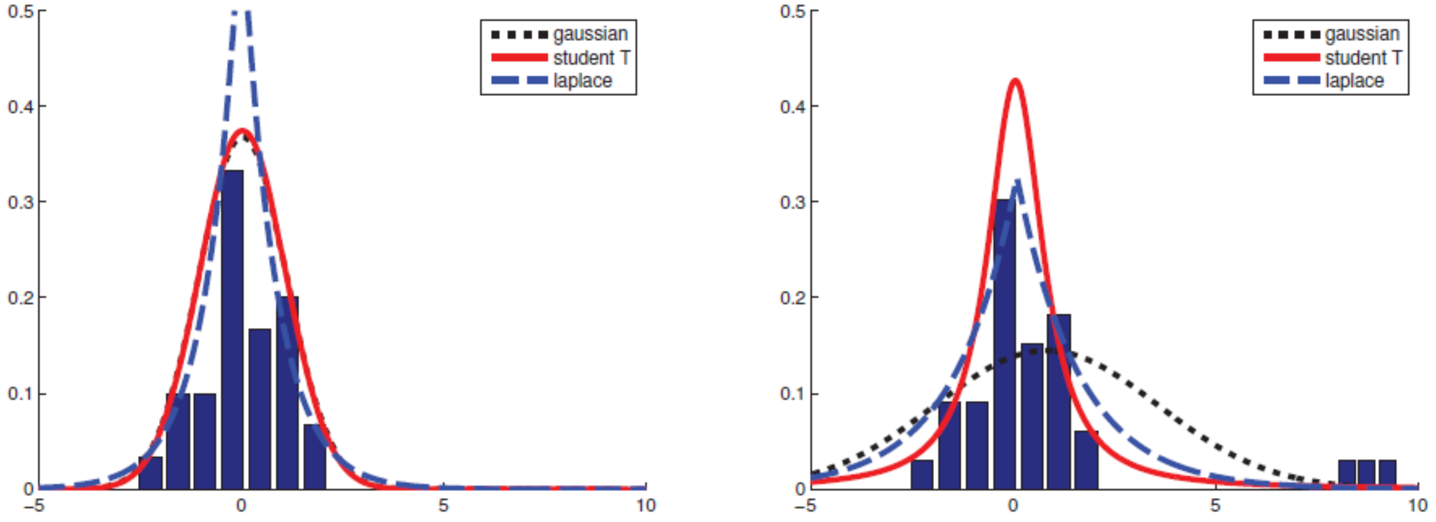


pdf and log(pdf)



(a) The pdf's for a $\mathcal{N}(0,1)$, $\mathcal{T}(0,1,1)$ and $\mathrm{Lap}(0, 1/\sqrt{2})$. The mean is 0 and the variance is 1 for both the Gaussian and Laplace. The mean and variance of the Student is undefined when $\nu = 1$. (b) Log of these pdf's. Note that the Student distribution is not log-concave for any parameter value, unlike the Laplace distribution, which is always log-concave (and log-convex...) Nevertheless, both are unimodal. Figure generated by `studentLaplacePdfPlot`.

Effect of Outliers



(a) (b)

Illustration of the effect of outliers on fitting Gaussian, Student and Laplace distributions. (a) No outliers (the Gaussian and Student curves are on top of each other). (b) With outliers. We see that the Gaussian is more affected by outliers than the Student and Laplace distributions. Based on Figure 2.16 of (Bishop 2006a). Figure generated by `robustDemo`.

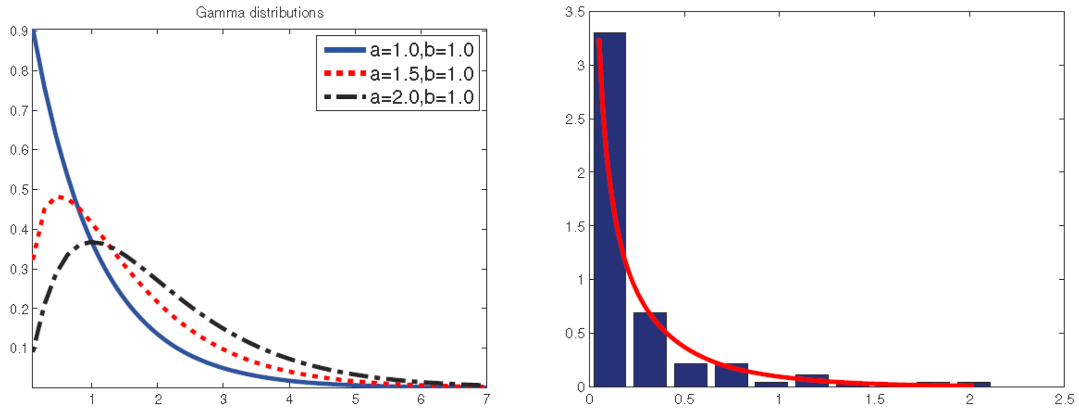
The Gamma distribution

- The Gamma distribution is a flexible distribution for positive real valued random variables

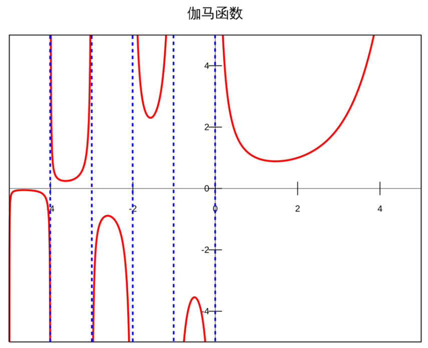
$$\mathrm{Ga}(T|\text{shape}=a,\text{rate}=b) \triangleq \frac{b^a}{\Gamma(a)} T^{a-1} e^{-Tb}$$

$$\Gamma(z)=\int_0^\infty t^{z-1}\mathrm{e}^{-t}\mathrm{d}t,\ \Re(z)>0.$$

$$\text{mean}=\frac{a}{b},\ \text{mode}=\frac{a-1}{b},\ \text{var}=\frac{a}{b^2}$$



(a) (b)



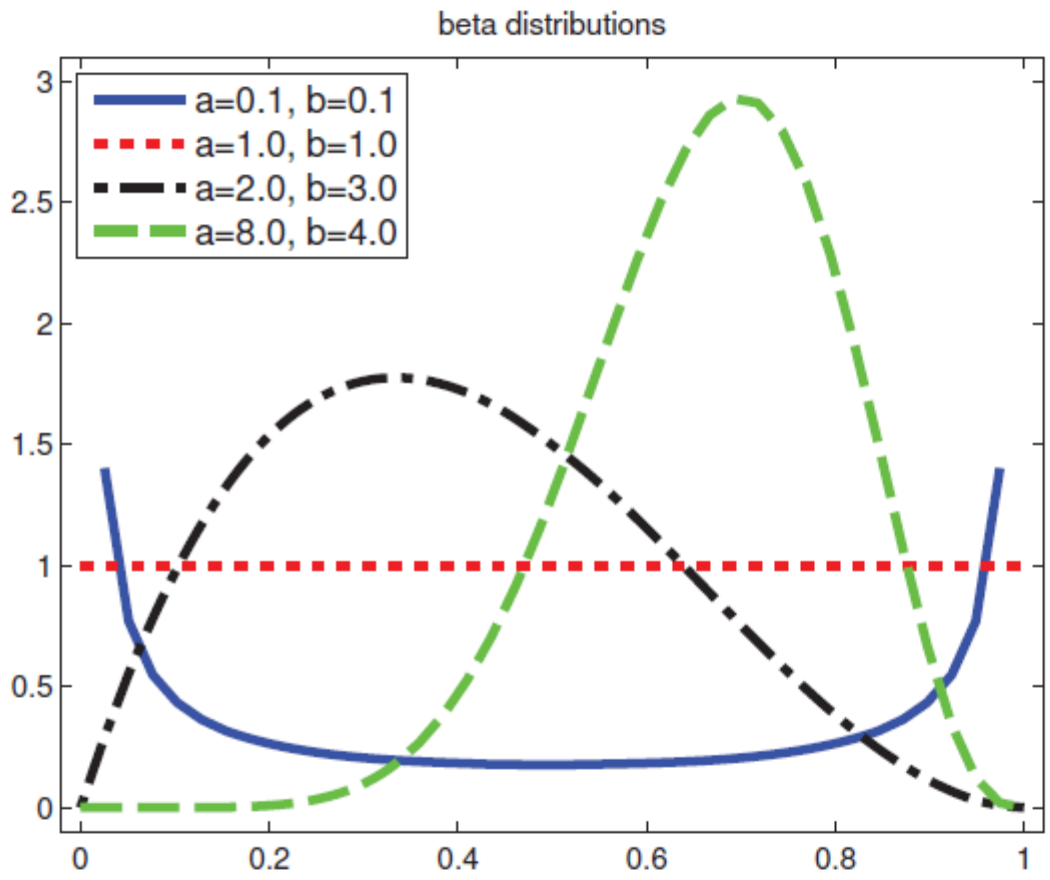
Beta distribution

- Interval [0,1]
- Probability density distribution

$$\boxed{\mathrm{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}}$$

- where B(a, b) is Beta function
  - mean = $\frac{a}{a+b}$
  - var = $\frac{ab}{(a+b)^2(a+b+1)}$
  - mode = $\frac{a-1}{a+b-2}$

Beta distribution



- *a*=*b*=1, uninform distribution
- *a* and *b* <1, bimodal distribution with the spikes at 0 and 1
- *a* and *b* >1, unimodal distribution

Conjugate prior

- Bayes in general

  $p(\theta \mid x) \propto p(x \mid \theta)\, p(\theta)$ -> $p(\theta \mid x)$ is often a complex integral

- With conjugate prior

  - **Prior:**

  $$\theta \sim \mathrm{Beta}(a, b)$$

  - **Likelihood** (for Bernoulli observations $x_1, \ldots, x_n$):

  $$x_i|\theta \sim \mathrm{Bernoulli}(\theta), \quad i = 1, \ldots, n$$

  - **Posterior:**

  $$\theta|x_1, \ldots, x_n \sim \mathrm{Beta}\left(a + \sum_{i=1}^{n} x_i,\ b + n - \sum_{i=1}^{n} x_i\right)$$

Just to update parameters, much easier!

Dirichlet distribution

- A multivariate generalization of the beta distribution
  - Probability simplex

$$S_K = \{\mathbf{x} : 0 \le x_k \le 1, \sum_{k=1}^{K} x_k = 1\}$$

Dirichlet distribution

- A multivariate generalization of the beta distribution
  - Probability density function

$$\boxed{\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K)}$$

where

$$\boxed{B(\boldsymbol{\alpha}) \triangleq \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)}} \qquad \boxed{\alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k}$$

Concentration parameter

Dirichlet distribution

- Property
  - Mean

$$\boxed{\mathbb{E}[x_k] = \frac{\alpha_k}{\alpha_0}}$$

  - Mode

$$\boxed{\mathrm{mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K}}$$

  - Variance

$$\boxed{\mathrm{var}[x_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}}$$

$$\mathrm{mean} = \frac{a}{a+b}$$

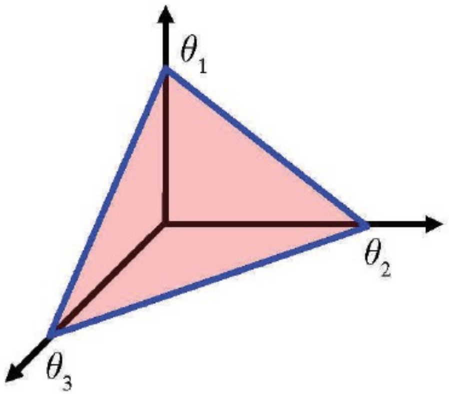
$$\mathrm{var} = \frac{ab}{(a+b)^2(a+b+1)}$$

$$\mathrm{mode} = \frac{a-1}{a+b-2}$$

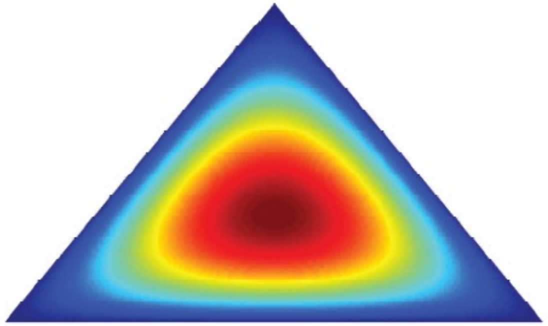
Dirichlet distribution

- Example

$$\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K} x_k^{\alpha_k - 1}\mathbb{I}(\mathbf{x}\in S_K)$$

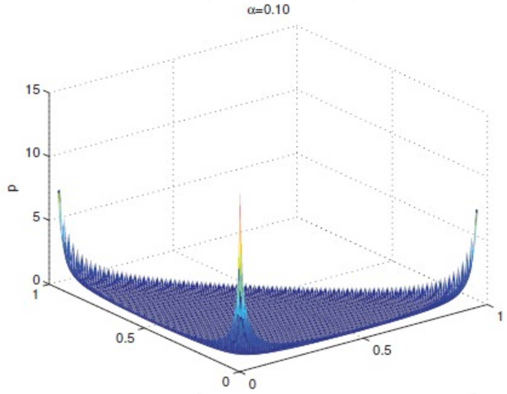


K=3



$\boldsymbol{\alpha} = (2, 2, 2)$

$\boldsymbol{\alpha} = (20, 2, 2)$



$\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$.

Transformation of random variables

- Question: If $X \sim p(\mathrm{x})$ is some random variable, and $Y = f(X)$, what is the distribution of Y ?
  - If f is linear transformation
  - If f is general transformation

Transformation of random variables

- Linear transformation
  - The function f() is a linear function

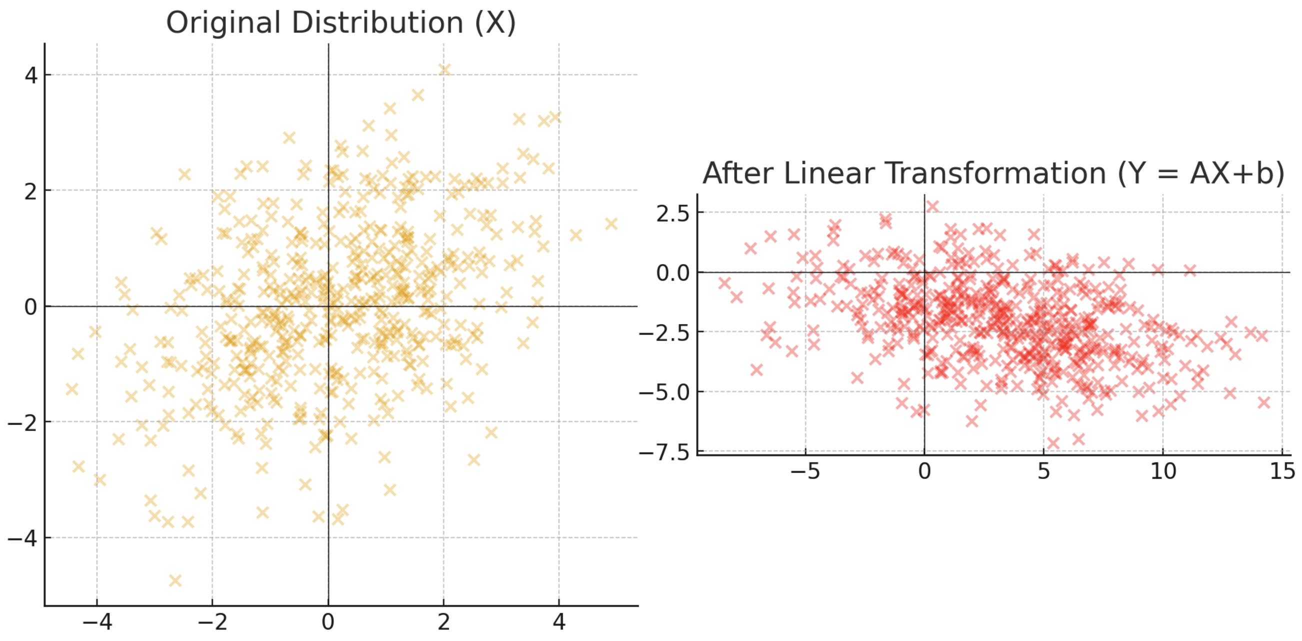
$$Y = f(X) = AX + b$$

where A is a matrix, and b is a vector

$$\boxed{\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[AX + b] = A\mu + b, \\ \mathrm{cov}[Y] &= \mathrm{cov}[AX + b] = A\Sigma A^T, \end{aligned}}$$

Transformation of random variables

- Linear transformation



Transformation of random variables

- Linear transformation
  - Special case
    - If f() is a scalar-valued function

$$Y = f(X) = a^T X + b$$

where a is a vector, and b is a value

$$\boxed{\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[a^T X + b] = a^T \mu + b, \\ \operatorname{var}[Y] &= \operatorname{var}[a^T X + b] = a^T \Sigma a. \end{aligned}}$$

Transformation of random variables

- General transformation
  - If X is discrete

$$p_y(y) = \sum_{x:f(x)=y} p_x(x)$$

    - Example:
      X={1,2,3,4,5}, uniform distribution
      f(X) = 1 if X is even, and f(X) = 0 otherwise

$$\Rightarrow \quad P_y(1) = \sum_{x\in\{2,4\}} P_x(x) = 0.4$$

$$P_y(0) = \sum_{x\in\{1,3,5\}} P_x(x) = 0.6$$

Transformation of random variables

- General transformation
  - If X is continuous

$$P_Y(y) = P(Y \le y) = P(f(X) \le y) = P(X \in \{x : f(x) \le y\})$$

- If f() is an invertible function (change of variables formula)

$$\boxed{P_y(y) = P(f(X) \le y) = P(X \le f^{-1}(y)) = P_x(f^{-1}(y))}$$

$$\boxed{p_y(y) \triangleq \frac{d}{dy} P_y(y) = \frac{d}{dy} P_x(f^{-1}(y)) = \frac{dx}{dy}\frac{d}{dx} P_x(x) = \frac{dx}{dy} p_x(x)}$$

$$\det\left|\frac{dx}{dy}\right|$$

Transformation of random variables

- General transformation
  - If X is continuous (change of variables formula)

$$\boxed{p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|}$$

- Example: X~ U(−1, 1), and Y = X^2. What is $p_y(y)$?

$$\boxed{p_x(x) = \frac{1}{2}}$$

$$\boxed{p_y(y) = p_x(x)\left|\frac{dx}{dy}\right| = p_x(x)\frac{1}{\left|\frac{dy}{dx}\right|} = \frac{1}{2}\frac{1}{|2x|} = \frac{1}{4}y^{-1/2}}$$

Central limit theorem (CLT)

Given i.i.d. $X_1, X_2, \dots, X_N$, each with mean μ and variance σ^2.

Let $S_N = \sum_{i=1}^{N} X_i$, the probability density function of S_N is

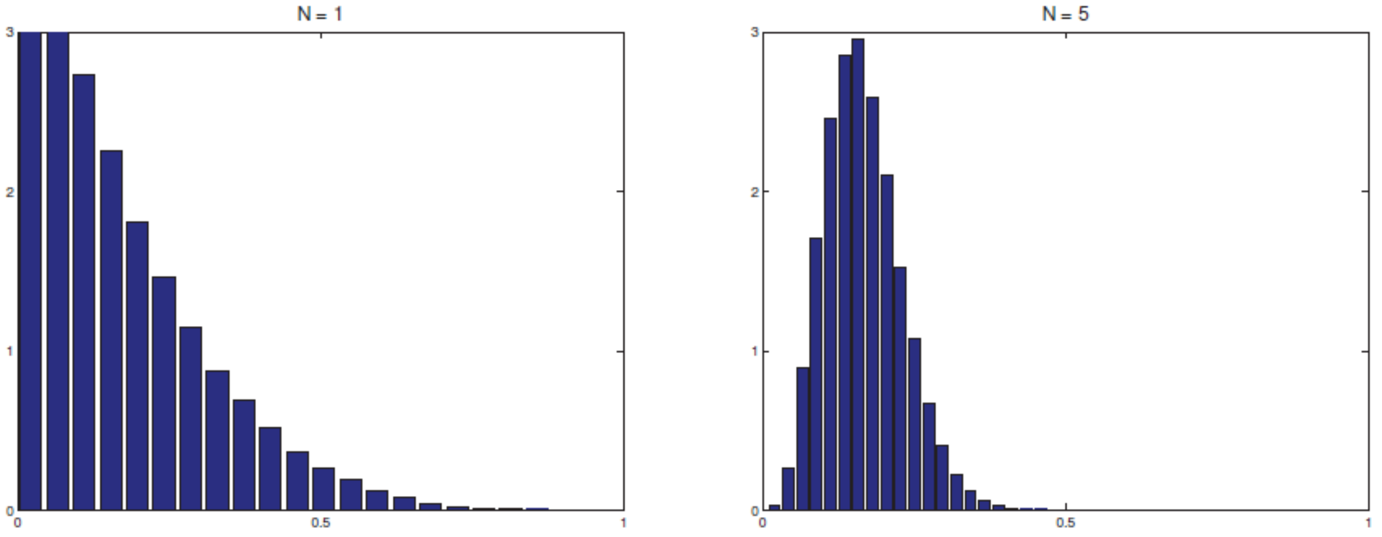
$$\boxed{p(S_N = s) = \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left(-\frac{(s - N\mu)^2}{2N\sigma^2}\right)}$$

Let $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, the probability density function of Z_N

$$Z_N \triangleq \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

converges to the standard normal $\mathcal{N}(0,1)$.

Central limit theorem (CLT)



(a) (b)

Figure 2.17 The central limit theorem in pictures. We plot a histogram of $\frac{1}{N}\sum_{i=1}^{N} x_{ij}$, where $x_{ij} \sim \mathrm{Beta}(1,5)$, for $j = 1 : 10000$. As $N \to \infty$, the distribution tends towards a Gaussian. (a) $N = 1$. (b) $N = 5$. Based on Figure 2.6 of (Bishop 2006a). Figure generated by `centralLimitDemo`.

Monte Carlo approximation

- Question: How to compute the distribution of a function of a random variable X?
  - Generate N samples from the distribution, call them $x_1, x_2, \ldots, x_N$
    - Markov chain Monte Carlo (MCMC)
  - Approximate the distribution of f(X) by using the empirical distribution of $\{f(x_1), f(x_2), \ldots, f(x_N)\}$

$$\boxed{p_{\mathrm{emp}}(A) \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}(A)}$$

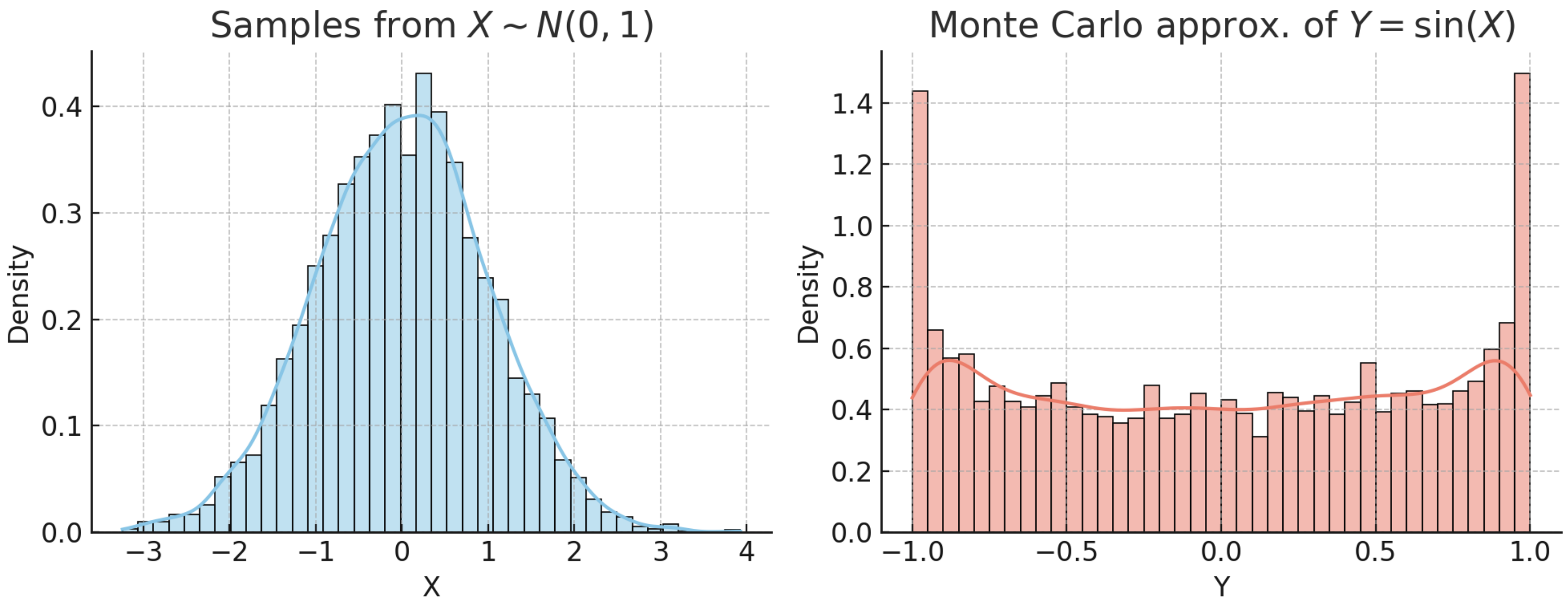
Monte Carlo approximation

- How to Compute $E(f(X))$?
  - Generate $x_1, x_2, \dots, x_N \sim p(X)$

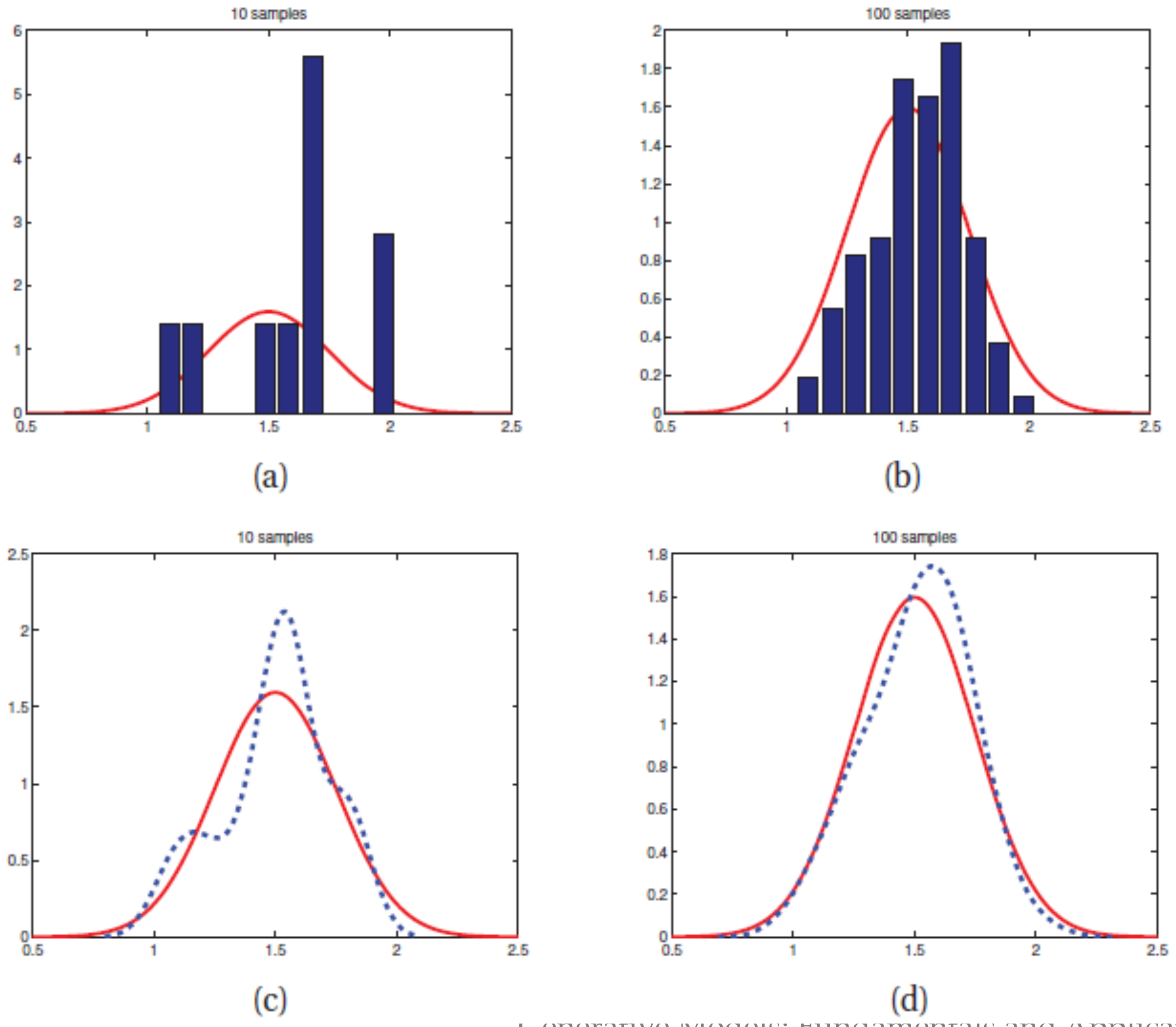
$$E(f(X)) = \int f(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

Monte Carlo approximation

- Example: X~ N(0, 1), and Y =sin(X). What is $p_y(y)$?



Accuracy of Monte Carlo approximation



10 and 100 samples from a Gaussian distribution, $N(\mu = 1.5, \sigma_2 = 0.25)$. Solid red line is true pdf. **Top line**: histogram of samples. **Bottom line**: kernel density estimate derived from samples in dotted blue

Accuracy of Monte Carlo approximation

- Denote by μ the exact mean of f(X)

$$\mu = E[f(X)]$$

- Denote by σ^2 the exact variance of f(X)

$$\sigma^2 = \mathrm{var}\left[f(X)\right] = \mathbb{E}\left[f(X)^2\right] - \mathbb{E}\left[f(X)\right]^2$$

- Denote by $\hat{\mu}$ the mean of MC approximation

$$\hat{\mu} - \mu \rightarrow \mathcal{N}(0, \frac{\sigma^2}{N})$$

- Denote by $\hat{\sigma}^2$ the variance of MC approximation

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - \hat{\mu})^2$$

Information theory

- Entropy ($\mathbb{H}(X)$ or $\mathbb{H}(p)$)
  - Measure of the uncertainty of a random variable X with distribution p

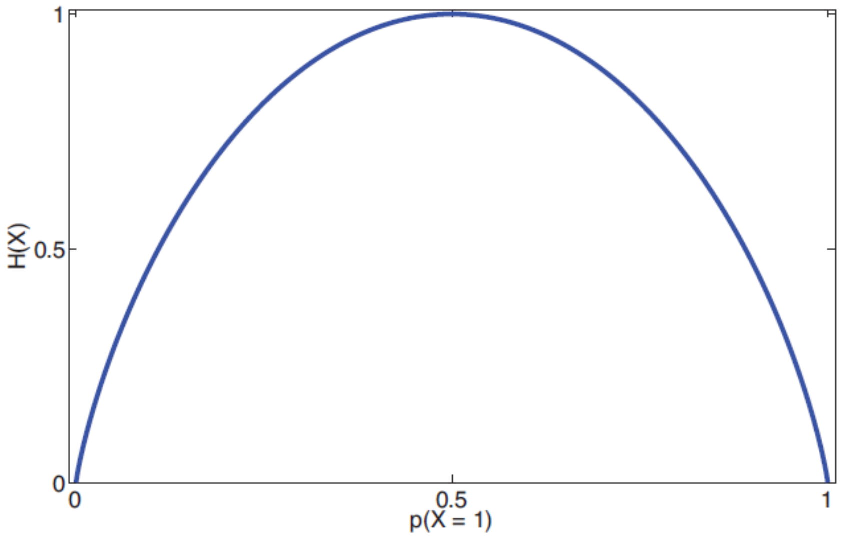
$$\boxed{\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k) \log_2 p(X=k)}$$

    - Entropy is maximized if $p(X=k) = \frac{1}{K}$ (uniform distribution)
    - Entropy is minimized if distribution with delta-function that has all its mass on one state.

Information theory

- Example: binary random variable $X \in \{0,1\}$
  - $p(X=1)=\theta$
  - $p(X=0)=1-\theta$

$$\begin{aligned}\mathbb{H}(X) &= -[p(X=1)\log_2 p(X=1) + p(X=0)\log_2 p(X=0)] \\ &= -[\theta\log_2\theta + (1-\theta)\log_2(1-\theta)]\end{aligned}$$



Information theory

- Conditional Entropy
  - The remaining uncertainty of X when Y is already known

$$H(X|Y) = -\sum_{x,y} p(x,y) \log p(x|y)$$

- Joint Entropy
  - The total uncertainty when considering variables X and Y together

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y)$$

Information theory

- Cross entropy
  - The average number of bits needed to encode data coming from a source with distribution p when we use model q to define our codebook

$$\boxed{\mathbb{H}(p, q) \triangleq -\sum_k p_k \log q_k}$$

  - Special case

$$\boxed{\mathbb{H}(p) = \mathbb{H}(p, p)}$$

Information theory

- Example:
  - Suppose tomorrow's weather has **true distribution**:
    - $p = \{P(Sunny) = 0.9，P(Rainy) = 0.1\}$

$$H(p) = -[0.9 \log_2 0.9 + 0.1 \log_2 0.1] \approx 0.47 \text{ bits.}$$

  - If we model q that assumes equal probability: $q = \{0.5，0.5\}$, the cross-entropy becomes

$$H(p, q) = -[0.9 \log_2 0.5 + 0.1 \log_2 0.5] = 1 \text{ bit.}$$ -> At more expense

Information theory

- KL divergence (relative entropy)
  - Measure the dissimilarity of two probability distributions p and q

$$\mathbb{KL}\left(p||q\right) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

$$= \sum_{k} p_k \log p_k - \sum_{k} p_k \log q_k = -\mathbb{H}\left(p\right) + \mathbb{H}\left(p, q\right)$$

Information theory

- KL divergence (relative entropy)
  - Theorem (Information inequality)

$$\mathbb{KL}\,(p||q) \geq 0 \text{ with equality iff } p = q.$$

  - Result: discrete distribution with the maximum entropy is the uniform distribution

$$\mathbb{H}\,(X) \leq \log|\mathcal{X}|$$

$$\begin{aligned} 0 \leq \mathbb{KL}(p||u) &= \sum_x p(x) \log \frac{p(x)}{u(x)} \\ &= \sum_x p(x) \log p(x) - \sum_x p(x) \log u(x) = -\mathbb{H}(x) + \log|\mathcal{X}|. \end{aligned}$$

Information theory

- Mutual information
  - Find out how much knowing one variable can tell us about the other

$$\boxed{\mathbb{I}(X;Y) \triangleq \mathbb{KL}\left(p(X,Y)||p(X)p(Y)\right) = \sum_x \sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)}}$$

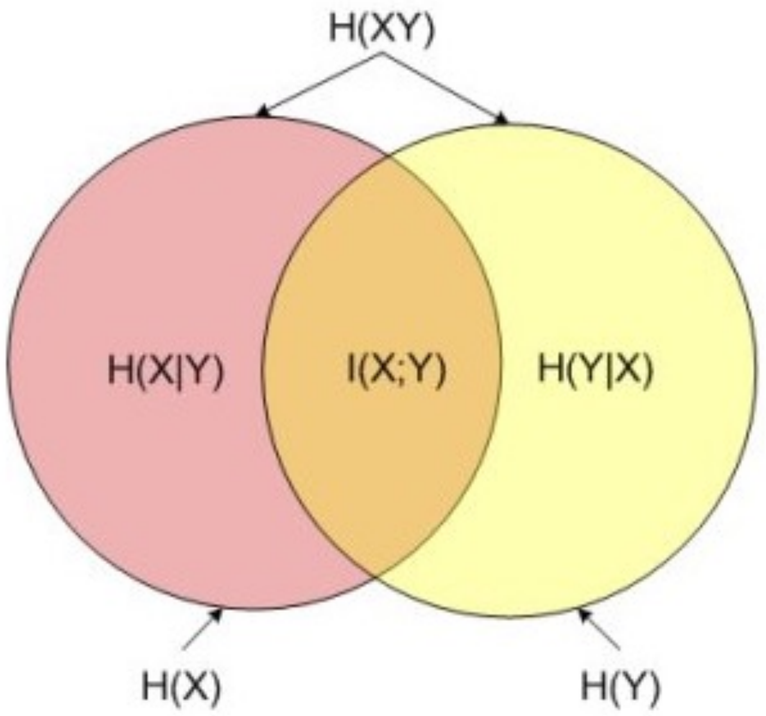
  - Symmetry: $\mathbb{I}(X;Y) = \mathbb{I}(Y;X)$
  - Non-negativity: $\mathbb{I}(X;Y) \geq 0$ with equality iff p(X,Y)=p(X) p(Y)

Information theory

- Mutual information
  - Measure the reduction in uncertainty about X after observing Y
  - Measure the reduction in uncertainty about Y after observing X.

$$\boxed{\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X)}$$

Information theory



$$
\begin{aligned}
\mathbb{I}(X;Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y) \\
&= H(X,Y) - H(X|Y) - H(Y|X)
\end{aligned}
$$

Summary

- Basic probability
  - Frequentist vs. Bayesian
  - Event, random variable, probability, conditional/joint probability, Bayes rule, independence, conditional independence, correlation
  - Common statistics: mean, median, mode, variance, standard deviation, covariance, correlation coefficient
- Distribution
  - Empirical, binominal/Bernoulli, multinominal/Multinoulli, uniform, Gaussian, multivariate Gaussian, Student t, Laplace, Poisson, Beta, Gamma, Dirichlet
- Transformation of variables
  - Linear transformation, general transformation, CLT, Monte Carlo approximation
- Information theory
  - Entropy, conditional/joint entropy, cross-entropy, KL divergence, mutual information

Thanks!

Questions?